KNOWING AND EXPRESSING OURSELVES

BENJAMIN WINOKUR

A DISSERTATION SUBMITTED TO THE

FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN PHILOSOPHY

YORK UNIVERSITY

TORONTO, ONTARIO

FEBRUARY 2021

©BENJAMIN WINOKUR, 2021

Abstract

This dissertation concerns two epistemologically puzzling phenomena. The first phenomenon is the authority that each of us has over our minds. Roughly, to have authority is to be owed (and to tend to receive) a special sort of deference when self-ascribing your current mental states. The second phenomenon is our privileged and peculiar self-knowledge. Roughly, self-knowledge is privileged insofar as one knows one's mental states in a way that is highly epistemically secure relative to other varieties of contingent empirical knowledge. Roughly, one has peculiar self-knowledge insofar as one acquires it in a way that is available only to oneself.

In Chapter One I consider several more detailed specifications of our self-ascriptive authority. Some specifications emphasize the relative indubitability of our self-ascriptions, while others focus on their presumptive truth.

In Chapter Two I defend a "Neo-Expressivist" explanation of authority. According to Neo-Expressivism, self-ascriptions are authoritative insofar as they are acts that put one's mental states on display for others, whether or not these mental states are also known by the self-ascriber with privilege and peculiarity. However, I do not dispute that we often have privileged and peculiar self-knowledge. This raises the question of what such knowledge does explain, if not the authority of our self-ascriptions.

In Chapter Three I examine several extant answers to this question, focusing on privileged and peculiar self-knowledge of the propositional attitudes. Each answer meets with objections.

In Chapter Four I develop a "Social Agentialist" account of the explanatory indispensability of privileged and peculiar self-knowledge. I argue that such knowledge enables at least three forms of 'social-epistemic agency': interpersonal reasoning, complex group action, and linguistic interpretation. Next, I argue that, even though privileged and peculiar self-

knowledge does not explain the authority of our self-ascriptions, it is importantly related to our (Neo-Expressively understood) authority.

In Chapter Five I consider possible sources of our privileged and peculiar self-knowledge, focusing again on propositional-attitudinal self-knowledge. I eventually defend a "Constitutivist" view. This is the view that, for agents who meet certain background conditions, self-knowledge is privileged and peculiar because it is metaphysically built into the attitudes self-known.

For my parents

Acknowledgements

I was not remotely confident in my philosophical abilities when I first arrived at York to work toward my doctorate. Frankly, there are still days where I feel like a deer in philosophy's headlights. This being said, I know that my personal and philosophical growth has been significant throughout my time here. So much of this owes to the unwavering efforts of my supervisor, Claudine Verheggen. I still remember awkwardly stumbling into her office on the day that I intended to ask for her supervision. I also remember her expression when she agreed, as if it was absurd for me to have worried about rejection. Since then, she has read an outrageously large amount of my work, always knowing when to criticize my arguments or offer some (far too) kind words. If it were not for her, I would never have finished this dissertation. There's just no other way to put. My gratitude toward her is immeasurable.

I am also deeply thankful for the efforts of Christopher Campbell and Henry Jackman. Their comments on drafts of each chapter in this lengthy dissertation have greatly improved their quality, even if I cannot be sure that I always did justice to their concerns. More importantly, they agreed to serve on my committee. Further thanks are owed to Shanya Rosenbaum and Victoria McGeer for rounding out my committee, and to Regina Rini for chairing my defense.

I also want to thank my mentors who did not serve on my committee, but whose influence on my dissertation has been significant nevertheless. First of these is Robert Myers, who has helped me to grappled with the work of several philosophers that figure heavily in my dissertation and broader philosophical thinking. He is also to be credited with reading and commenting drafts of some of my papers, portions of which have been adapted to this dissertation.

Second, I want to thank Rockney Jacobsen. When I studied at Wilfrid Laurier University, prior to arriving at York, Rockney was on sabbatical and so I never encountered any of his work. Indeed, I did not even encounter the topics on which this dissertation was written. Fortunately, once I found my way to the topics of self-knowledge and self-expression, he dedicated a great deal of time to chats and email correspondences with me.

Third, and finally, I owe a great deal of thanks to Dorit Bar-On. I first met her over lunch when she was visiting York to deliver a talk to our department. She humoured many poorly-formed questions about her book, *Speaking My Mind*. Years later, we resumed our correspondence and have continued it actively up to the present day. Her views are among the most prominently discussed in this dissertation. Were it not for her generosity, I would not understand them nearly as well as I (hopefully) do.

Support from one's professors and mentors is indispensable to finishing a dissertation. But one also needs the support of one's friends. In my case, I am happy to think of just about every one of my fellow graduate students as a friend. Here I name some of those to whom I have grown closest over the years: Alex Leferman, Olivia Sultanescu, Sam Steadman, Ali Karbaeli Mahdi, Dylan Ludwig, Lauren Edwards, Brandon Tinklenberg, Dennis Papadopoulos, Jef Delvaux, Bogdan Florea-Alexandru, and David Rocheleau-Houle. I have learned so much from them. Perhaps more importantly, I have partied with them, travelled with them, and commiserated with them.

My soul has a strong Dionysian component. That aspect of me would be dead, or close to it, if it were not for my many friends outside of the academy who know how to relax. Listing all of them would be impossible. Fortunately, they know who they are.

Finally, I thank my parents for the unparalleled support they have always provided for me.

Table of Contents

Chapter One—Authority

§1.1.1—Introduction

Spring has bloomed. You and I are walking through a park and see two women hunched over a chess board. Their brows are furrowed, deeply focused. I look at you and say "I want to learn to play chess someday. I believe it is the finest game of intellects." As we walk onward, we see a homeless person resting against a tree. I look at you again and say "I believe that our nation's economic policies are responsible for our homelessness crisis, and I intend to do something about it!" Later on, the sun begins to set. I look to you once more and say "I'm tired and hungry: I want to go home." You agree, and we soon part ways.

Is there anything remarkable about this scenario? Well, perhaps it is remarkable that I want to learn to play plain old chess in a time where I could just as easily do battle with intergalactic space pirates on my computer. And perhaps it is remarkable that I am reporting my economic beliefs to you, as well as my bold intentions to change economic policies in my homeland, when I never took a single economics course as an undergraduate and have the mathematical reasoning skills of an eighth grader. This is a philosophy dissertation, however, and so my question is really whether there is anything *philosophically* remarkable going on here.

On first pass, it is hard to see what could be remarkable here: the scenario is apt to strike one as a perfectly mundane conversation between two friends. Still, philosophers are adept at uncovering puzzling features of the everyday, and aspects of the above scenario have struck many of them as no exception. As we shall see, what is puzzling here is difficult to elucidate precisely. Indeed, the primary function of this chapter will be to elucidate it. Still, it will be helpful to begin with some initial characterizations of our puzzle.

What we are interested concerns the sorts of thoughts and utterances that are at issue in statements like those above. These are *self-ascriptions* of mental states: thoughts and utterances that ascribe beliefs, desires, pains, fears, and all the rest of one's mental goings-on, to oneself. What is so puzzling about them? Here, philosophers have bandied around various terms. But perhaps the most common term used is that such self-ascriptions are, in some sense, puzzlingly *authoritative*. Here is a small sample of claims about the authority of self-ascriptions:

(1) We are justified in, or are obligated to recognize, or actually do cede "deferential trust" to self-ascriptions (Parrott 2015, p. 2219)

(2) Self-ascriptions "carry so much more weight than anyone else's pronouncements on the same matters…" (Bar-On 2004, p. 10).

(3) Self-ascriptions enjoy "relative indubitability" (Bar-On 2004, p. 310), such that "[i]f you are sincere and competent with respect to the concepts you use to express your mental states, nobody can—rationally—cast any doubt" on them (Coliva 2016, p. 62).

As we will see, there are important differences between these characterizations, provisional as they are. But a puzzle should already be apparent, if only in an equally provisional way. For, just like any other empirical claims, self-ascriptions ascribe totally contingent states of affairs. It is a contingent fact that I want to help the homeless; it is also a contingent fact that there are homeless people. Why, then, should there be significant epistemic differences between thoughts or claims about these facts?

Philosophers often oscillate between talking about our *actual practices* of treating self-ascriptions as authoritative, on the one hand, and our *justifications for* or *obligations to* treat them as authoritative, on the other. This is not (always) an illicit oscillation, since many philosophers believe that authority has both descriptive and normative dimensions. In this

dissertation I too will be interested in authority as both a normative and descriptive phenomenon, and though I too will often oscillate between normative and descriptive treatments of it, it should be kept in mind, unless explicitly stated otherwise, that the ensuing discussions apply—or will be made to apply, in time—to both.

The reader who has followed me this far may have been stifling a good deal of uneasiness, however. For it might seem that authority admits of obvious counter-examples. Thus, time and again we find that it is not us, but our loved ones, who are in the best position to say what is on our minds. Likewise, we might distrust what people say about their minds because we take them to be self-deceived or cognitively impaired. In other cases, a listener might defer to a speaker's self-ascriptions despite the fact that she *ought not* to, perhaps because she is not aware that the speaker's capacity to reliably self-ascribe her mental states has been compromised. Owing to these and other cases, the days are long gone when philosophers took self-ascriptions of mental states to be *universally* authoritative, whether in the sense that they universally ought to be accorded authority, or in the sense that they are universally treated as such. But if self-ascriptions are not universally authoritative, how should we home in on genuinely authoritative cases?

Here is a platitudinous starting point: self-ascriptions are first-person authoritative in *good self-ascriptive conditions*, and ought to be treated as such when these conditions obtain. Good self-ascriptive conditions are conditions where, e.g., one is not under cognitive duress, enjoys overall cognitive well-functioning, and where there is no reason to suspect that one's self-ascription is insincere, or that one does not actually understand the concepts involved in one's self-ascription, or that one has suffered a slip of the tongue. Equally platitudinously, we might say that *bad* self-ascriptive conditions are conditions in which any number of these cognitive impairments or mundane missteps *do* occur. However, there is a further complication: sometimes

we may take a speaker's self-ascriptions to be authoritative in light of our failure to recognize that it was issued in bad self-ascriptive conditions. So our actual practices of taking self-ascriptions to be authoritative will not perfectly co-vary with facts about the authority we are justified in or obligated in taking them to have. In this way, since authority has descriptive and normative dimensions, it may be that a single self-ascription can be authoritative in one sense (e.g., in being treated as such) while lacking authority in another sense (e.g., not being owed any recognition of authority).

Even with these broad qualifications in place, one might wonder whether all of this is *still* too quick, for it might now seem that many self-ascriptions lack authority even in good self-ascriptive conditions, and even when they are rightly taken to be issued in good self-ascriptive conditions. Consider, for example, self-ascriptions of mental states that one *used to have*. One might question whether such self-ascriptions are authoritative because they are prone to all the failures that infringe on our memory processes generally. Exceptions may be self-ascriptions of *recently passed* mental states, such as the self-ascription of a desire for tea that one had just a moment before one took one's first sip of it. These cases aside, we all recognize that memory is fallible in a host of ways. Self-ascriptions about past mental states seem like poor candidates for being first-person authoritative.[1] They need not be automatically owed any sort of deference, for one can easily imagine that they are mistaken.

For this reason, philosophers typically agree that intuitions about authority pertain mostly (if not entirely) to *present-tense* self-ascriptions of mental states. But it is not *just* present-tense self-ascriptions of mental states that will be my focus. Rather, it is also present-tensed self-ascriptions of mental states that are also sufficiently *first-personal* in character, hence the

[1] Similar worries can be brought to bear on self-ascriptions of mental states one *predicts* one will have in the future.

common label of 'first-person authority' that is used instead of our current shorthand, 'authority', by some authors (Davidson, 1984). What it is to self-ascribe a mental state from the first-person point of view is an open question at this stage, but whatever it is, it cannot be that it is issued from the sort of point of view that others have on us. The reason is straightforward: if only self-ascriptions are authoritative, then another person cannot authoritatively ascribe a mental state to you, and so your self-ascription should not be made from a perspective that is relevantly similar to the perspective of another person.

Because we can now see that there are self-ascriptions that are *not* authoritative, and because I wish to avoid having to use cumbersome phrases like *self-ascriptions of present mental states made from the first-person point of view* to refer to the class of self-ascriptions in which I am interested, it will be prudent to consolidate our language. For this reason, I will adopt the term 'avowal' from Gilbert Ryle (1949) to refer to the relevant class of self-ascriptions.[2] Note, however, that I do not mean the term 'avowal' to signify self-ascriptions that are by definition authoritative (though context will show that by 'avowal' I *almost* always mean 'authoritative avowal'). For I want to leave open that one might issue a present-tense self-ascription of a mental state, from a sufficiently first-person point of view, and yet be deprived of authority in any number of respects, perhaps because one's avowal is misunderstood by its receiver and so is not *treated* with the deference it is properly owed. Still, on my usage, it is only avowals that are *proper candidates* for being (taken as) authoritative.

§1.1.2—Clarifying Our Explanandum, First Pass: Authority

I eventually hope to defend an account of the authority of avowals. Avowals are self-ascriptive thoughts or utterances of a certain sort, and so one way of putting our question is: what special

[2] I do not claim that Ryle and I have the *exact* same meaning of 'avowal' in mind.

property or properties do avowals possess, in virtue of which they are authoritative? However, some have argued that we cannot account for authority by focusing on avowals themselves. These authors propose, instead, that we should consider properties of avowing *agents* that explain authority. This dissertation will function, in part, as a sustained meditation on whether avowals themselves should be our focus when trying to explain authority. But for the moment, my aim is not to make a start on this question. This is because, so far, we have only considered a small sampling of brief claims about what authority amounts to. The most pressing matter, then, is to clarify exactly what authority amounts to in the first place.

Here is the most bare-bones version of the thesis we must develop:

> **Authority Thesis**: avowals of one's current mental states enjoy a kind of authority that other-ascriptions and other contingent empirical claims lack.[3]

The Authority Thesis (AT) is just a skeleton, one that does not have many bones. Still, it is not *altogether* boneless, for it does stipulate an important difference between avowals and two other types of empirical claims. First, there are ordinary contingent claims about various non-mental empirical goings-on, e.g., that there are seventeen buses at the bus terminal. Second, there are *other-ascriptions*, which are ascriptions of mental states to some person, *P*, made by some *other* person, *O*, rather than by one's self—for example, Tim's claim that I believe that there are seventeen buses at the bus terminal.

What AT does not say, and what I want to emphasize now, is that authority is a *relational* property: a self-ascription is authoritative if someone does, ought to, or is justified to treat it as such. The simplest way of developing the contrast is overtly social: authority is (or ought to be) conferred by *hearers* on the self-ascriptions of *avowing speakers*. As a universalizing

[3] This formulation is a slightly modified version of Barz's, who describes authority thus: "Self-ascriptions enjoy a kind of first-person authority that other-ascriptions lack" (2018, p. 126). I believe that Barz would accept this modification (namely, the "and other empirical claims…" qualifier) without argument.

characterization, however, this is too strong. Three reasons are these. First, it is possible for the avower to recognize her own avowal as authoritative. Second, and more interestingly, avowals can be made in silent thought, and they can be authoritative all the same. To make sense of this, we can follow Dorit Bar-On in saying that the authority of an avowal-in-thought "may be construed as pertaining to how *we*, who envisage the subject's mental act of avowing, regard it" (2004, p. 310, fn. 20). Third, and finally, it is possible for avowals to be authoritative even if they are not *directly* issued from speaker to hearer in a live social context. To see this, consider the following case:

> **Bieber's Diary**: Justin Bieber, frustrated that his new song "Yummy" is doing poorly on the Billboard Top 100 chart, writes in his diary: "I'm so angry. 'Yummy' is an incredible song!"

We can reasonably (and should) dispute the claim that "Yummy" is an incredible song. But if prevailing intuitions about authority are on the right track, then his avowal of anger cannot be so easily disputed. And yet, plausibly, Bieber never intended for anybody to see his diary.

With these points in mind, it is now time to consider what exactly authority—understood, we might say, as a *quasi-social* phenomenon—amounts to. This is the task of the rest of this chapter. First, I will examine a range of specifications of authority that have either been explicitly offered by philosophers, or that can be charitably reconstructed from the philosophical literature on authority. Many of these will meet with objections, and at one point it may appear to some readers as though we can have no adequately specified definition of authority after all. In response to these objections, I will offer several improved specifications of authority. The result, as we shall see, is that authority amounts to a *cluster* of properties: there turn out to be multiple senses in which avowals can be authoritative. It is only once we have adequately specified authority, in its various permutations, that we will be in a position to consider its explanations.

At the close of this chapter, I will return to the question of the methodology we should adopt in aiming to explain it. After describing my preferred methodology, I will defend an account of authority (in all its permutations) in Chapter Two.

§1.2.1—Barz's Dilemma

We should explain authority only if it warrants explanation, and we can only determine whether authority warrants explanation if we have adequately specified its purported nature(s). To get us started, I will I will be responding to Wolfgang Barz's (2018) recent argument that authority—specified in various ways—does not warrant explanation, at least not *philosophical* explanation. Barz is skeptical about whether authority warrants philosophical explanation because, as he sees it, extant specifications succumb to a dilemma: they all turn out to be (1) philosophically puzzling but false, or (2) true but philosophically unpuzzling. Either horn spells trouble for me: the first entails that I have been studying a chimera, and the latter entails that I—qua philosopher—have no business trying to explain authority (as I will try to do in Chapter Two).

Against Barz, I will argue that his dilemma can be circumvented by several carefully specified authority theses. Thus, while the rest of this chapter is partly intended to vindicate common philosophical puzzlement over how authority is to be explained, it also argues that philosophers ought to take up my specifications of authority (to the extent that they are not already doing so). My defenses of each specification will involve various combinations of (1) appeals to intuition, (2) textual evidence that other philosophers have in fact accepted them, (3) reasons to view these specifications as philosophically puzzling, and (4) responses to objections, where applicable. Before we get there, two tasks are in order. First, we must examine Barz's three requirements that he thinks any adequate specification of authority must meet. Second, we must consider those specifications of authority that Barz himself develops and rejects.

§1.2.2—Barz's Three Requirements

Barz sets up his dilemma by putting forward three requirements for any adequately specified conception of authority, foreshadowing that no specification of authority meets all three. According to him, any adequately specified conception of authority must:

(1) "…delineate the alleged authority as a feature that at least a sufficiently large number of *self-ascriptions have*."
(2) "…delineate the alleged authority as a feature that…*other-ascriptions…*lack."
(3) "…delineate the alleged authority as a thesis that *gives rise to a kind of puzzlement distinctive of philosophical problems*." (2018, p. 127)

As we will see, Barz's dilemma turns on the claim that no extant specification of authority meets all three of these requirements. So it is important to carefully consider each of them.

Barz's requirement (1) is intended to rule out specifications of authority that only account for a problematically narrow range of authoritative avowals. However, because narrowness is a threshold property (as is *problematic* narrowness), there might be reasonable disagreement about its instances. On my view, this introduces some initial unclarity into the requirement. Still, to see what Barz is after, we can take up a case that he discusses. Say, then, that I utter "I am thinking that the sky is blue". Many philosophers agree that avowals like this are *self-verifying* (Burge 1988). This is because they are made true "*in* and *through*" their being thought: in the very thinking of the content, my thought that I am thinking it is true (Borgoni 2018b, p. 683). Now, one might construct a specification of authority according to which avowals are authoritative because self-verifying, whereas other-ascriptions and other empirical claims are not authoritative because not self-verifying. However, even if this works for the relevant cases, most authoritative avowals are not self-verifying. This is because most avowals attribute specific *kinds of states* to the speaker—states that are not possessed simply in virtue of one's entertaining a thought with the same *content* as a possible state. For example, uttering 'I want to eat cookies' is not self-

verifying. It is only true if one desires to eat cookies, and having this desire is not guaranteed simply in and through thinking a thought that self-attributes it.[4] Similarly, thinking "I am in pain" is not sufficient for being in pain.

Barz is right that self-verifying thoughts comprise a narrow range of putatively authoritative avowals. However, whether a specification of authority that is only suited to such avowals is *problematically* narrow remains unclear.[5] The reason concerns the possibility of pluralism about authority. By 'pluralism' I do not mean the view that there are many adequate authority theses (though, as we will see, I do hold this view), for that would leave open the possibility that each specification must account for every kind of authoritative avowal. Rather, I mean the view that avowals of different types of mental states might be authoritative in different ways or to different degrees. Such positions are not unprecedented (Wright 2001; Parrott 2015), and so it may be possible to account for self-verifying avowals in one way while accounting for other kinds of authoritative avowals in another. We will return to this issue later on.

The case for Barz's requirement (2) is more straightforward; it simply expresses the demand that one preserve the asymmetry in authority, baked into our initial version of AT, between avowals and other-ascriptions. If a given specification of authority entails that other-ascriptions are also authoritative, then we will not have successfully isolated anything distinctive about avowals.

Finally, there is Barz's requirement (3), in defense of which he offers a metaphilosophical view. On this view, a philosophical puzzle is resistant to purely empirical resolution and so tends to persist despite advances in the natural sciences (2018, p. 127). The idea behind requirement

[4] Burge also takes some thoughts of the form *I judge that p* to be self-verifying, as when they are 'used to execute [and] not merely describe a judgement' (1996, p. 92).
[5] Barz does take this specification to be problematically narrow (2018, pp. 130-131).

(3), then, is that we should wonder why philosophers have cared so much about explaining authority "if it [leads] to questions answerable by empirical means alone" (2018, p. 128). Despite the fact that neither Barz nor myself offer an exact account of what it means for a question to be answerable by empirical means alone, nor an exact account of what constitutes a genuinely philosophical solution to a puzzle (perhaps conceptual analysis or the method of cases are viable methodologies), I will assume that the basic idea is clear enough to my target audience of academically trained philosophers in the analytic tradition.[6]

It is worth noting that Barz sometimes oscillates between claiming that authority must be philosophically puzzling and that it must be philosophically *interesting*. Barz might take these to amount to the same thing, but they are not obviously so. Perhaps, then, a philosopher could be in a special position to show us that, philosophically unpuzzling though it may be, authority is nevertheless philosophically interesting, perhaps because of its connections to other philosophically interesting or puzzling phenomena.[7] To take two examples, Borgoni (2019) has argued that denying an avowal's authority can constitute an epistemic injustice, while Victoria McGeer (1998, 2007, 2015) has argued that authority is important to each agent's moral development and "mindshaping" capacities. With this in mind, we might grant that authority is not philosophically puzzling but deny that this undermines the legitimacy of our philosophical interest in it.

To Barz's credit, however, many philosophers have expressed full-blown puzzlement about authority: they have asked what explains its truth and have advanced sophisticated theories, while employing many fine-grained philosophical distinctions, to answer this question.

[6] At any rate, it has been espoused by well-regarded philosophers (see, e.g., Bonjour 2009, pp. 1-2).
[7] Thanks to an anonymous reviewer at *Journal of Philosophical Research* for suggesting this point.

Moreover, I myself believe that authority is philosophically puzzling. So, my aim is to show that philosophers need not settle only for authority's philosophical interestingness.

But even this does not mean that Barz's requirement (3) is unobjectionable. For one might also wonder why something that is philosophically puzzling requires a distinctively philosophical (read: non-empirical) explanation. Barz does not say why a properly philosophical puzzle cannot be one that is simply *puzzling from a philosophical perspective*, even if it is not *soluble using purely philosophical resources*. His metaphilosophical approach builds in the idea that a philosophical puzzle is soluble using purely philosophical resources, but perhaps this could be rejected. Indeed, authority might be puzzling from an epistemological perspective even if it is explicable from an empirical-psychological perspective. If this is right, then we can vindicate common philosophical puzzlement about authority without snuffing out the explanatory potential of other disciplines. To be clear, I myself will eventually argue that we can explain authority by drawing on resources from the philosophies of language, mind, and epistemology. Still, others may disagree, and so this point may be of service to them. With all of this said, it is now time to consider various specifications of authority.

§1.2.3—Authority as Indubitability[1]

We are presently seeking to specify the phenomenon of avowals' authority. We want to capture what a wide range of philosophers have been so adamant to explain, while being specific enough to avoid the charge that we are offering only "hackneyed phrases" that are easily defeated by counter-examples or being dismissed as philosophically unpuzzling (Barz 2018, p. 139).

Barz himself tries to meet this goal. He does so by homing in on two terms that are commonly deployed in descriptions of authority, namely, that avowals are authoritative because they (1) are *relatively indubitable* and (2) enjoy a *presumption of truth*. The attentive reader may

notice that these terms roughly track the three brief descriptions of authority that I offered in §1.1.1. The first two, which made claims about the 'deference' owed to avowals, or of the 'special weight' that they carry, might be understood as corresponding to the idea that avowals are owed a distinctive presumption of truth, whereas the latter is pretty clearly a claim about the immunity avowals have to doubt, at least given certain conditions. Barz considers the latter sort of specification first.

Very few philosophers contend that avowals are immune to *every possible* form of doubt. As we saw in §1.1.1, some say that avowals are immune to *rational* doubt, or that they "are not easily subjected to doubt" (Bar-On 2004, p. 3). These are our reasons for thinking that philosophers construe avowals as *relatively* indubitable. Drawing on suggestions such as these, Barz proposes the following specification of authority:

> **Indubitability$_1$**: Necessarily, if a person *S* tells another person *H* at time *t* that she (*S*) is currently in mental state ϕ, then *H* cannot rationally doubt at *t* that *S* is currently in ϕ—provided that *H* assumes that (i) *S* is sincere, (ii) *S* made no slip of the tongue, and (iii) *S* is not conceptually confused. (2018, p. 129)[8]

Indubitability$_1$ is framed in a way that involves *S*'s (a speaker) *telling H* (a hearer) that she is in some mental state. However, as per our discussion in §1.1.2, I want to caution us against understanding this in a way that is necessarily overtly social. That is, we should keep in mind that *S* can avow by writing down what counts as an avowal in her diary, or by simply avowing in thought. As such, we need to recognize that *H* can be someone who stumbles upon *S*'s diary, or who considers *S*'s silent avowals. So, even though I will not mess with Barz's template (I say 'template' because, as we will see, the other specifications of authority that he considers all share this framing), I want it to be understood going forward that, unless otherwise specified, our

[8] Barz cites Bar-On as a proponent, who writes that avowals are "not easily subject to doubt" (2004, p. 3). Perhaps a better example is Coliva, who argues that "[i]f you are sincere and competent with respect to the concepts you use to express your mental states, nobody can—rationally—cast any doubt on your avowals." (2016, p. 62)

discussions of authority should apply to avowals-in-thought as well as avowals encountered outside of overt speech acts in interpersonal communicative contexts.

Barz is willing to acknowledge that Indubitability$_1$ is at least prima facie philosophically puzzling, since *epistemologists* might reasonably find it strange that there could be *any* class of contingent, empirical claims that is ever immune to reasonable doubt given (i)-(iii). Nevertheless, Barz argues that it is false if taken to apply to avowals of propositional attitudes. This is because it seems that avowals of beliefs, desires, and other propositional attitudes can be rationally doubted even given (i)-(iii). His example of such a case is self-deception. For instance, I might avow a belief that my mother loves me in full sincerity, without conceptual confusion, and without slippage of the tongue, and yet be taken as saying something false because I am taken to be self-deceived. Thus, if Indubitability$_1$ is plausible at all, Barz thinks it best to restrict it to avowals of *conscious experiences* that are, qua conscious, not repressed.

Even when restricted in this way, however, Barz finds that Indubitability$_1$ is vulnerable to a counter-example. As an initiation trick, a person is told that she is going to be branded with a hot poker. She is then blindfolded and, soon after, an ice cube is pressed to her body. Upon feeling the ice cube, she (i) sincerely, (ii) without slippage of the tongue, and (iii) without conceptual confusion cries out "I am in pain!" Plausibly, however, her avowal can be reasonably doubted since, after all, one might reasonably doubt that the ice cube really caused *pain* as opposed to *cold*. Granted, the cold may *be* painful. However, what is at issue is whether *doubt* is *reasonable* here, not whether it is *veridical*.[9] Thus, even though she avows "I am in pain!" while meeting conditions (i)-(iii), her avowal can reasonably be doubted.[10]

[9] Barz adapts this case from Locke (1967). See also Bar-On for a similar case (2004, pp. 322-323).
[10] I do not consider cases like failing to authoritatively avow what one's conscious experiences are *caused by* or *represent* as counter-examples worth exploring (cf. Schwitzgebel 2008). Instead, I simply concede that we might lack authority about these matters.

Now, one might recognize this case as highly abnormal, because the subject issues her avowal under psychological duress—she issues it anxiously, while anticipating a painful event, all while deprived of a cherished perceptual modality. So it might be suggested that Indubitability$_1$ is underspecified and that, when it is better specified, the counter-example will be adequately addressed. Specifically, if we add that we must not take *S*'s avowal to be issued under conditions of psychological duress, it should come as no surprise that the initiate's avowal can reasonably be doubted, but only because it is not supposed to be immune to reasonable doubt.

Barz anticipates this sort of reply and admits that we could introduce "normality conditions" alongside (i)-(iii).[11] He does not say much about what these conditions are, and so I can only venture what I think is a fair guess, namely, that they are the conditions in which one enjoys overall (i.e., not superhuman) cognitive well-functioning, such that one is able to lucidly engage one's attentional resources. From here we get something like the following, where the italicized portion constitutes the added content, and the rest is the same as Indubitability$_1$:

> **Indubitability$_{1*}$**: Necessarily, if a person *S* tells another person *H* at time *t* that she (*S*) is currently in mental state ϕ, then *H* cannot rationally doubt at *t* that *S* is currently in ϕ—provided that *H* assumes that (i) *S* is sincere, (ii) *S* made no slip of the tongue, (iii) *S* is not conceptually confused, *and (iv) S issues her avowal under normal cognitive conditions*.[12]

Unfortunately, Barz takes Indubitability$_{1*}$ to succumb to a different counter-example, one which he draws from David Thompson (who, in turn, draws it from Daniel Dennett):

> The visual field is coloured, at least for those of us who are not colour-blind. It seems as if it is coloured from one periphery to the other. Yet if you stare straight ahead while someone moves a coloured playing card from directly behind you into your peripheral field, you are unable to say what colour it is. Try it! The card can be moved surprisingly close to the center of your visual field before you can name the colour. Since most of us know already that only the cones of the macula are sensitive to colour while the rods on the rest of the retina are not, we should not be surprised by this experience.

[11] Barz (2018, p. 130).
[12] To be absolutely clear, this is my formulation, not Barz's.

> Yet we are surprised! It seems to us that the whole visual field is coloured, though only the center of it really is. (2009, p. 26)

This passage invites the following counter-example to Indubitability$_{1*}$. You, having read the above, have knowledge of the fact that the periphery of one's visual field is not coloured. Now suppose that your friend believes that the whole of one's visual field *is* coloured. Finally, suppose that you place a red playing card in front of her and ask her about the colour at a point on the periphery of her vision. Finally, suppose that your friend answers: "the point on the periphery of my vision is red-coloured". She answers this on the basis of her knowledge of the redness of the card that is conferred by the *center* of her vision, and her reasonable yet false belief that her *whole* visual field is coloured. Alas, you—being in the know—can reasonably doubt her avowal. This is so even if we stipulate that you take her avowal to be issued in normal cognitive conditions. So Indubitability$_{1*}$ is false. We have yet to adequately specify authority, even for avowals of conscious experiences specifically.

§1.2.4—Authority as Indubitability$_2$

Barz concedes that your friend might issue a different avowal that is immune to counter-example, namely, that *she thinks* that the point on the periphery of her vision is red-coloured. Thus, we get Indubitability$_2$, which is exactly like Indubitability$_{1*}$, except with an additional qualifier, (v): *S's avowal is flanked by an 'I think that…' clause*.

Barz accepts Indubitability$_2$. This is because when an agent says that she is thinking that P, her avowal is made true by the very act of incorporating the content P into her avowal. Such avowals are self-verifying (as defined in §1.2.2) and so immune to reasonable doubt.[13] Unfortunately, despite this fact, Barz takes Indubitability$_2$ to violate two of his requirements for

[13] Burge (1988) discusses self-verifying avowals at length.

an adequate specification of authority. First, it is said to violate his requirement (1), because it applies only to a narrow range of avowals. I have already gone over why this is so in §1.2.2, but I have also already explained that this requirement is potentially suspect, since Barz has not ruled out the possibility of a pluralist approach to specifying authority. However, that point does not help with Indubitability$_2$'s violation of Barz's requirement (3), since the manner in which avowals are authoritative on this proposal is indeed quite trivial. Such avowals are authoritative because self-verifying, where self-verification is a simple result of the fact that thinking a thought means that your self-ascription of that thought is true. If this is not philosophically puzzling, then there is not much to be gained by sustained philosophical reflection on it.

§1.2.5—Authority as Presumptive Truth$_1$

In light of the above failures, Barz moves on to the second genre of possible specifications of authority, namely, those that focus on the *presumptive truth* of avowals. He begins with the following specification:

> **Presumption of Truth$_1$**: If (1) a person *S* tells another person *H* at time *t* that she (*S*) is currently in mental state ϕ and (2) *H* lacks any good reason to doubt what *S* says, then *H* is justified at *t* in believing that *S* is currently in ϕ—even if *H* possesses no positive evidence that *S* is currently in ϕ (aside from the fact that *S* tells him so). (2018, p. 131)

Now, since Barz's requirement (2) reminds us that *other*-ascriptions should *not* be understood as authoritative, he notes that this specification is incomplete until it is supplemented with a "complementary claim":

> **CC to Presumption of Truth$_1$**: By contrast, if (1) a person *P* tells another person *H* at time *t* that a third person *S* is currently in mental state ϕ and (2) *H* possesses no positive evidence that *S* is currently in ϕ (aside from the fact that *P* tells him so), then *H* is not justified at *t* in believing that *S* is currently in ϕ—even if *H* lacks any good reason to doubt what *P* says. (2018, p. 131)

Presumption of Truth_1 itself, which Barz attributes to Donald Davidson (1984) and McGeer (2007), is not disputed by him. However, he rejects CC to Presumption of Truth_1, and so takes the conjunction of Presumption of Truth_1 and its CC to violate his requirement (2).

His argument asks us to think about testimonial justification. It is intuitive to many philosophers that we are often justified in taking the testimony of strangers at face value even without possessing independent positive evidence of its truth. We often seem justified, for example, in accepting the testimony of a passerby at a train station who has answered our request for the time of an incoming train, even if we lack additional positive information about her trustworthiness.[14] Barz's worry is that there are analogous cases involving other-ascriptive testimony. Thus, suppose that I enter a party looking for a friend. I ask a stranger where she is, and the stranger replies that my friend left because she believes that staying too long at parties brings bad juju.[15] Suppose that I lack any evidence of the stranger's insincerity and have no evidence to contravene her claim. Many find it intuitive that I can presume the truth of her other-ascription. But I do so in exactly the sorts of conditions that entitle me to presume the truth of avowals. This defeats CC to Presumption of Truth_1, in violation of Barz's requirement (2).

One might respond by rejecting this 'liberal' conception of testimonial justification. Denying this conception in favour of a 'conservative' one, the thought goes, might allow us to say that testimonial justification *does* hinge on one's having independent positive evidence for its truth, countervailing intuitions be damned.[16] On this view, I am not justified in presuming the truth of the stranger's other-ascription, and so CC to Presumption of Truth_1 is not violated.

[14] Burge (1993), for instance, thinks that we are *a priori* entitled to accept any testimony that is intelligible and for which we lack defeating evidence.
[15] This is a less violent spin on Barz's case (2018, p. 132).
[16] See, e.g., Lackey (2008), who argues that we need additional inductive evidence about *S*'s trustworthiness.

Unfortunately, this reply is in tension with Presumption of Truth$_1$ rather than its CC. After all, because avowals are also testimonials about one's mental states,[17] accepting conservatism about testimonial justification will require us to say that we are justified in deferring to avowals only if we have independent positive evidence of their trustworthiness, which is precisely what Presumption of Truth$_1$ denies. So, on the one hand, defending Presumption of Truth$_1$ requires denying conservatism about testimonial justification while, on the other hand, defending CC to Presumption of Truth$_1$ requires accepting it. No matter which way we go, Barz concludes that we must reject this specification of authority.

§1.2.6—Authority as Presumptive Truth$_2$

Barz also tries his hand at a second specification of the presumption of truth idea, along with its CC:

> **Presumption of Truth$_2$**: If (1) a person *S* tells another person *H* at time *t* that she (S) is currently in mental state ϕ and (2) *H* lacks any good reason to doubt what *S* says, then, given normal circumstances, *H*'s willingness to question whether *S* really is in ϕ is rather low—even if *H* possesses no positive evidence that *S* is currently in ϕ (aside from the fact that *S* tells him so).
>
> **CC to Presumption of Truth$_2$**: By contrast, if (1) a person *P* tells another person *H* at time *t* that a third person *S* is currently in mental state ϕ and (2) *H* possesses no positive evidence that *S* is currently in ϕ (aside from the fact that P tells him so), then, given normal circumstances, *H's* willingness to question whether *S* really is in ϕ is not as low—even if *H* lacks any good reason to doubt what *P* says. (2018, p. 131)

[17] Barz briefly considers the view, sometimes attributed to Wittgenstein, that self-ascriptions are not testimonials because they merely express one's mental states and so have no truth-value (except where the expressed states are beliefs). But he notes that this view is incompatible with the idea that self-ascriptions are presumptively true. Others (e.g., Thompson 2012) think that expressive self-ascriptions are not testimonials because they don't have testimony as their intentional aim. However, I believe we can make sense of the expressive *and* truth-evaluable status of self-ascriptions (Jacobsen 1996, 1997, Falvey 2000, Bar-On 2004; see also Owens 2006 for more on how testimony can transmit knowledge by expressing one's mental states; there, he focuses only on belief-expression, but logical space is surely open to a more expansive view). Moreover, it is possible that an utterance can count as testimony even if this is not its intentional aim (Lackey 2006; Doyle 2015, p. 84, fn. 64).

Barz cites Bar-On and Douglas Long (2003, p. 181) as proponents of this specification of authority. This specification differs from the previous one because it does not claim an asymmetry concerning the *justification* we have to defer to avowals as opposed to other-ascriptions. Instead it claims that our *actual deferential practices* involve *greater willingness* to presume the truth of avowals than other-ascriptions.

Barz has two objections to this specification of authority. The first is that it may be empirically unfounded, since he has not encountered any data in support of it. I think this is unfair, because it is reasonable in this context to take our reactions to possible cases as relevant data. Consider, for example, the introductory remarks of David Finkelstein's (2003) *Expression and the Inner*. There, Finkelstein asks you to imagine your willingness to doubt your friend Max's self-ascription of an intention, relative to your willingness to doubt his wife's ascription of that same intention to Max. Finkelstein declares that, in such a case, 'it doesn't even occur to you to think that Max, rather than Sarah, might be mistaken…' (2003, p. 1). While it is often problematic to draw empirical-psychological generalizations from reactions to possible cases, this is a very special context, since the very empirical-psychological phenomenon in which we are interested has to do with our reactions to a kind of testimony. For Barz's objection to stick, he would have to argue that our reactions to cases like the above are unlikely to correspond to our reactions to actual similar cases. But it seems to me that there is no relevant difference between real and imagined cases here that would ground this skeptical suggestion.

It is Barz's second objection, however, that is more serious. The objection is that the above specification of authority violates his requirement (3): even if true, Presumption of Truth_2 and its CC would not be philosophically puzzling. This, he argues, is because our deferential practices can be explained by facts about "temperament, prejudice, bias, social pressure, and other factors

that are often beyond rational control" (2018, p. 136), all of which are ripe for empirical-psychological (and perhaps sociological) investigation rather than anything distinctively philosophical (where 'philosophical' is defined in terms of the metaphilosophical view described in §1.2.2). Succinctly put: "that the extent of our willingness to question a statement might depend on *who* utters it should not astonish us: it is perfectly compatible with existing knowledge about human psychology…there is no reason to consult philosophers" (2018, p. 136).

Granted, this objection merely tells against the philosophically puzzling status of our actual deferential practices regarding avowals, and so one could argue that the justificatory status of these practices might still be philosophically puzzling. Alas, Barz quickly reminds us that this normative rendering of authority is exactly what the conjunction of Presumption of Truth_1 and its CC tried and failed to produce.

Because Barz takes the above specifications of authority—$\text{Indubitability}_{1/1*}$, Indubitability_2, Presumption of Truth_1, and Presumption of Truth_2—to (nearly[18]) exhaust the options that might be charitably extracted from the literature, he concludes with his dilemma for proponents of authority:

> Either one arrives at a specification [of authority] that is philosophically interesting but false ($\text{Indubitability}_{1[/1*]}$, Presumption of Truth_1 including its CC), or one arrives at a specification that is true but of minor philosophical interest (Indubitability_2, Presumption of Truth_2 including its CC). (2018, p. 137)

Barz does not say that no viable specifications of authority could possibly be developed. Instead, he only concludes that his dilemma shifts the burden onto believers in authority to explain, in sufficient detail, what exactly the phenomenon is and why it is philosophically puzzling. In what follows I will try to meet this burden.

[18] Barz quickly considers two further options in passing, one of which he thinks changes the subject, and another which he thinks is sufficiently similar to Presumption of Truth_2 to be rejected without additional argument. I will turn to these specifications in §1.3.3.

§1.3.1—Avoiding Barz's Dilemma: Indubitability$_{1+}$

There are at least two strategies one could pursue in response to Barz's dilemma. First, one could attempt to steer any of the previously criticized specifications of authority around the horns of his dilemma. Second, one could offer novel specifications of authority that avoid his dilemma. In this section, after trying the first strategy, I will ultimately pursue the second.

Let us consider indubitability-based specifications of authority first, and let me quickly remind the reader of the path that we took to our current position. Initially we came up against the case of the person who, during an initiation ritual, issued a dubious self-ascription of pain when pressed with a cold ice cube, the result of which was that we were required to modify Indubitability$_1$ by introducing a "normality conditions" qualifier that could protect it against counter-examples involving cognitive duress. This yielded Indubitability$_{1*}$. Indubitability$_{1*}$ then faced a counter-example in which an agent in normal cognitive conditions self-ascribes a particular visual appearance that can be reasonably doubted. Barz then suggested that the best reply was to modify Indubitability$_{1*}$ into a thesis concerned strictly with avowals of the form "*I think that…*", thus yielding Indubitability$_2$. This led to the further objection that Indubitability$_2$ violates his requirement (1). Despite my misgivings with that objection, I conceded that Indubitability$_2$ may violate his requirement (3) as well: the self-verifying nature of such avowals is nearly trivial and, hence, neuters the philosophically puzzling status of their authority. For this reason, I will abandon the defense of Indubitability$_2$. Instead, I will focus on whether we can rehabilitate something closer to Indubitability$_{1*}$. To reiterate:

> **Indubitability$_{1*}$**: Necessarily, if a person *S* tells another person *H* at time *t* that she (*S*) is currently in mental state ϕ, then *H* cannot rationally doubt at *t* that *S* is currently in ϕ—provided that *H* assumes that (i) *S* is sincere, (ii) *S* made no slip of the tongue, (iii) *S* is not conceptually confused, and (iv) S issues her avowal under normal cognitive conditions.

I think that there is a way to deal with Barz's counter-example to Indubitability$_{1*}$ (the dubitable self-ascription of red in the corner of one's visual field). This will admittedly amount to a restriction on its scope. However, unlike Indubitability$_2$, it will not restrict its scope by narrowing the types of self-ascriptions to which it applies, and so Barz's requirement (1) is not obviously violated. Specifically, I think the right thing to do is to side with philosophers like Annalisa Coliva (2016, pp. 63-64) who argue that *sufficiently coarse-grained* avowals of conscious experiences are relatively indubitable.

Coliva does not discuss Thompson's case specifically. Instead, she discusses a case in which an agent self-ascribes her awareness of a number of dots in her visual field. Coliva concedes that the agent might simply *miscount* these dots, and so her avowal can reasonably be doubted even if it is made in normal conditions:

> …one can say that one's after image contains five red dots, when in fact it contains only four. However, authority can be maintained even in this case because determining the number of dots involves counting and one may go astray in doing it. Still, one would be authoritative with respect to the fact of having an after image. (2016, p. 64)

I suspect that, in addition to being able to say simply that "I have an afterimage", one can also self-ascribe *some* of its content in a way that is not open to reasonable doubt, given (i)-(iv), e.g., "my visual field includes multiple dots". What is being suggested is that reasonable doubt is harder to generate as the contents of one's avowals become coarser-grained. I now want to advance an analogous reply to the Thompson case: Barz has given us no reason to doubt that one's avowal of a *red visual impression* is indubitable given (i)-(iv) even if we concede that avowals of the *extent* or *location* of its redness can be reasonably doubted.[19]

[19] Notice that a case like colour-blindness is not a counter-example to this reply, since it is more charitable to reinterpret what the colour-blind agent means by 'red' than it is to take her to be making a false claim.

The reasonable idea behind these concessions, I take it, is that avowals about locational and quantitative visual properties are vulnerable to the same sorts of errors as empirical claims about similar properties of *objects in the world*. Crispin Wright puts the point nicely when he notes that, because one must hold a particular visual impression in mind over time in order to track certain of its quantitative and locational properties, this amounts to treating it as an object on the model of external perception.[20] But we are not authoritative about the features of objects in our external environment; nor, then, is it obvious that our explanandum must include authority over features of visual impressions to be (quasi-)perceptually scrutinized, at least where these are understood as sufficiently fine-grained. So, whatever we *do* end up saying about the range of authoritative avowals of conscious experiences, there are good reasons to be cautious here, and so to provide space for a view according to which Thompson-style avowals fall outside the scope of an adequately specified authority thesis. Note also that this qualification does not require us to restrict the proper domain of authoritative conscious-experience-avowals to those flanked by an '*I think that…*' clause, since avowals like "I am experiencing a red afterimage" contain no such clause. If all of this is right, then even though Indubitability$_{1*}$ is strictly speaking false, the following remains plausible:

> **Indubitability$_{1+}$**: Necessarily, if a person *S* tells another person *H* at time *t* that she (*S*) is currently in mental state Φ, then *H* cannot rationally doubt at *t* that *S* is currently in *Φ*—provided that *H* assumes that (i) *S* is sincere, (ii) *S* made no slip of the tongue, (iii) *S* is not conceptually confused, (iv) *S* issues her avowal under normal cognitive conditions, *and (v): S's avowal is sufficiently coarse-grained*.

Note that, while this proposal is intended primarily to rescue the authority of avowals of conscious experiences of visual impressions with intentional contents, it also applies readily to

[20] Wright's (2015, p. 65) point is made in a discussion of criticisms about first-person authority due to Snowdon (2012). See Zimmerman (2008, p. 344) for a similar discussion.

non-intentional experiences like pains and tickles. It applies on account of the fact that these experiences are maximally coarse-grained because altogether lacking in propositional content.[21]

As a final thought, let me address the concern that Barz's requirement (1) may take issue with Indubitability$_{1+}$'s restriction to conscious experience avowals. My response is the same as what I proposed in §1.2.2: Barz has given us no reason to think that Indubitability$_{1+}$ does not apply to a sufficiently large set of avowals, even though it may only apply to conscious experience avowals. That is because it applies to conscious experience avowals of many kinds, whereas Indubitability$_2$ only applies to self-ascriptions flanked by "*I think that…*" clauses.

§1.3.2—Avoiding Barz's Dilemma: Indubitability$_{\text{Brute-Error}}$

I have argued that there is a case to be made for the Indubitability$_{1+}$ of conscious experience avowals. Are there other avowals that are authoritative qua (relatively) indubitable? I believe so.

One might attempt to make a start here by noting, first, that avowals of propositional attitudes do not have visually perceptible or locational properties (coarse-grained or otherwise) about which one could go wrong in the first place.[22] In that event, Barz's counter-example (based on the Thompson case) would get no grip on the authority of such avowals. However, we should recall the reason for which Barz thought to restrict indubitability-based specifications of authority to conscious experiences in the first place, namely, that avowals of other types (at least, of propositional attitudes) are vulnerable to a different counter-example in the form of self-deception. Going back to Indubitability$_1$, recall his objection that avowals of propositional

[21] There is the possibility of a further counter-example as regards sensations, namely, Williamson's (2000) argument that we can doubt self-ascriptions of simple sensations when they are borderline determinate (e.g., at the specific time slice where cold begins to shade into warmth). To this, we might (once again following Coliva) add yet another qualifier, i.e., *(vi): S's self-ascription is taken to be of a sufficiently determinate state*. There is certainly more that could be said about this (and, in another context, we will return to Williamson's anti-luminosity argument in Chapter Five). For now, I hope that this suffices to adequately specify Indubitability$_{1+}$.
[22] Davidson (1987).

attitudes by an agent who is (i) sincere, (ii) makes no slip of the tongue, and (iii) is not conceptually confused can still be reasonably viewed as self-deceived.

But what about Indubitability$_{1+}$, which includes qualification (iv): S issues her avowal under normal cognitive conditions, as well as (v): S's avowal is sufficiently coarse-grained? Qualification (iv) might seem relevant, for if self-deception takes place in abnormal cognitive conditions, then (iv) accounts for it.

Surely, self-deception can cause problems for an agent's cognitive functioning, and yet it has been argued that self-deception often serves to *protect* the agent from other psychological harms.[23] The situation is sure to be complicated, and so simply stipulating that self-deception must be a mark of abnormal cognitive functioning is contentious. To see this, note that an agent might avow a belief across myriad cognitive contexts, and across long swaths of time, all while being self-deceived. For example, I might remain self-deceived about my real attitudes toward my mother for my whole life, even as I manage to navigate various cognitive contexts in reasonably good cognitive order. Moreover, even if these cognitive contexts index to different degrees of cognitive well-functioning, *some* of which might be reasonably viewed as abnormal, we cannot simply stipulate that self-deception *only* happens in the abnormal contexts. So it seems that Indubitability$_{1+}$ is not easily extended to avowals of propositional attitudes.

Recently, Coliva (2016, p. 65) has flagged three strategies by which one might address this concern. First, we might reconfigure Indubitability$_{1+}$ so that it includes an additional qualifier, i.e.: *the hearer must not take the speaker to be self-deceived*.[24] Second, we might grant that Indubitability$_{1+}$ can admit of exceptions, yet argue that it is "present in a significant amount of avowals…" (2016, p. 62). Finally, we might devise an account of *self-deception* that is

[23] For more about the complicated interplay between the harms and benefits of self-deception, see Bagnoli (2012).
[24] Wright (2001, p. 324).

compatible with Indubitability$_{1+}$ instead of modifying Indubitability$_{1+}$ to accommodate self-deception.

Like Coliva, I am not fond of the first strategy, since adding this qualifier to Indubitability$_{1+}$ seems unsettlingly ad hoc.[25] Now, while Coliva expresses some sympathy for the second strategy (which I will discuss below), she favours the third. Thus, she argues that cases of self-deception are cases in which an agent *really does* have the attitudes she self-ascribes. On her view, the irrationality of self-deception does not consist—as is often supposed—in the fact that she is somehow ignorant of her real attitudes. Instead, it consists in the fact that, unbeknownst to her, she *also* has countervailing attitudes.[26]

I am not concerned here with whether this is the correct account of self-deception (we will return to this question in Chapter Five). Rather, I am concerned with whether this account can rescue Indubitability$_{1+}$ or something similar. Coliva seems to think that it can. The problem, however, is that her account can at best explain why self-deception does not threaten the reliability of our *self-knowledge*. This is because Coliva's argument establishes, at most, that self-deception does not threaten the *truth* of our avowals of propositional attitudes. The problem we are considering, however, is different: we are considering whether *doubts* about such avowals are *reasonable*.[27] It might happen that a self-deceived agent's avowal is true and yet its hearer can, in judging the speaker to be self-deceived, reasonably take her avowal to be false.

Perhaps Coliva would argue that such doubts are unreasonable because they are underpinned by a mistaken account of self-deception, according to which self-deceived avowals are false. But rejecting (or simply not knowing of) Coliva's account of self-deception, while

[25] Coliva (2016, p. 65).
[26] Coliva (2016, chapter 7) develops this view. See also Bilgrami (2006a) and Parent (2017).
[27] See Coliva's (2019b) reply to Borgoni (2019), in which she applies her account of self-deception to a defense of an indubitability thesis. I take it that Coliva's defense is as inadequate there as it is here.

adopting another, need not automatically make one unreasonable here. Moreover, even if hearers *are* aware of and accept her account, they might also have good reason to embrace a pluralistic view of self-deception, such that some but not all cases where an agent's avowals are taken to be self-deceived are dubious from the hearer's point of view.

This leaves us with the second strategy, which actually requires denying Indubitability$_{1+}$. It requires this because it amounts to denying the *necessity* of the indubitability of propositional attitude avowals, given (i)-(v). In its place, the second strategy advocates for a weaker view, i.e., that we understand such avowals as *generally* indubitable, given (i)-(iv). The idea, then, is that we cannot view agents as generally self-deceived, perhaps because such agents would be inconceivably irrational. We now arrive at the following specification of authority:

> **Indubitability$_{\textbf{prop-att.}}$**: *Generally*, if a person *S* tells another person *H* at time *t* that she is currently in mental state $\Phi_{prop\text{-}att.}$, then *H* cannot rationally doubt at *t* that *S* is currently in $\Phi_{prop\text{-}att.}$—provided that *H* assumes that (i) *S* is sincere, (ii) *S* made no slip of the tongue, (iii) *S* is not conceptually confused, and (iv) *S* issues her avowal under normal cognitive conditions.[28]

I accept this thesis, for it seems to me that there really are limits on how often we can reasonably doubt avowals of beliefs, desires, and intentions before it is no longer reasonable to attribute agency to the speaker or thinker. Moreover, for what it is worth, this view has some historical pedigree. Wright, for instance, claims that "*[w]holesale* suspicion about my attitudinal avowals—where it is not a doubt about sincerity or understanding—jars with conceiving of me as an intentional subject at all" (2001, pp. 324-325, emphasis mine), while Coliva claims that "if one were *systematically* proven wrong in one's psychological avowals, doubt would be cast upon one's possession of the relevant concepts" (2016, p. 66, emphasis mine).

[28] The abbreviation $_{prop.\text{-}att.}$ flags the fact that this thesis is concerned with the indubitability of avowals of propositional attitudes specifically. Thanks to Wolfgang Barz for pressing me to clarify this.

Even still, Indubitability$_{\text{prop-att.}}$ is not promising as a specification of *authority*, since the insulation of such avowals from *systematic* suspicion does not entail that one's avowals will be treated as *especially* indubitable (Bettcher 2009, p. 100). To see this, note that a similar thesis about perceptual reports also has significant historical pedigree and independent plausibility. I have in mind the Davidsonian thesis that interpreting the perceptual reports of others is governed by a principle of charity, such that we cannot treat them as systematically mistaken.[29] If this is right, and if we want to call avowals authoritative *in virtue of* their Indubitability$_{\text{prop-att.}}$, then we will have to concede that perceptual reports are also authoritative in a similar sense.[30] But we are trying to isolate what makes avowals *uniquely* authoritative.

One might try to dispute this putative parallel between perceptual reports and avowals. For example, consider a case Wright uses to dispute it. He writes that "I may have such poor colour vision that you rightly come to distrust my testimony on matters of colour" (2001, p. 325) and that, *unlike* avowals, wholesale suspicion about my colour-reports does not jar with taking me to be an agent. Wright takes this to indicate that perceptual reports are not quite like avowals in respect of being immune to systematic doubt. However, it may be that systematic errors on matters of colour are intelligible only because such reports comprise a sufficiently narrow kind of perceptual report. And this, it will be said, is why they can be systematically doubted without placing too much doubt on the agent's general perceptual capacities. But now the thought is that we have room for an analogy with avowals: perhaps I can be seen to go systematically wrong in

[29] This is one reading of the lesson of Davidson's reflections on radical interpretation (1973). On another reading, what the interpreter must maximize in his conception of the interpretee is not correctness, but intelligibility. Still, I suspect that there is some degree of important correlation between intelligibility and correctness—agents who we take to be increasingly wrong about their perceptual environment will also be increasingly unintelligible to us.
[30] In effect, this is a violation of Barz's requirement (2), except that the requirement does not explicitly mention that avowals should be distinguished from empirical claims other than other-ascriptions.

my avowals of, e.g., *beliefs about my mother*, while still being taken as an agent, but only because such avowals comprise a highly narrow kind of avowal.

It may be possible to undermine the parallel in some other way, perhaps by showing that colour-perception reports and avowals of beliefs about my mother differ importantly in their narrowness. If so, then $\text{Indubitability}_{\text{prop-att.}}$ might be defensible as a specification of authority. At this point, however, I will simply set it aside. This is because I believe there is another indubitability thesis available that better distinguishes avowals from perceptual reports (with respect to authority), and that can principally accommodate self-deception. I have in mind the idea that avowals, unlike perceptual reports, are uniquely *immune to brute error*. Here, to get us started, is Bar-On:

> …although some perceptual reports and some non-mental self-reports can exhibit some first-person/third-person asymmetries, they are all open to what may be called brute error—an error that is simply due to the world failing to cooperate, rather than being due to some kind of failure of the subject's conceptual, perceptual, or psychological faculties. (2010, p. 2)

Bar-On follows Tyler Burge, who describes brute errors as those which "depend on the independence of physical objects' natures from how we conceive or perceive them, and on the contingency of our causal relations to them" (1988, p. 657). Put differently, the thought is that our relationship to our own psychology is one in which we cannot go wrong about our mental states simply because they trick us into misidentifying them.[31] There are no mirages or "ringers"[32] for one's mental states, as the story goes. Instead, mistakes in avowals seem to

[31] Bar-On (2004, p. 183).
[32] Doyle (2019).

always be attributable to the speaker or her psychology.[33] Erroneous avowals are sourced in problems with the agent's mind, not in any brute interference in the mind-world interface.

Plausibly, self-deception is one such case: one is not self-deceived because one is *duped* by a mental *appearance* into mis-ascribing an attitude. Rather, one engages in motivated reasoning, or represses one's true attitude, and so the fault is ultimately with oneself. Now, this point concerns the sorts of errors to which one's *self-beliefs*—one's beliefs about one's attitudes—are immune, and so it may be complained that we have now simply changed the subject from a peculiar testimonial phenomenon to its first-person epistemology. But I believe that we can reconstrue such talk as follows. When somebody issues a perceptual report, one might reasonably wonder whether the person has seen things aright, no matter how closely they are attending to their environment, simply because the environment may have failed to cooperate with the agent's cognitive faculties. But when a speaker is taken to mis-ascribe a propositional attitude, the hearer must take the mistake to be psychological in nature. Thus, unlike Indubitability$_{1+}$, reasonable doubt about propositional attitude avowals is not ruled out. This seems to be exactly what we want, since self-deception does not seem to be a case of brute error.

This brings us to what we can call Indubitability$_{\text{brute-error}}$:

> **Indubitability**$_{\text{brute-error:}}$ Necessarily, if a person *S* tells another person *H* at time *t* that she (*S*) is currently in mental state ϕ, then *H* cannot rationally doubt at *t* that *S* is currently in ϕ—provided that *H* assumes that (i) *S* is sincere, (ii) *S* made no slip of the tongue, (iii) *S* is not conceptually confused, (iv) *S* issues her avowal under normal cognitive conditions**,** ***and (v) S has not succumbed to some psychological failing***.

If Indubitability$_{\text{brute-error}}$ is true, then we have a novel indubitability thesis which handles self-deception in a principled way while preserving a relevant difference between avowals of

[33] Burge sometimes writes as if avowals only enjoy *near*-immunity to brute error, i.e., that they *"normally* seem to involve some malfunction" (1996, p. 104, emphasis mine). I do not see what sort of exceptions Burge could have had in mind. Accordingly, I follow Bar-On (2004) in taking brute errors of self-ascription as impossibilities.

propositional attitudes and ordinary perceptual reports. In fact, it is a thesis that I take to be true of *all* avowals. Moreover, such a thesis seems philosophically puzzling: epistemologists should naturally wonder why avowals exhibit immunity to brute error while no other kind of empirical claims seems to exhibit such immunity.

§1.3.3—Avoiding Barz's Dilemma: Presumption of Truth$_3$

In the next several subsections I will advance various versions of the thesis that avowals are authoritative in virtue of their presumptive truth.

I begin with the thought that there seems to be a difference in the *degree* to which we are justified in presuming the truth of avowals as opposed to their other-ascriptive counterparts (and other empirical claims), at least in good avowing conditions. This suggestion is dialectically novel in light of the fact that Presumption of Truth$_1$ and its CC ignore *degrees* of justification (see §1.2.5), while presumption of Truth$_2$ and its CC ignore justification altogether (see §1.2.6).

Interestingly, near the end of his paper, Barz briefly considers the proposal that '*S's self-ascriptions are (or are treated as) more reliable, or more strongly justified, than anyone else's ascriptions of mental states to S*' (2018, p. 138). In fact, there are two proposals here. First, there is the proposal without parentheses: that *S*'s beliefs about her own current mental states are more strongly justified than anyone else's beliefs about them. But Barz reminds us that this is not our subject: authority has to do with the epistemic status of *H*'s justification to defer to *S*, rather than the epistemic quality of *S*'s avowal. Second, there is the proposal with parentheses: that avowals are *treated as* more reliable or justified than other-ascriptions. Here, Barz replies that, like Presumption of Truth$_2$ and its CC, this proposal describes "a behavioral pattern that is perfectly compatible with existing knowledge about human psychology" (2018, p. 138) and so is philosophically unpuzzling.

Notice, however, that Barz appears to have missed another proposal:

> **Presumption of Truth$_3$**: If (1) a person *S* tells another person *H* at time *t* that she (*S*) is currently in mental state ϕ and (2) *H* lacks any good reason to doubt what *S* says, then *H* is *more justified* at *t* in believing that *S* is currently in ϕ *than H would be in believing an ascription of* ϕ *to S by some other person, P*—even if *H* possesses no positive evidence that *S* is currently in ϕ (aside from the fact that *S* tells him so).

The italicized portion of this thesis is crucial, first, because it highlights a contrast with other-ascriptions and so automatically equips Presumption of Truth$_3$ with a CC and, second, because it distinguishes Presumption of Truth$_3$ from the two proposals just considered. It distinguishes Presumption of Truth$_3$ from the first proposal because it refers to the justification that *H* has for presuming the truth of an avowal rather than to epistemic status of *S*'s avowal, and it distinguishes Presumption of Truth$_3$ from the second proposal because it does not describe differences in how *H* actually happens to *treat* avowals compared to other-ascriptions.

Presumption of Truth$_3$ may also track extant proposals. Thus, going back to Bar-On's remark that avowals seem to 'carry so much more weight' (2004, p. 10) than other-ascriptions of the same mental states, she never remarks that other-ascriptions necessarily carry *no* weight (at least when they are presumed sincere and so on). Likewise, I believe there is a perfectly charitable normative reading of the above passage from Finkelstein (§1.2.6) on which one driving intuition is that we are more justified in presuming the truth of Max's avowal than his wife's other-ascription. Finally, consider Matthew Parrott's claim that '…no matter how much epistemic expertise someone else has on my attitudes, another person's assertions about them cannot be entitled to deference to the same degree as my own' (2015, pp. 2220). None of these authors need to be read as claiming that *H* ought *never* to presume the truth of *P*'s ascription of ϕ to *S*, given the right background conditions. It suffices to read them as claiming that any such presumption of *H*'s will be less justified than a presumption made toward *S*'s avowal of ϕ.

Presumption of Truth_3 also seems philosophically puzzling. For, despite whatever purely empirical psychological explanation we may provide for what *motivates* hearers to presume the truth of other avowals more readily than they presume the truth of other-ascriptions, it is surely *epistemologically* puzzling if we are especially *justified* in presuming the truth of avowals.

Nevertheless, an objection to Presumption of Truth_3 has recently been put to me.[34] It begins with the banal observation that neither *S*'s avowal nor *P*'s other-ascription provides *H* with *maximal* justification to believe that *S* is in ϕ. What is needed for *H*'s belief to be maximally justified is further physiological, neurological, and non-linguistic behavioural evidence about *S*'s mental state. Indeed, *just the same* evidence will be needed for *H* to reach maximal justification whether *H*'s *initial* justification for her belief about *S*'s mental state is conferred by *S*'s avowal or, instead, by *P*'s other-ascription. The problem now is that, since the additional evidence needed is the same in both cases, it seems that *S*'s avowal does *not* justify *H*'s belief more than *P*'s other-ascription. For, if it did, then it would not be the case that *H* would need the same evidence to reach maximal justification in both cases. Instead, *less* evidence would be needed in the avowing case than in the other-ascription case.

To my mind, this objection makes a crucial and unmotivated presupposition: if the same *quantity* and *kinds* of evidence will be needed to maximally justify *H*'s belief about *S*'s mental state (whether *H* starts from *S*'s avowal or from *P*'s other-ascription), then this evidence must also have the exact same justificatory force for *H* in both cases. For, if the evidence does not have the same justificatory force for *H* in both cases, then it might happen that all the same evidence is acquired by *H* in either case, such that *H* becomes maximally justified in believing that *S* is in ϕ, but that the additional evidence (same in quantity and kind) adds less justification

[34] Thanks to Wolfgang Barz.

to *H*'s belief when *H* begins from *S*'s avowal than it does when *H* begins from *P*'s other-ascription. I think that we should reject the presupposition, and hence to allow for differences in the justificatory force conferred by the same body of evidence given different contextual factors.

Since this is a rather abstract response, consider one possible explanation of it (we will take up this sort of view more seriously in Chapter Two). On this explanation, avowals are acts that *express* the avowed mental state, whereas other-ascriptions merely *report that* another person is in a mental state. After all, you cannot express *my* mental states, and so your other-ascriptions can at most express *your views about* my mental states. Now, in expressing my mental state in the avowing act, one view is that I *show you* my mind (Bar-On 2000, 2004). The idea is not that my mental state has a visual appearance in the way that colors or cookies do; rather, we can say that you perceptually encounter my mental state *by hearing my avowal*. At any rate, however we fill in the details, the warrant you have to believe me is especially strong: indeed, you have a factive reason to believe that I am in ϕ, whereas *P*'s other-ascription does not provide you with a factive reason. Now my contention is that the acquisition of additional evidence about my mental state will not count for much if you begin from my avowal of ϕ: if you already have a factive reason to believe that I am in ϕ, coming to possess additional justification in the form of (e.g.) neurological evidence will have a diminished epistemic value for you, given the strength of the justification you already have. Inversely, if your belief that I am in ϕ is initially based on *P*'s other-ascription of ϕ to me, this neurological evidence will not have diminished epistemic value for you. Behind this is a simple suggestion: the justificatory value of some piece of evidence can vary depending on how much one needs it. I am not absolutely convinced that this reply to the objection is right, but I do hope to have shifted the burden back to the skeptic.[35]

[35] One final note is worth making here, which is that Presumption of Truth_3 leaves open the possibility that different types of avowals justify *H*'s belief about *S*'s mind to different degrees. For example, *H* may be more justified in

§1.3.4—An Interlude Regarding the Groundlessness of Avowals

In §1.3.5-1.3.6 I will develop two more specifications of authority. According to both specifications, the presumptive truth of avowals is philosophically puzzling in light of the fact that a further thesis about avowals is also true. In this subsection, I discuss this further thesis.

Consider this passage from Wright:

> The demand that somebody produces reasons or corroborating evidence for [avowals]…is always inappropriate. There is nothing they might reasonably be expected to be able to say. In that sense, there is nothing upon which such claims are based. (2001, p. 321).

In other words, while it is often appropriate to ask people to give reasons for their empirical claims, it seems that it is not appropriate to ask people to give epistemic reasons for their avowals, nor can avowers reasonably meet such demands. Asked, for example, how you know that you believe that the sky is blue, the best you will be able to say is "I just do!". This is hardly to offer an epistemic reason for your avowal, and yet it seems that the question rather than the answer is problematic.

I will abbreviate Wright's thesis as follows, where (as usual) *S* is an avowing speaker and *H* is a hearer (or, more neutrally, 'receiver') of *S*'s avowal:

> **Groundlessness**: there is no epistemic reason that *S* can reasonably provide for her avowals if requested to do so by *H* (such that any request from *H* is inappropriate)

Several points about Groundlessness are in order. First, I understand Groundlessness as a thesis about the *epistemic* inappropriateness of such requests—their being epistemically out of place as opposed to being, say, prudentially or morally out of place. Moreover, because this is a *normative* epistemological thesis, it is distinct from a superficially similar thesis due to Bar-On:

deferring to phenomenal avowals than intentional ones (cf. Wright 2001), even though both sorts of avowals justify *H*'s belief about *S*'s mind more than any of *P*'s corresponding other-ascriptions.

> Avowals are not normally subjected to ordinary epistemic assessment…it would be highly irregular to request of someone issuing an avowal to offer reasons for her pronouncement (2004, p. 3).

This is because Bar-On's thesis concerns a *descriptive* phenomenon, i.e., the *infrequency* with which avowals are epistemically assessed, and the *irregularity* with which hearers ask speakers to epistemically justify their avowals. I believe that Bar-On's thesis is interesting in its own right, but I also suspect that it is *symptomatic* of what I am calling Groundlessness: if avowals are indeed irregularly subject to requests for epistemic support, this plausibly follows from our general, shared sensitivity to the inappropriateness of soliciting such support.

Second, Groundlessness is not a thesis about the nature of the self-knowledge a speaker may have of the mental states she avows. This means that the inappropriateness of asking *S* to produce epistemic reasons for her avowals is compatible with the possibility that she has self-knowledge of the avowed mental state, and that this self-knowledge has some sort of basis, albeit one that is beyond her ken.[36]

Third, like authority, Groundlessness is intended to apply to first-personal self-ascriptions—i.e., genuine avowals—only. Thus, it is not about the *impossibility* of providing epistemic reasons for self-ascriptions more generally. To see what I am getting at, consider the following case from Parrott:

> Suppose I tell you that I believe someone in my department is out to get me and that I think this (about myself) because (i) my long-time therapist has diagnosed me as paranoid and (ii) while at work, I have noticed myself furtively watching colleagues. How should you respond to my avowal of this belief? Should you trust that I do, in fact, believe someone in the department is out to get me? Should you defer to what I say about my beliefs over what my therapist says about them? The evidence I appeal to seems good enough; indeed it is evidence that you might have become aware of independently of anything I said. But, in this case, it seems more natural for one to be hesitant and not take what I say to be true. (2015, p. 2219)

[36] For example, their avowals may be grounded in the reliable deliverances of an 'inner-scanner' (cf. Armstrong 1968), even if one cannot knowingly cite this mechanism as the source of her self-knowledge. I consider this view in more detail in Chapters Three (see especially §3.2) and Chapter Five.

This is an example of a third-personal self-ascription; a case where one cites the sort of evidence about oneself that others might cite in ascribing a mental state to you. Parrott notices that no deference is owed in this case, because the perspective from which one produces one's avowal is too much like that of another person's perspective on you to count as authoritative. Groundlessness captures this observation: one cannot be *reasonably* expected to provide such evidence, not because it is impossible to cite any, but because it is epistemically counterproductive to do so when one's self-ascription is a genuine avowal.

Fourth, although Groundlessness claims that *S* cannot reasonably provide epistemic reasons for her avowals (and that *H* cannot appropriately request them), such that the question "how do you know?" is an inappropriate question to issue toward an avowal, there are still epistemic challenges that *S* might be appropriately confronted with. For example, as Casey Doyle points out, *S* might be challenged as being "insincere, confused, or deluded" in her avowal (2019, p. 354).[37] All of this is consistent with the various qualifiers that most of our authority theses have so far taken on board. Perhaps there are also limits on the extent to which such challenges are applicable to avowals, but I do not discuss these possibilities here.

A fifth and final point will take us through to the end of this subsection. To begin, note that Groundlessness has a sweeping scope, since it denies the appropriateness of requests for epistemic reasons for *any* kind of (first-personal) avowal. Now, one might think that this is misguided. For, as Quassim Cassam argues, there do seem to be epistemic reasons that we can reasonably offer for *phenomenal* avowals, such as avowals of tickles or pains.[38] For example, one may claim to know that one is in pain *by feeling it* (2009, pp. 10-11). This is a potentially

[37] Thanks to an anonymous reviewer for the *Journal of Philosophical Research* for prompting this point.
[38] *Pace* Hampshire, who claims that: "he who reports that he is currently experiencing a certain sensation cannot intelligibly be asked how he knows it" (1979, pp. 282–283).

important point even if it does not generalize to avowals of propositional attitudes or other mental states that lack a proprietary phenomenology, or that cannot be avowed simply on the basis of whatever phenomenological salience they do have.[39] We can even agree that one's feeling pain, say, does not count as *evidence* for one's avowal of pain, much in the way that seeing a burrito on one's table, being a factive reason for believing as much, means that seeing this is not mere evidence for one's belief. The point is that citing what one feels, whether this is evidence or not for one's avowal, may serve as a perfectly legitimate epistemic reason to provide for it. Accordingly, *H*'s request for such a reason need not be inappropriate.

In reply, some philosophers have insisted that this sort of reason is no reason at all. This is because they think that "being in pain and feeling pain are one and the same thing" (Shoemaker 1994, p. 128), such that explaining one's avowal of being in pain by citing one's feeling of pain amounts to repeating oneself. However, Cassam notes that feeling pain is not equivalent to *feeling that* one is in pain, since the latter but not the former requires one to possess and apply the concept of pain. For this reason, he concludes that there is indeed sufficient "ontological distance" (2009, p. 11) between being in pain and *feeling that* one is in pain for a speaker to cite feeling that she is in pain as her epistemic reason for avowing pain.[40] *Mutatis mutandis*, so the story may go, for other phenomenal avowals.

Perhaps Cassam's appeal to *feeling that* one is in pain is innocuous enough as a way of highlighting the difference between being in pain and conceptualizing one's pain, such that there is surely *something* that happens in the transition from feeling pain to avowing one's pain. But I

[39] See, e.g., Bar-On (2004, p. 104) who argues that the *contents* of avowals are not phenomenally accessible (*pace* Pitt 2004), even if their *being believed* is. Note, moreover, that even if we agree with Pitt that *occurrent* beliefs have a proprietary phenomenology (this being the phenomenology of conscious *judging*), phenomenal introspection of standing beliefs and other standing attitudes is another matter entirely (Doyle 2019, p. 355). See Arango–Muñoz (2019) for more recent skepticism about cognitive phenomenology.
[40] Cassam's idea of epistemic feeling is inspired by Dretske's (1973) idea of epistemic seeing.

am not confident that there is a real objection to Groundlessness here. This is because what actually goes on in feeling that one is in pain is, for everything Cassam has said, totally opaque. It is (by definition) not just a matter of feeling pain, but for this reason "I feel that I am in pain" reads like a placeholder for whatever it is that puts us in a position to avow pain, and so sounds like no real answer when provided in response to a request to supply an epistemic ground for one's avowal of pain. It is not even as informative as saying you see that p as your reason for declaring p. After all, in that case you are at least saying something about the sensory modality that has provided you with knowledge, even if you know very little else about the deeper architecture of visual perception. On the other hand, if 'feeling that' is being suggested as a *sui generis* sensory modality for enabling us to avow our pains, we should surely need to hear more about this. As a final point, I doubt that many laypersons have a grip on the feeling that/feeling distinction, and so it is unlikely that ordinary hearers can reasonably be expected to provide 'I feel that I am in pain' as an epistemic ground for their avowals of pain.

But even if we concede that phenomenal avowals are counter-examples to Groundlessness, the thesis still seems to hold for non-phenomenal avowals: if the presumptive answer to how one knows one's pain is one's feeling that one is in pain, this seems unfit as an answer to how one knows one's beliefs, hopes, intentions, and desires, and there do not seem to be any other answers available from the first-person point of view for how one knows that one is in these mental states. Thus, one might simply concede the objection and argue that it won't apply to a specification of authority that exploits Groundlessness, so long as that specification applies only to non-phenomenal avowals.

At this point, the question will be whether a specification of authority that depends on Groundlessness is bound to violate Barz's requirement (1), since any such specification will

apply only to a limited range of authoritative avowals. I already cast some provisional doubt on that requirement in §1.2.2, precisely by questioning the monistic conception of explanations of first-person authority that seems to underlie it. But we can also strengthen the case against requirement (1) by way of two observations. First, there is some empirical evidence that different cognitive systems govern our processing of phenomenal and non-phenomenal mental state ascriptions (Knobe & Prinz 2008; Gray et al. 2011), and some have taken this as evidence that different sorts of deference might be owed to each.[41] Second, pluralism about the sources of self-knowledge is gaining traction in nearby discussions about the so-called "privileged access" we have to our mental states (Boghossian 1989).[42] If those discussions have any implications for discussions of authority (seeing as privileged access may explain authority), then we might reasonably expect different specifications of authority to cover different sorts of avowals. This reply would surely be unsatisfying if we thought that there is no way to adequately specify authority for phenomenal avowals. But this is not our predicament, at least if you agree with me that Presumption of Truth_3 applies to all authoritative avowals (see §1.3.3).

In fact, I will not offer a pluralistic explanation of authority in Chapter Two, so the present reply does not help me specifically. Still, I believe that my first reply to Cassam is sufficient. Readers should take the second reply as a possible fallback position in the event that my preferred explanation of authority fails.

I have spilled a lot of ink clarifying and defending Groundlessness. But I have not yet shown how it combines with any claims about the presumptive truth of avowals to generate plausible specifications of authority. Now it is time to do so.

[41] I owe my awareness of these references to Parrott (2015, p. 2217).
[42] See especially Boyle (2009), Coliva (2016), and Komorowska-Mach (2019). See Byrne (2018, pp. 157-158) for resistance.

§1.3.5—Avoiding Barz's Dilemma: Groundless-Authority$_1$

Consider the following specification of authority (note that one may wish to restrict the scope of this specification to non-phenomenal avowals, depending on one's thoughts about §1.3.4):

> **Groundless-Authority$_1$**: If (1) a person *S* tells another person *H* at time *t* that she (*S*) is currently in some mental state ϕ and (2) *H* lacks any good reason to doubt what *S* says, then *H* is justified at *t* in believing that *S* is currently in ϕ even if *H* possesses no positive evidence that *S* is currently in ϕ (aside from the fact that *S* tells *H* so), *and even though there is no epistemic reason S can reasonably provide for her avowal if requested to do so by H (such that any request from H is inappropriate)*

> **CC to Groundless-Authority$_1$**: By contrast, if (1) a person *P* tells another person *H* at time *t* that a third person *S* is currently in some mental state ϕ, and (2) *H* possesses no positive evidence that *S* is currently in ϕ (aside from the fact that *P* tells *H* so), then—even if *H* lacks any good reason to doubt what *P* says, *and even if H is justified at t in believing that S is currently in ϕ—there must be epistemic reasons S can reasonably provide for her avowal if requested to do so by H (such that any request from H is appropriate)*

Groundless-Authority$_1$ itself is simply Presumption of Truth$_1$ plus Groundlessness (the italicized portion). In other words, the idea is that *H's* justification to presume the truth of *S*'s avowals obtains, absent reason for doubt and without additional positive evidence, even when *S* is not reasonably able or expected to say anything in defense of her avowals. Contrariwise, CC to Groundless-Authority$_1$ is really quite different from CC to Presumption of Truth$_1$. There are two senses in which it is different. First, the italicized *even if* clause makes room for the possibility that *H* can justifiably presume the truth of other-ascriptions in the right circumstances (though, precisely because it is an *even if* clause, it doesn't *entail* this). Second, it *denies* Groundlessness for other-ascriptions. So the basic asymmetry is between presumptively true, *Groundless* avowals and (potentially) presumptively true, *Grounded* other-ascriptions.

To get a better sense for this specification, let us return to the scenario in which *H* enters a party and sees her friend *S* leaving. Next a stranger, *P*, explains that *S* believes that staying too long at parties brings bad juju. Let us say that *H* has no positive evidence that *P* is right (beyond

her say so). Let us also say that *H* takes *P*'s other-ascription to be sincere and all the rest. CC to Groundless-Authority$_1$ grants that *H* may be justified in believing *P* (contra CC to Presumption of Truth$_1$). But now imagine that, out of simple curiosity or distrust, *H* asks *P* how *P* knows this about *S*. My claim is that *P* should be able to offer an explanation: perhaps she heard *S* loudly proclaim that staying too late at parties brings bad juju. Moreover, and crucially: in the event that *P* cannot produce any such justification, *H*'s justification to believe *P* will be undercut, for *H* should now begin to wonder how *P* knows what she claims to know about *S*.

Now consider an alternative scenario in which *H* enters a party and sees *S* leaving. *H* asks her why she is leaving and *S* tells *H* that she believes that staying too long at parties brings bad juju. Groundless-Authority$_1$ claims that there is nothing she can reasonably say in response to the question "how do you know you believe that?" In this sense, the question is inappropriate. In all likelihood, *S* would balk at the question. Nevertheless, *H* is justified in believing what *S* says about her belief, so long as *H* takes her to be sincere and all the rest. *S*'s inability to reasonably provide an answer is not a reason to doubt what she says, despite *H*'s lacking any further positive evidence of *S*'s belief beyond her say-so. Her avowal is, thus, authoritative in the sense of being justifiably presumed true despite its Groundlessness, whereas *P*'s other-ascription is not authoritative because, even if it can be justifiably presumed true, it is not also insulated from reasonable requests for epistemic support.

Now, for all that Groundless-Authority$_1$ and its CC claim, it can happen that *H* is *equally* justified in deferring to *S* and *P* (bracketing Presumption of Truth$_3$). This is true despite the fact that only other-ascriptions can actually be backed by epistemic reasons (at least in the non-phenomenal case). How could this be? How can we be just as justified in presuming the truth of avowals as we are in presuming the truth of other-ascriptions, even though in the latter case the

agent (*P*) can ideally meet demands to justify herself, whereas in the former case she (*S*) cannot? Here resides an epistemological and hence philosophical puzzle: empirical claims for which epistemic support can be provided are generally (taken to be) epistemically superior to claims for which no such support can be provided, and yet avowals are an exception.

§1.3.6—Avoiding Barz's Dilemma: Groundless-Authority$_2$

Now consider one final specification of authority (note again that one may wish to restrict the scope of this specification to non-phenomenal avowals, depending on one's thoughts about §1.3.4):

> **Groundless-Authority$_2$**: If (1) a person *S* tells another person *H* at time *t* that she (*S*) is currently in some mental state ϕ and (2) *H* lacks any good reason to doubt what *S* says, then, given normal circumstances, *H*'s willingness to question whether *S* really is in ϕ is rather low, *and even though there is no epistemic reason S can reasonably provide for her avowal if requested to do so by H (such that any request from H is inappropriate)*
>
> **CC to Groundless-Authority$_2$**: By contrast, if (1) a person *P* tells another person *H* at time *t* that a third person *S* is currently in mental state ϕ and (2) *H* possesses no positive evidence that *S* is currently in ϕ (aside from the fact that *P* tells him so), then, given normal circumstances, *H*'s willingness to question whether *S* really is in ϕ is not as low—even if *H* lacks any good reason to doubt what *P* says—*and even though there are epistemic reasons S can reasonably provide for her avowal if requested to do so by H (such that any request from H is appropriate)*

Groundless-Authority$_2$ is simply Presumption of Truth$_2$ plus Groundlessness (the italicized portion). What is new, in other words, is that *H's* willingness to question *S*'s avowals is rather low in normal circumstances, absent reason for doubt and without additional positive evidence, even when S is not reasonably expected to say anything in defense of her avowals. Contrariwise, CC to Groundless-Authority$_2$ tells us that *H*'s willingness to question *P*'s other-ascription will be higher in normal circumstances (1) even absent reason for doubt, (2) even though it would be

entirely appropriate for *P* to epistemically support her other-ascription and, I should add, (3) even if *P* does actually provide such support.

Recall that Barz's main objection to Presumption of Truth_2 and its CC was that they are not philosophically puzzling, since they are explicable along empirical-psychological lines and are not epistemologically mysterious. I argue that no such objection follows for Groundless-Authority_2 and its CC. First, it is epistemologically puzzling that *H* is more likely to defer to *S* than to *P* in normal circumstances, despite the fact that, in other areas of discourse, we tend to be more willing to defer to the person who we think is better able to epistemically support her claims. For, even though this is itself a fact about the psychological dispositions of *H*, and so counts as an empirical rather than normative phenomenon, an epistemological point remains salient: epistemologists should puzzle over the fact that *H*'s belief-forming practices diverge from a basic epistemic norm—one that tells us to place greater trust in the person who can epistemically support her claims—when *H* is confronted with avowals.

Perhaps we will eventually explain this disposition without recourse to any distinctively philosophical problem-solving methodology. But we should recall a point that I made in §2: the mere fact that some phenomenon is puzzling for philosophical (here, epistemological) reasons does not obviously entail, as Barz presupposes, that it must be solved using distinctively philosophical resources. For, while we may eventually acquire a full empirical psychological explanation of the puzzle behind *H*'s disposition to defer more readily to Groundless avowals than non-Groundless other-ascriptions, this will not show that the puzzle did not have a distinctively epistemological dimension to it in the first place, since it will have been perfectly

reasonable for epistemologists to wonder why hearers are so readily disposed to presume the truth of avowals despite their Groundlessness.[43]

§1.3.7—A Point of Speculation

Barz himself is aware that Groundlessness is a popular thesis about avowals. Why, then, did he fail to consider either variant of Groundless-Authority and their respective CCs? One possible diagnosis is that he focused too narrowly, in developing his arguments, on discussions of authority in which the groundlessness of avowals is not explicitly noted. No doubt, it is often possible to find passages where philosophers seem to be interested *merely* in the apparent fact that avowals 'carry so much more weight' than their other-ascriptive counterparts (Bar-On 2004, p. 10). Passages like this can be read as suggesting that Groundlessness is irrelevant to characterizing the puzzle(s) of authority. Moreover, Groundlessness is logically independent of any claims about the presumptive truth of avowals, and so it may seem obvious that we should treat these properties of avowals in theoretical isolation.

Nevertheless, when Davidson (1984) observes that evidentially unsupported claims do not in general carry more weight than evidentially well-supported ones (or even the same weight, I add), he is trying to highlight the fact that this observation admits of a puzzling exception in the case of avowals. Likewise, as we saw, Parrott (2015) puzzles over the fact that *expertise* seems to be irrelevant to the deference we cede to avowals.[44] As I read these and other authors, they are at least partly puzzling over the fact that we (justifiably or especially readily) *presume the truth of Groundless avowals*.

[43] My view, to be articulated in Chapter Two, is that we can use distinctively philosophical argumentation to address Groundless-AT_2 and its CC. My point is only that others, who disagree with my preferred explanation, might take advantage of this thought.
[44] See also Ayer (1963).

§1.4.1—Finalizing Our Explanandum

Barz's (2018) paper is titled "Is There Anything to the Authority Thesis?" I have answered in the affirmative. In fact, there is something to *several* such authority *theses*, for there are multiple plausible and philosophically puzzling specifications of authority. So that the reader can view each of my specifications of authority in one place, the remainder of this subsection compiles them, with some brief reminders about their intended scope where appropriate.

First, we have:

> **Indubitability$_{1+}$**: Necessarily, if a person *S* tells another person *H* at time *t* that she (*S*) is currently in mental state Φ, then *H* cannot rationally doubt at *t* that *S* is currently in *Φ*—provided that *H* assumes that (i) *S* is sincere, (ii) *S* made no slip of the tongue, (iii) *S* is not conceptually confused, (iv) *S* issues her avowal under normal cognitive conditions, and (v): *S*'s avowal is sufficiently coarse-grained.

Indubitability$_{1+}$ is intended to range over avowals of conscious experiences. This extends to conscious experiences with intentional contents (where qualification (v) is doing its most important work) and conscious experiences that lack intentional contents altogether, such as purely sensory states.

Next, we have:

> **Indubitability$_{\text{brute-error}}$**: Necessarily, if a person *S* tells another person *H* at time *t* that she (S) is currently in mental state ϕ, then *H* cannot rationally doubt at *t* that *S* is currently in ϕ—provided that *H* assumes that (i) *S* is sincere, (ii) *S* made no slip of the tongue, (iii) *S* is not conceptually confused, (iv) *S* issues her avowal under normal conditions of cognitive well-functioning, and (v) *S* has not succumbed to some psychological failing.

Indubitability$_{\text{brute-error}}$ is, I submit, true of avowals generally.

The first presumption-of-truth-based specification on which we settled was:

> **Presumption of Truth$_3$**: If (1) a person *S* tells another person *H* at time *t* that she (*S*) is currently in mental state ϕ and (2) *H* lacks any good reason to doubt what *S* says, then *H* is *more justified* at *t* in believing that *S* is currently in ϕ than *H* would be in believing an ascription of ϕ to *S* by some other person, *P*—even if *H* possesses no positive evidence that *S* is currently in ϕ (aside from the fact that *S* tells him so).

My view is that this specification of authority ranges over avowals of all sorts. Recipients of avowals have, in general, a superior epistemic warrant for presuming their truth than for presuming the truth of other-ascriptions.

The first of our two 'Groundless-Authority' theses is:

> **Groundless-Authority$_1$**: If (1) a person *S* tells another person *H* at time *t* that she (*S*) is currently in some mental state ϕ and (2) *H* lacks any good reason to doubt what *S* says, then *H* is justified at *t* in believing that *S* is currently in ϕ even if *H* possesses no positive evidence that *S* is currently in ϕ (aside from the fact that *S* tells *H* so), and even though there is no epistemic reason *S* can reasonably provide for her avowal if requested to do so by *H* (such that any request from *H* is inappropriate)
>
> **CC to Groundless-Authority$_1$**: By contrast, if (1) a person *P* tells another person *H* at time *t* that a third person *S* is currently in some mental state ϕ, and (2) *H* possesses no positive evidence that *S* is currently in ϕ (aside from the fact that *P* tells *H* so), then—even if *H* lacks any good reason to doubt what *P* says, and even if *H* is justified at t in believing that *S* is currently in ϕ—there must be epistemic reasons *S* can reasonably provide for her avowal if requested to do so by *H* (such that any request from *H* is appropriate)

I believe that this specification applies to avowals generally, though we have seen that some might disagree about whether it applies to phenomenal avowals.

Finally, we have:

> **Groundless-Authority$_2$**: If (1) a person *S* tells another person *H* at time *t* that she (*S*) is currently in some mental state ϕ and (2) *H* lacks any good reason to doubt what *S* says, then, given normal circumstances, *H*'s willingness to question whether *S* really is in ϕ is rather low, and even though there is no epistemic reason *S* can reasonably provide for her avowal if requested to do so by *H* (such that any request from *H* is inappropriate)
>
> **CC to Groundless-Authority$_2$**: By contrast, if (1) a person *P* tells another person *H* at time *t* that a third person *S* is currently in mental state ϕ and (2) *H* possesses no positive evidence that *S* is currently in ϕ (aside from the fact that *P* tells him so), then, given normal circumstances, *H*'s willingness to question whether *S* really is in ϕ is not as low—even if *H* lacks any good reason to doubt what *P* says—and even though there are epistemic reasons *S* can reasonably provide for her avowal if requested to do so by *H* (such that any request from *H* is appropriate)

So much by way of reiteration. In the final subsection of this chapter, I briefly discuss two different broad methodological strategies for explaining authority.

§1.4.2—Setting Up the Discourse—Self-Knowledge and Avowals

A few times throughout this chapter I have distinguished between avowals and our *self-knowledge* of the mental states we avow. What can we say about the relationship between avowals and self-knowledge?

In fact, it is overwhelmingly common for philosophers to try and explain the authority of avowals in terms of self-knowledge. But not just *any* self-knowledge: what philosophers typically appeal to is a form of self-knowledge that is "privileged" (Boghossian 1989) and "peculiar" (Gertler 2011a). Self-knowledge is privileged when it is more reliable than any other agent's knowledge of your mind, and it is peculiar when it is acquired in a way that is available to no one else. Thus, the puzzle of why the recipients of avowals tend to (or do) confer authority (of whatever sort) on avowals is to be explained as a consequence of the fact that avowers have a special (and especially reliable) way of knowing the mental states they avow. We can call this *the epistemic strategy* for explaining authority, and we can call an account of authority based on this strategy an *epistemic account of authority*.

While the epistemic strategy is by far the most popular strategy for explaining authority, it has been argued by some philosophers that it is not the only strategy. To take just two examples (that I do not favour personally): one might argue that it is part of the "grammar" of mental discourse that we treat avowals as authoritative, whether or not avowers have privileged and peculiar self-knowledge. Alternatively, one might argue that there are properties of avowals themselves—understood as thoughts or speech-acts—that warrant hearers in taking avowals as authoritative. In Chapter Two I will be pursuing a version of this second non-epistemic strategy.

In pursuing this strategy, I will not be denying that we have privileged and peculiar self-knowledge, and that this is something that may need to be explained over and above the authority of avowals. In other words, I do not think that explaining authority along non-epistemic lines explains everything there is to explain about the uniqueness of our relationship to our own minds. Indeed, the possibility of privileged and peculiar self-knowledge (as well as its function, nature, and origins) will be a major focus of the final three chapters. Accordingly, I do not accept the following view that Lukas Schwengerer attributes to some proponents of the non-epistemic strategy:

> **Linguistic View**: The peculiar nature of self-knowledge should be described exclusively by features of linguistic practice, syntax, semantics, and pragmatics. (2019, p. 2)

I reject Linguistic View because I take it that authority—understood as a cluster of properties of avowals qua semantically structured acts and artifacts—is not identical to the phenomenon of privileged and peculiar self-knowledge, and so I resist views like (Linguistic View) which would purport to reduce the privilege and peculiarity of self-knowledge to the authority of avowals. The authority that agents accord to avowals is one thing, and the privileged and peculiar status of an agent's self-knowledge is another. Moreover, even though both may call out for explanation, it need not be that the one must explain the other. Or so I shall eventually argue.

My pursuit and defense of a non-epistemic strategy for explaining authority in Chapter Two will not be original in the main. My main contribution will be to address several recent objections to it. In defending my preferred explanation of authority, we will occasionally encounter questions about privileged and peculiar self-knowledge as it relates to avowals. But the real task of thinking more about such self-knowledge will take us through the remaining three chapters, where I will try to say something about how my account of authority fits together

with my preferred accounts of the functions, nature, and origins of such knowledge. My eventual account will be that our capacity to *express* our first-order mental states by avowing them, coupled with a capacity for privileged and peculiar self-knowledge of the very mental states we avow, are jointly necessary conditions on our capacities for certain forms of social-epistemic agency. The end result, I hope, will be a contribution to our understanding of why it is that we are social-epistemic agents who can avow authoritatively as well as know the mental states we avow with privilege and peculiarity.

Chapter Two—Expressivist Authority

§2.1.1—Introduction

In this chapter I will defend a non-epistemic account of authority in its various specifications. This, as mentioned at the end of Chapter One, is an account that does not seek to explain authority in terms of the avower's *self-knowledge* of the mental states she avows, where self-knowledge is understood as a matter of possessing warranted, true *second-order* beliefs about one's mind (self-beliefs). Instead, the *expressivist* explanation that I will defend here makes the following core claim: the authority of avowals resides in the fact that we ourselves, but not others, are able *reveal* our *first-order* mental states to other people through our avowals. Of course, this claim requires clarification that will be provided in due course: what it is to express a mental state, what is required of an agent in order to be able to do so, what it is for an expression to reveal that which is expressed, and how exactly expressing a mental state through an avowal secures its authority, are matters of detail that vary across particular versions of expressivism.

In §2.2 I will begin with a brief overview of a view that I call 'Traditional Expressivism'. While we will see in §2.3 that Traditional Expressivism has been subject to powerful criticisms, it will be shown that key elements of it can be rehabilitated and improved upon. This will lead us to the sort of expressivist view that I favour—one that I will call 'Neo-Expressivism', following Dorit Bar-On. After articulating Neo-Expressivism in §2.4, I will show in §2.5 how it can explain the various specifications of authority that we settled on in Chapter One. In §2.6 I will turn to some further explanatory applications of Neo-Expressivism. In §2.7 I will address several objections to Neo-Expressivism. As we will see, taking up the last of these objections will serve as a launchpad the last three chapters of this dissertation.

§2.2.1—Traditional Expressivism Expounded

Famously, Ludwig Wittgenstein opposed the idea of a *private language*. A private language is, roughly, a language that is understandable in principle only by one person. For example, suppose I experience a certain sensation and attach a certain term to it—call it 'S_1'. Suppose I continue to do this for my subsequent sensations $S_2 \ldots S_n$, until I have built up a stock of referential terms that constitute the core of my language. I do this by attaching terms to my own conscious experiences. It might seem as though a private language has hereby been constructed: I have generated a language via a technique of inner ostension, a technique that nobody else can use to establish terms that refer to my conscious experiences. According to Wittgenstein, the problem is that these terms in my language would lack objective conditions of correct application, and hence would not really have fixed meanings at all. And so it follows that this imaginative exercise does not deliver a coherent result after all.

Why not? Essentially, the concern is that there could be nothing beyond my say-so, beyond my take on whether I had applied the terms correctly, to ensure that I had in fact applied them correctly; there could be no *agent-external standard* by which I could check whether my terms were successfully applied. If the correct use of my terms $S_1 \ldots S_n$ is beholden only to my take on whether they correspond to the right experiential episodes, then no standard besides whatever standard I decide can assure their proper use. But this is just another way of saying that they lack objective conditions of correct application: it is only my say-so, and nothing else, that assures their proper use. And this is what exposes the incoherence of a private language: if $S_1 \ldots S_n$ mean whatever I like, then they in fact mean nothing, because there are no objective conditions that govern their use and so no meaning that can be specified in terms of how they are used.

Like Wittgenstein, I have focused on experiential (sensory) terms in this brief exposition of his private language argument, but his thought seems to be that we should draw a similar lesson about any and all terms in a language: if a putative method of meaning-determination is in principle only available to one person, then the putative speaker of a language comprised of such terms would once again be beholden to nothing but her own whims in governing their use, and this would show once again that such terms are not really meaningful—no such language or speaker is really possible.

Let us continue to focus on mental terms, since we are ultimately interested in thinking about avowals. Here, Wittgenstein's dismissal of the possibility of a private language leaves us with an important question. After all, if we cannot understand sensation language (or any mental language) as set up via routines of inner ostension, then we need some other way to understand its formation. It is at this stage that we can introduce a key passage for expositing Wittgenstein's expressivism:

> There doesn't seem to be any problem here; don't we talk about sensations every day, and give them names? But how is the connection between the name and the thing named set up? This question is the same as: how does a human being learn the meaning of the names of sensations?—of the word 'pain', for example. Here is one possibility: words are connected with the primitive, the natural, expressions of sensation and used in their place. A child has hurt himself and he cries: then adults talk to him and teach him exclamations and, later, sentences. They teach the child new pain-behaviour. (1953, §244)

The basic idea that has been traditionally extracted from passages like this is that avowals are learned analogues of non-linguistic expressive behaviours such as wincing, crying, and clapping, but also non-self-ascriptive linguistic behaviour like utterances of "ouch!" and "yay!". As children grow and learn, these more primitive expressions are replaced by "I'm in pain!", "I am sad!", and "it tastes great!". In this way, we learn to avow what was heretofore expressible by non-self-ascriptive means. The story of our psycholinguistic development is such that we are

effectively taught to *express* our states of mind by avowing them instead of or in addition to whatever other sorts of hard-wired or less sophisticated expressions we otherwise would or could express our mental states through.

Later in the same passage Wittgenstein adds that "…the verbal expression of pain replaces crying and does not describe it." Elsewhere, he makes similar suggestions:

> The words 'I am happy' are a bit of the behaviour of joy. (1953, §450[45])
>
> If he says 'I believe he's coming'…then he is acting, he is speaking, according to that condition, not reporting that it is to be found in him. (1953, §832)
>
> When someone says "I hope he'll come"—is this a *report* about his state of mind, or a *manifestation* of his hope?—I can, for example, say it to myself. And surely I am not giving myself a report. (1953 §585)

Claims such as these have been the source of much controversy in Wittgenstein scholarship. The traditional interpretation is this: the lacuna left by the private language argument in our understanding of mental talk is not bridged by invoking some special introspective mechanism which one might use to set up referential relations between, e.g., one's pains and their referring terms. One simply expresses one's pain directly by means of an avowing form of expressive behavior that one has been taught to use. In this way, "I am in pain!" serves as a way of expressing one's pain, *rather than* as a way of *describing* or *reporting* a pain that one has *discovered* in one. Of course, while reports also express mental states, what they express are beliefs *about* what one is reporting *on*. Reports are not direct expressions of the very beliefs, desires, fears, or pains they are about.

Now, in supposing that Wittgenstein was advancing the strictest of analogies between non-linguistic and linguistic expressive behaviour, the following claim has also frequently been

[45] *Remarks on the Philosophy of Psychology*, Volume I. Sections correspond to the 4th edition of *Philosophical Investigations*, edited by Hacker & Schulte (see Wittgenstein, L. 1953 in bibliography).

treated as part of Wittgenstein's views: inasmuch as a wince expresses pain but is not itself truth-evaluable, so too are avowals like "I am in pain" *mere* expressions of pain rather than truth-evaluable claims. This is one way of reading Wittgenstein's claim that expressions are not reports: they are not reports, because they do not make claims about states of affairs. How, then, can they be *true* or *false* (cf. Malcolm 1954, 1958; Rosenthal 1993[46])?

This reading is supported by Wittgenstein's description of avowals as *replacing* (rather than, say, *supplementing*) ontogenetically prior forms of expressive behaviour. Granted, avowals can and do carry Gricean 'natural' meanings—an avowal that one is in pain "means" that one is in pain by *revealing* or *manifesting* pain, just as bolts of lightning can "mean" that rain is coming and smiles can "mean" that someone is happy. But this does not suffice for truth-evaluability. Rather, it suffices for a kind of non-truth-evaluable signification. The lightning is not false if the rain never comes, because lightning is not truth-evaluable. Nor, on this reading of Wittgenstein, is one's avowal false if one is not actually in a particular state of mind.[47]

As I have been rehearsing a traditional reading of Wittgenstein as an expressivist about avowals, let us call this 'Traditional Expressivism'. Traditional Expressivism consists, roughly, of these core claims:

(1) Avowals are behaviour that express one's first-order mental states themselves

(2) Insofar as avowals express first-order mental states, they are just like non-linguistic behaviours that express first-order mental states (cries, winces, hugs)

(3) In being just like non-linguistic expressive behaviours, avowals are not truth-evaluable

[46] See also Wright (1998), who does not think that Wittgenstein is an expressivist but does understand the expressivist reading of him this way.

[47] García Rodríguez argues that self-ascriptions are necessarily "claims or judgements about oneself" (2019, p. 141) and so do not express the very mental states they semantically represent. This, I think, is an entirely unnecessary way of framing things: it follows only if one defines self-ascriptions as reports.

(4) Nevertheless, because avowals express mental states, they can signify them (in a broadly Gricean natural meaning sense of 'signify')

§2.3.1—A Problem for Traditional Expressivism as an Explanation of Authority

If, as Traditional Expressivism maintains, avowals are behaviours that express mental states in the way that winces and cries do, how does this explain authority in any of its permutations? Generally, the story is as follows: they *show H* that *S* is in a given mental state, and this is enough for *H* to justifiably defer to *S*'s avowal, and to explain why *H* does in fact defer, or for *S*'s avowal to be relatively indubitable by *H*'s lights. Traditional Expressivism also accounts for the groundlessness of avowals (§1.3.4): the fact that there is no epistemic reason that *S* can reasonably provide for her avowals if requested to do so by *H* (such that any request from *H* is inappropriate). For, as Wright points out, "if the avowal 'I'm in pain' is not a statement, true or false, then naturally it is inappropriate to ask its author for grounds for it" (1998, p. 35).

But does this sketch of an account of authority really speak to any of our specifications of it? Here is the most glaring question: how can avowals be authoritative in virtue of enjoying a distinctive *presumption of truth* if they are not so much as truth-evaluable? Perhaps the problem does not carry over to indubitability-based specifications of avowals, if only for the trivial reason that avowals—being non-truth-evaluable—cannot reasonably be *doubted* as *false*. But if this is indeed a way of explaining the relative indubitability of avowals, it does not extend any further than these specifications of authority. And insofar as it does not, it seems that Traditional Expressivism is hopeless to offer a more complete account of authority. Indeed, it is not even amenable to *combination* with some other account, precisely because it conceives of avowals as non-truth-evaluable and so cannot be combined with any explanation of authority that treats avowals as truth-evaluable.

§2.3.2—Traditional Expressivism as Radically Logico-Semantically Revisionary

Should we accept that avowals are not truth-evaluable, thereby denying that they are owed a presumption of truth? Unfortunately, it is widely agreed that the costs of this strategy are massive. A major problem is that the view destroys any semantic continuity between avowals and other utterances, including self-ascriptions that are not avowals.[48] For example, while avowals like "I am in pain" have no truth-conditions according to Traditional Expressivism, past-tense self-ascriptions that don't express mental states (because those states are no longer there to be expressed) *are* truth-evaluable. Suddenly, then, it seems as if our claims about our mental lives can be assessed for veridicality only when they are not made in the avowing mode. This is plainly bizarre, if nothing else.[49]

Relatedly, consider what the Traditional Expressivist must say about avowals in wider semantic and logical contexts. For example, when a speaker says, in an avowing mode, "I am not hungry, but Penny is hungry," Traditional Expressivists must take this utterance's truth-conditions to be given entirely by the latter conjunct. But this is, in effect, to deny that the utterance really *is* conjunctive, and hence to deny that avowals can be embedded in truth-functional compounds. Similarly, an inference like "if I am hungry, and if Penny is hungry, then John is hungry", where *I am hungry* is avowed, has only one antecedent rather than two. This radically alters the semantics of the conditional and any inference that might take its form. All of this is radically revisionary and goes to show just how unpalatable the consequences are if we are to try to hold on to the Traditional Expressivist view.

[48] See, for an early instance of this objection, Geach (1971).
[49] Hacker (1986, p. 90; 1990, p. 303-304) and Wright (1992, pp. 19, 29) have attempted to work around these issues. But see Jacobsen (1996) for what I take to be convincing responses.

§2.4.1—Neo-Expressivism Expounded

Traditional Expressivism faces severe objections, but what does this show? To my mind, and to the minds of some other philosophers, this does not show that Wittgenstein's basic insights are all for nothing. Rather, it shows that we need a revised picture, perhaps one that strays from the letter of Wittgenstein's own thinking, or perhaps one that charitably reinterprets him. Here I will explicate a "Neo-Expressivist" view that improves upon Traditional Expressivism in several respects. The central way in which Neo-Expressivism improves upon Traditional Expressivism is in its semantic treatment of avowals, in that it enables us to preserve their truth-evaluability. While my development of this view will take bits and pieces from multiple expressivists, not all of whom call themselves Neo-Expressivists, I will follow Dorit Bar-On in referring to each of these authors as Neo-Expressivists (and to the general picture I will develop as a Neo-Expressivist one) in virtue of their sharing a core body of claims that I will set out below.

Let us begin by carving out these core claims. First, Neo-Expressivists follow Wittgenstein in suggesting a contrast between expressive and reportive speech. Thus, Bar-On claims that "avowals, like natural expressions, serve to express the self-ascribed mental states" rather than report them (2004, p. 241), and Rockney Jacobsen agrees with Wittgenstein that "self-ascriptions express rather than report or describe the psychological states indicated by their main verb" (1996, p. 14). Finally, focusing on avowals of beliefs specifically, Kevin Falvey writes that:

> The notion of reporting is cognate with the notion of observation: reporters are people who are sent off to far-flung places to observe what is happening and report back to us on their findings…[Contrariwise], [p]rovided the avowal of a belief is sincere, which may be assumed to be the norm in interpersonal communication, the speaker is giving direct expression to, making manifest, what he believes. (2000, p. 77)

As I have said, Neo-Expressivists also argue that avowals are truth-evaluable. Thus, a central task for these philosophers is to address the *cognitivist* position that "if self-ascriptions

are…truth-evaluable, then they are assertions and express beliefs, and so they are not…expressions of the very states indicated by their psychological terms" (Jacobsen 1996, p. 19). The basic motivation for cognitivism is intuitive. We do not evaluate pains, desires, intentions, and tickles as true or false, but we do evaluate beliefs as such. So, if avowals express pains, desires, intentions, and tickles themselves rather than beliefs about these states, how are they truth-evaluable? Cognitivists sidestep this problem by rejecting expressivist views.

Fortunately, Jacobsen thinks that Wittgenstein himself flags a better strategy in the following passage (and for this reason, he does not think that Wittgenstein is actually a Traditional Expressivist, contra the prevailing consensus):

> One may have the thought: 'How remarkable that the *single* meaning of the word 'to feel' (and of the other psychological verbs) is compounded of the heterogeneous components, the meanings of the *first* and *third* person.' But what can be more different than the profile and the front view of a face; and yet the concepts of our language are so formed, that the one appears merely as a variation of the other. (*RPP* I §45)

Jacobsen takes Wittgenstein to be saying that "the Janus-faced nature of our psychological concepts consists in just this fact: that they combine semantic univocality with both expressive and non-expressive employments" (1997b, p. 19). This is supposed to signal a way out from the cognitivist's rejoinder, but how exactly?

For Jacobsen, we can see how a concept might have both expressive and non-expressive employments, alongside "semantic univocality", by embracing a minimalist conception of truth. According to this conception, a truth-evaluable sentence is any sentence that can be meaningfully fitted into a disquotational rule ("P" is true ≡ P) and that has the syntactic structure needed to be embedded in wider semantic and logical contexts such as negations, conditionals, and so on. Jacobsen believes that avowals have this sort of structure, and therefore that psychological

concepts can figure into truth-evaluable avowals. Avowals do not express second-order beliefs *about* the mental states indicated by their psychological verbs, but they are truth-evaluable in virtue of their syntactic structures and availability for meaningful disquotation.

To better see how this works as a reply to the cognitivist, consider the following cognitivist argument:

(1) Avowals are suitable substitutions for '*p*' in the disquotational rule ("*p*" is true ≡ *p*), so;
(2) Avowals are truth-evaluable, so;
(3) Avowals are assertions – i.e., expressions of second-order beliefs – that avowers have the first-order mental states they self-ascribe, so;
(4) Avowals are not expressions of the first-order mental states indicated by their psychological terms[50]

In this argument, cognitivism (3) is supposed to be a natural upshot of (1)-(2), both of which Jacobsen accepts. It is supposed to be a natural upshot of (1)-(2) given that we think of truth-evaluable utterances in general as assertions, and assertions as expressions of beliefs about their content clauses. Step (4), the falsity of expressivism, follows as a result.

With the minimalist conception of truth in play, however, Jacobsen argues that the move from (1)-(2) to (3) is a bad one. For, as Jacobsen notes, "no party to the debate [about minimalism] supposes that truth-evaluability *suffices* for assertoric status," and this is what one would need to secure the move at issue. For, as we have just seen, one part of minimalism about truth is that sentences are truth-evaluable if they can be meaningfully embedded in wider semantic contexts, such as being the antecedent of a conditional. And yet, conditional sentences like "if I eat pie, I will feel sick" do not express *assertions* as their antecedents. Nothing is asserted even though the antecedent is a declarative sentence. Nor do sentences "uttered on stage

[50] This is adapted, with cosmetic adjustments, from Jacobsen (1996, p. 23)

in the recital of a script, or uttered during the practice of elocution" (1996, pp. 23-24) express assertions. Nevertheless, they are truth-evaluable: they are true or false, whether or not they are put forward in an assertoric spirit.

To be sure, we can agree that we *typically* take utterances to be assertions when we take them to be truth-evaluable. But this only requires us to accept what Jacobsen calls:

> **Default Assumption**: "an utterance of a truth-evaluable sentence meaning that *p* is an act of asserting that *p*, *unless* otherwise explicitly divested of assertoric force by etiolating features of context." (1996, p. 24)

Default Assumption enables the following rejoinder to cognitivism: because even cognitivists can admit of cases where declarative utterances are divested of assertoric force, there is no principled reason to deny that *another* such case is the *expressive* one. Thus, "a truth-evaluable sentence can be uttered for a variety of expressive—i.e., non-assertoric—purposes, with the same meaning and same truth-value it would have had in a descriptive use" (1996, p. 25). Minimalism tells us how this could be so. Therefore, Neo-Expressivists "can in principle deny that our typical self-ascriptions have assertoric character on the purely mundane ground that the utterance *is already fully employed* in the business of expressing, say, a desire, a sensation or a feeling" (1996, p. 24).

The picture we now have is this. When I *avow*, I think or speak a sentence like "I believe that *p*" or "I desire to ϕ". *Qua avowal*, my utterance simply expresses the avowed state without making an assertion *about* that state (which would express a second-order belief about that state). As such, avowing is not a way of merely making a report about a mental state. It is a way of expressing that mental state itself, and is truth-evaluable despite not expressing a belief about my state (given minimalism). Indeed, one can see that this is just an application of the Fregean distinction between force and meaning: my utterance can at once have the force of expressing the

very first-order state it is about and, at the semantic level, *be about* that state, even if my utterance is not an assertion about that state and so does not express a second-order belief about that state. In just the way that we might ask, of someone's utterance "I am going to the library tomorrow" whether they are *predicting* their behaviour or *promising* to behave that way, while holding fixed the *meaning* of that sentence and its potential truth or falsity, we can do the same when we encounter self-ascriptions and wonder whether they are reports about or expressions of the state self-ascribed.[51]

A somewhat different route to the same result is traversed by Bar-On. She too tends to think of avowals as non-assertoric, "if we understand assertion as a specific kind of speech-act, with a relatively well-defined point or purpose and felicity conditions, on a par with making a request, issuing a command, asking a question." If this is how we must understand assertions, "then we may insist that at least some acts of expressing one's feeling, thought, etc. in language are not acts of making an assertion, issuing a statement, or delivering a report" (2004, p. 247). Perhaps this is clearest when we think about avowals made in thought, as when you enter a messy room and think to yourself "I'm disgusted" with no audience in mind. It is unintuitive to think of this as an assertion, in any standard sense of that term, though it is a paradigmatic (silent) avowal.[52]

---

[51] As Jacobsen (1997, p. 135) points out, promises are themselves a challenge for cognitivism.

[52] Alston (1967, p. 16) argues that "I can express my enthusiasm for your plan just as well by saying 'I'm very enthusiastic about your plan' as I can by saying 'What a tremendous plan'" and takes this to show that "expressing and asserting are not mutually exclusive in the way commonly supposed." Finkelstein (2003, p. 95) agrees, and takes this to be the key to reconciling truth-evaluability with expressibility. What I have drawn from Jacobsen and Bar-On is that, *whether we think of avowals as assertions or not*, they can be construed as truth-evaluable. Disagreements among Finkelstein and other Neo-Expressivists on this point likely comes down to the idiosyncratic ways in which they each define assertion. But we can now see that there is no need to be hung up on this dispute. For the record, I agree with Parrott (2010, p. 38) that we can follow Williams in saying that "if a speaker comes out with a declarative sentence not as part of a larger sentence (as one might say, by itself) and there are no special circumstances, then he is taken to have asserted what is meant by that sentence" (2002, p. 74); see also Roessler (2015, en. 7). So understood, many expressive utterances and thoughts can be counted as assertions, and what is rejected above as an assertoric conception of avowals is a highly specific way of understanding assertion.

Though Bar-On rejects the view that avowals are assertions (understood in a certain way), her strategy for rescuing the truth-evaluability of avowals is not to push for minimalism about truth. Instead, she appeals to a threefold distinction between kinds of expression, imparted to us by Wilfrid Sellars (1963):

> **EXP**$_1$ the *action* sense: a *person* expresses a state of hers by intentionally doing something.
>
> **EXP**$_2$ the *causal* sense: an *utterance* or piece of behavior expresses an underlying state by being the culmination of a causal process beginning with that state.
>
> **EXP**$_3$ the *semantic* sense: e.g., a *sentence* expresses an abstract proposition, thought, or judgment by being a (conventional) representation of it. (2004, p. 216)

Utilizing this distinction, Bar-On says more than Jacobsen does about how avowals are (dis)continuous with non-linguistic expressive behaviour. Specifically, she highlights the fact that avowals are continuous with much of our non-linguistic expressive behaviour in *expressing*$_1$ and *expressing*$_2$ mental states, even though they are discontinuous with such behaviour in *expressing*$_3$ those states. Moreover, she notes that avowals are, qua expressive$_1$, akin to hugs that express joy and non-avowing utterances and thoughts like "that's gross!" that express disgust, while being (in this way) unlike cries and winces that simply overcome one.

Most crucial, for Bar-On, are the action-expression (EXP$_1$) and semantic-expression (EXP$_3$) prongs of Sellars's distinction, at least when thinking about avowals. While both Bar-On and the Traditional Expressivist agree that we express our mental states through acts of avowing, she disagrees with the Traditional Expressivist insofar as she also takes avowals *as products* to express propositions by semantically representing them. So, when everything is in order, one's avowal action-expresses the very same state that is represented by the proposition it semantically expresses, and this is what accounts for its truth-evaluability (and, indeed, truth) despite the fact

that it expresses one's first-order state rather than (or in addition to—more on this later) a belief about one's first-order state.

We can now see that there are at least two ways to secure the truth-evaluability of avowals, and hence very little reason to prefer Traditional Expressivism—with its radical logico-semantic consequences—over Neo-Expressivism. In the next subsection I fill in some further features of Neo-Expressivism, features that will eventually help us to see what is at stake in developing and defending the Neo-Expressivist explanation of authority.

§2.4.2—Neo-Expressivism Expounded: Neo-Expressivism as a Non-Epistemic Account

At the beginning of this chapter, I described expressivist accounts as providing a *non-epistemic* route to explaining authority. Here I will explain that the Neo-Expressivist's opposition to epistemic explanations of authority concerns their resistance to treating avowals as the upshots of *introspection*. For our purposes, 'introspection' can be understood as a placeholder for whatever method we might use to *recognize* or *detect* the mental states we are in.

We saw that Jacobsen, in denying cognitivism, denies that avowals express self-beliefs. For this reason, it is reasonable to see him as opposed to an introspective account of avowing.[53] Bar-On writes that "[t]hough I am genuinely ascribing a thought with a particular content to myself [in avowing], *neither* the content *nor* the state need be in any way an epistemic target for me" (2004, p. 214). Finally, David Finkelstein (2003, pp 20-21) argues that having to introspect one's mental

[53] Indeed, Jacobsen (ms.) seems clear about this.

states renders one *alienated* from them.[54] Alienated self-ascriptions express one's *consciousness of* one's mental state, but they do not express one's first-order mental states conscious*ly*.[55]

One way of capturing the shared idea here is that avowals are, like other expressive behaviours, *spontaneous*. I smile when I see you, and my doing so expresses my joy. But, intuitively, my joy is not something that I first detect and then communicate to you through my smile. I simply smile (or avow), acting as a joyous person does. So, just as we have reason to resist thinking of smiles as achievements of introspection, we have reason to resist thinking of avowals—these being akin to smiles in their expressive function—as achievements of introspection. Indeed, insofar as introspecting one's mental states means detecting their presence, we should understand any self-ascription backed by introspection as a *report*, one that speaks *from the perspective from which the mental state was detected* rather than *from that mental state itself*.

Thinking of avowals as non-introspective self-ascriptions also provides us with an account of the difference between relating to your mental states first- versus third-personally: relating to your mental states first-personally means being in a position to express them without introspective effort, whereas being in a third-personal relation to your mental states means merely being able to issue introspection-backed reports about your mental states. Indeed, this simultaneously explains what makes avowals *first-personal* rather than *third-personal* self-ascriptions.

[54] While I take it that Finkelstein would agree that these ways of speaking correspond to the sort of report/expression contrast with which we have been working, Finkelstein later acknowledges that one can report *by* expressing. Any such report, however, will not be a *mere report*, which he glosses as "*an attempt (or merely apparent attempt) to inform someone of a fact that the speaker has learned or ascertained*" (2003, p. 97).

[55] To take an example of a particular introspective model that he rejects, Finkelstein (2003, pp. 23-24) opposes the idea that avowals are products of inner-perception, since *perceiving one's anger* is not tantamount to *occupying one's anger*, and this prevents one from *expressing* one's anger.

Interestingly, however, Bar-On distinguishes between *kinds* of avowals at a more fine-grained level than other Neo-Expressivists do, by acknowledging a continuum along which avowals can shade into reports. At the one end we have "avowals proper":[56]

> In this subclass [of avowals], we have verbal acts in which a subject volunteers a present-tense mental self-ascription spontaneously and unreflectively, not in response to an invitation to describe their state, nor even (we may suppose) with the aim of informing their audience of what is going on in them. We also have in this subclass self-ascriptions produced in utter silence, and not in the presence of any (real or imaginary) audience, as when someone says to herself "Boy, I can't stand him"…As I understand them, avowals proper do not plausibly represent the self-ascriber's opinion, or carefully formed judgment, that she is in the ascribed mental state. They are not reasonably regarded as the culmination of the subject's inwardly directed truth-targeting reflection. Furthermore, avowals proper are not acts that subjects deliberately undertake to perform with a specific audience-directed goal in mind, such as convincing, informing, pleasing, etc. (2004, p. 242)

Avowals proper are those on which we have been primarily focusing thus far.[57] It is important to note that avowals proper, being spontaneous, are not therefore just like *outbursts*. Avowals, for all Neo-Expressivists, can be intentional acts—they are, after all, often *speech acts* (possible exceptions being avowals as passing thoughts). What matters is that they are not easily understood as intentional actions on a traditional Davidsonian model of action explanation; they are not produced by beliefs (that one is in some mental state) and desires (to express some mental state) (cf. Hursthouse 1991). Rather, they may be intentional simply in the sense that one has a certain measure of control over them.

Next, contrasting with avowals proper, Bar-On writes:

> There are also what may be described as *non-evidential reportive avowals*. Having mastered the use of "I want the teddy" as a linguistically articulate expressive device, Jenny may put that device to partially reportive use. For instance, she may offer that kind of self-ascription in answer to such questions as "Why are you looking in that drawer?" or "What do you want most right now?". In such cases, the self-ascriptive

[56] Bar-On & Long (2001) is the original source of this term.
[57] See also Green (2007, §3.3-3.4), for an extended critical discussion of Grice's claim that speaker meaning necessarily requires audience-directed intentions in the way that assertions do (or, if not all assertions, then at least those assertions that qualify as "tellings", in the sense of Fricker 2006).

> utterance may seem to shade readily into ordinary reports. If it is still regarded as more secure than such reports, we may wonder why. I suggest that, if we regard non-evidential reportive avowals as more secure than other reports, theoretical self-reports included, this is still because, or *to the extent that*, we regard them as directly expressive of the self-ascribed state. (2004, p. 301)

Here the idea is that, because some avowals are offered in response to questions, they are not entirely spontaneous. Instead of simply coming out with one's mental state by avowing, some avowals are embedded in a conversational enterprise, whereby part of their point is to answer a question. But Bar-On's use of the term *non-evidential reportive avowal* should signify that she does not think of these avowals as *mere* reports. Thus, if I am asked what I want most right now, my reply can still express the very desire I avow, all without its being based on prior introspection of my desire that would allow me to express a belief to the effect that I have a desire. We can better understand this possibility by pointing to this passage:

> If a question arises about one particular aspect of what is going on at a given moment, we may have to pause to be able to offer an answer. But this isn't necessarily because we must survey our mental scene with an inner eye, or theorize about the causes of our behavior; the purpose of reflection may just be to eliminate what can be described as background noise and let the right condition come to the surface, as it were. (2004, pp. 302-303)[58]

Here we have in play a notion of 'reflection' that is intended to be distinct from introspection. Whereas introspection has been understood here as an epistemic process through which we detect our mental states, reflection is more akin to a meditative effort to *clear away* mental clutter that allows one's first-order mental state to "come to the surface" than it is an attempt to *focus in on* one's current mental condition. One is not trying to discover oneself, at least not exactly. Rather, one is trying to put oneself in a position to spontaneously come forth with oneself. Insofar as one has to reflect before avowing, one's avowal will lack the spontaneity

[58] Bar-On makes a similar claim about *attending*: "Attending need not be construed as an epistemic method of obtaining information or discovery; it can be seen instead as a psychological device for putting oneself in a position to *speak directly from* one's condition" (2004, p. 408).

needed to count as an avowal proper. But it is nevertheless not a mere report, because it is not backed by introspection.

Here, then, are the core claims of Neo-Expressivism as I am currently understanding it:

(1) When all goes well, avowals action-express the very mental states they semantically-express

(2) Avowals are truth-evaluable

(3) Avowals are, to varying degrees, spontaneous, epistemically unmediated acts

There are many further issues of detail that might divide Neo-Expressivists. Some of these will crop up throughout §2.6-§2.7. For now, however, it is time to apply Neo-Expressivism to the task of explaining authority.

§2.5.1—Neo-Expressive Authority: Indubitability$_{1+}$

To begin, consider:

> **Indubitability$_{1+}$**: Necessarily, if a person *S* tells another person *H* at time *t* that she (S) is currently in mental state ϕ, then *H* cannot rationally doubt at *t* that *S* is currently in ϕ—provided that *H* assumes that (i) *S* is sincere, (ii) *S* made no slip of the tongue, (iii) *S* is not conceptually confused, (iv) *S* issues her avowal under normal cognitive conditions, and (v): S's avowal is sufficiently coarse-grained.

In Chapter One we saw that Indubitability$_{1+}$ ranges, most plausibly, over avowals of conscious experiences, such as visual experiences or purely sensory states. How can Neo-Expressivists explain Indubitability$_{1+}$?

To begin, notice that the *non-epistemic* character of the Neo-Expressivist account is doing significant work here. In issuing what I called *coarse-grained* avowals in Chapter One, I was referring to avowals where the *content* avowed was not something that, presumably, could only be self-ascribed on the basis of some sort of epistemic activity like *counting* the number of dots in one's visual field. To the extent that coarse-grained avowals do not require introspective

acts like this, it is possible to understand avowals of visual experiences as more or less spontaneous expressions of the agent's state of mind, just as Neo-Expressivism claims. Thus, as Jacobsen says, "[f]irst person immunity to error rests on no observational or inferential achievements and is not any sort of epistemic accomplishment" (1997b, p. 136). Likewise, Bar-On says that avowers cannot be regarded as "*mistaking*" one mental state for another (2004, p. 200).[59] So, if *H* cedes proper uptake to *S*'s avowals, thereby recognizing their expressive character, *H* will take it as a more or less spontaneous action-expression of the very conscious experience it semantically represents. No room for doubts about introspective failure are reasonable (again, given sufficiently coarse-grained content), and so there is no further room for reasonable doubt, given (i)-(v). [60]

§2.5.2— Neo-Expressive Authority: Indubitability$_{\text{Brute-Error}}$

Consider next:

> **Indubitability**$_{\text{brute-error:}}$ Necessarily, if a person *S* tells another person *H* at time *t* that she (*S*) is currently in mental state ϕ, then *H* cannot rationally doubt at *t* that *S* is currently in ϕ—provided that *H* assumes that (i) *S* is sincere, (ii) *S* made no slip of the tongue, (iii) *S* is not conceptually confused, (iv) *S* issues her avowal under normal conditions of cognitive well-functioning, and (v) *S* has not succumbed to some psychological failing.

Recall that Indubitability$_{\text{Brute-Error}}$ was taken to range over all sorts of avowals. With this in mind, the Neo-Expressivist explanation of Indubitability$_{\text{Brute-Error}}$ is still similar to the Neo-Expressivist explanation of Indubitability$_{1+}$. Once again, the key idea is that because avowals are not ordinary

[59] Bar-On's discussion of immunity to brute error is far more nuanced than I have set out here. She also makes a big deal of how avowals exhibit *immunity to errors of misidentification* of the subject and *immunity to errors of misascription* of the avowal's content (2000, 2004).

[60] One might wonder about avowals of visual experiences. Compare: "I can see your anger" and "I can see your visual experience". One might think that the first is felicitous whereas the second is not. Let this be so, but the point is not to assimilate *all* ways of *showing* one's mental states (via avowals) to others as on the model of *seeing* the state or *seeing that* one is in it. Rather, when a speaker avows her visual experience, her articulation of its content is what is put on display for you, and your linguistic uptake allows you to access this fact about her.

reports, they are not subject to the ordinary trappings of observation (recall Falvey's point that the concept of reporting is cognate with the concept of observation). This is relevant because brute errors are easily understood as failures of observation: one takes oneself to be looking at a barn despite, unbeknownst to one, being in fake barn country. In this case, the world "fails to cooperate" and the agent is not epistemically culpable, nor is anything about her psychological condition to blame. Intuitively, such errors are not possible in avowing, and Neo-Expressivism—being a non-epistemic account of authority—explains why. For, when we give proper uptake to avowals, we take them to be issued from the very mental states avowed, and so we do not take them to express beliefs about those states that are based on introspection, and so do not take them as able to go wrong due to an uncooperative (inner rather than external) world.

§2.5.3—Neo-Expressive Authority: Presumption of Truth$_3$

Let us now move on to our presumption-of-truth-based specifications of authority, beginning with:

> **Presumption of Truth$_3$**: If (1) a person *S* tells another person *H* at time *t* that she (*S*) is currently in mental state ϕ and (2) *H* lacks any good reason to doubt what *S* says, then *H* is *more justified* at *t* in believing that *S* is currently in ϕ than *H* would be in believing an ascription of ϕ to *S* by some other person, *P*, even lacking any reason to doubt P, and even if *H* possesses no positive evidence that *S* is currently in ϕ (aside from the fact that *S* tells him so).

Neo-Expressivism explains Presumption of Truth$_3$ as follows. If *H* gives proper uptake to *S*'s avowal, then *H* takes it to action-express the very state it semantically represents. In so doing, *S* effectively *shows* or *reveals* her mental state to *H*. *H*'s justification to believe the avowal is therefore quite robust, for *S*'s avowal ideally provides *H*'s belief about *S*'s mind with a *factive* source of justification—it justifies her in taking the avowal to be true on *perceptual* grounds.[61]

[61] Perhaps, for extratheoretic reasons, we should eventually want to distance ourselves from a perceptual account of knowledge of other minds. Bar-On acknowledges the space for this possibility when she writes that "What is

Contrariwise, *H*'s justification to believe what *P* says about *S*'s mental state is not factive, perceptual justification, because *P*'s other-ascription does not express *S*'s mental state. *H* may not doubt that *P* has evidence, perhaps even overwhelming evidence, that *S* is in ϕ. Nevertheless, because this evidence does not add up to anything as direct as *showing S*'s mental state to *H*, *H*'s justification to presume the truth of *P*'s other-ascription is inevitably inferior to her justification to defer to *S*'s avowal.

§2.5.4—Neo-Expressive Authority: Groundless-Authority$_1$

Consider, next, the first of our two 'Groundless-Authority' theses:

> **Groundless-Authority$_1$**: If (1) a person *S* tells another person *H* at time *t* that she (*S*) is currently in some mental state ϕ and (2) *H* lacks any good reason to doubt what *S* says, then *H* is justified at *t* in believing that *S* is currently in ϕ even if *H* possesses no positive evidence that *S* is currently in ϕ (aside from the fact that *S* tells *H* so), and even though there is no epistemic reason *S* can reasonably provide for her avowal if requested to do so by *H* (such that any request from *H* is inappropriate)
>
> **CC to Groundless-Authority$_1$**: By contrast, if (1) a person *P* tells another person *H* at time *t* that a third person *S* is currently in some mental state ϕ, and (2) *H* possesses no positive evidence that *S* is currently in ϕ (aside from the fact that *P* tells *H* so), then—even if *H* lacks any good reason to doubt what *P* says, and even if *H* is justified at t in believing that *S* is currently in ϕ—there must be epistemic reasons *S* can reasonably provide for her avowal if requested to do so by *H* (such that any request from *H* is appropriate)

What would justify *H* in believing *S*'s avowal despite the fact that *S* cannot be reasonably expected to provide any epistemic support for it? Neo-Expressivism explains this neatly. Thus, when *H* interprets *S*'s avowal aright, *H* takes it to reveal the very state avowed. *H* does not take it as requiring any sort of epistemic support, because it directly expresses the relevant state. Moreover,

---

distinctive of behaviors that give direct expression to mental states could also be preserved if there was a way to fund the epistemic contrast between immediate versus evidence-based uptake [on the witnesses' part]—a contrast that is independently useful in epistemology (with perception being only one species of the genus 'immediate, non-evidence-based uptake')" (2010, p. 62). See also Dain (2019) for ways of thinking about possible differences between *kinds of seeing* that may be relevant here, as he takes them to be for Wittgenstein.

because *S*'s avowal springs from the state itself, more or less directly and spontaneously, there are no epistemic reasons that *S* has for avowing it. So Neo-Expressivism explains the Groundlessness of avowals, in terms of both *S*'s and *H*'s perspectives: it explains why *S* has nothing to say in defense of her avowal, and why *H* would be out of bounds to request any such defense. Contrariwise, *P*'s other-ascription does need to be supportable epistemically, even if demands for epistemic support are never actually made by *H* on *P*. This is because *P*'s other-ascription does not directly express *S*'s mental state, and so justification for the *belief about S*'s mind that *P*'s other-ascription expresses must be available. Perhaps *P* can see that *S* is in ϕ or has inferred that *S* is in ϕ. If *P* fails to deliver some such response (if and when asked), then *H*'s justification to believe *P* will be forfeit. Not so with avowals, since no such epistemic support need be given for them in the first place.

§2.5.5— Neo-Expressive Authority: Groundless-Authority$_2$

Finally, consider:

> **Groundless-Authority$_2$**: If (1) a person *S* tells another person *H* at time *t* that she (*S*) is currently in some mental state ϕ and (2) *H* lacks any good reason to doubt what *S* says, then, given normal circumstances, *H*'s willingness to question whether *S* really is in ϕ is rather low, and even though there is no epistemic reason *S* can reasonably provide for her avowal if requested to do so by *H* (such that any request from *H* is inappropriate)
>
> **CC to Groundless-Authority$_2$**: By contrast, if (1) a person *P* tells another person *H* at time *t* that a third person *S* is currently in mental state ϕ and (2) *H* possesses no positive evidence that *S* is currently in ϕ (aside from the fact that *P* tells him so), then, given normal circumstances, *H*'s willingness to question whether *S* really is in ϕ is not as low—even if *H* lacks any good reason to doubt what *P* says—and even though there are epistemic reasons *S* can reasonably provide for her avowal if requested to do so by *H* (such that any request from *H* is appropriate)

Why do we tend to more readily presume the truth of avowals than their counterpart other-ascriptions? Once again, Neo-Expressivists have an answer at the ready. Here is Bar-On:

> The avowing subject can do what no one else can: she can speak from the very states of mind she ascribes to herself. Indeed, I would suggest that there is a direct correlation between the degree of security we would be prepared to assign to a present-tense mental self-ascription and the extent to which we take the subject to be speaking her mind. (2004, p. 302)

> To regard a linguistic act as an avowal is to take it as an expression rather than a mere report of the ascribed condition. It is to take the avowing subject to be speaking directly from her condition, where the self-ascription tells us *what* condition is to be ascribed to her. All that we as audience need to know to identify the condition being expressed is linguistic uptake (2004, pp. 316-317).

When we cede the appropriate linguistic uptake to avowals, we acknowledge their direct, expressive character. This explains why we are not disposed to seek out an epistemic basis for them, and why we are disposed to take them as true. Contrariwise, in the case of *P*'s other-ascriptions, grasping their reportive nature only gives one a window into the fact that *P* has expressed a belief about *S*'s mind. Thus, while we may sometimes presume the truth of *P*'s other-ascriptions, we will simultaneously recognize that this is reasonable only insofar as *P* is in possession of good epistemic reason(s) for what she says. And for this same reason, we will be more likely to question *P*: we might question her epistemic grounds, whereas we will not do so for *S*, since requests for epistemic grounds are besides the point if *H* understands the expressive nature of *S*'s avowal.

§2.6.1—Additional Features of Neo-Expressivism: Moore's Paradox

Having offered a Neo-Expressivist explanation of authority in its various specifications, I spend this section advancing the case for Neo-Expressivism by pointing to two further explanatory fruits of the view. After doing so, I will address several objections to Neo-Expressivism in §2.7.

To begin, we can take a look at the Neo-Expressivist analysis of what has come to be called Moore's Paradox. The paradox, first pointed out by and hence named after G. E. Moore, has to do with thoughts and utterances like: "*p*, but I don't believe that *p*". These thoughts and

utterance are paradoxical, on their face, because they seem highly irrational despite the fact that they are not formally contradictory (after all, it is possible for *p* to be true despite my not believing it).

Neo-Expressivists have noticed that their conception of avowals lends itself to a "smooth account" of Moore's Paradox (Jacobsen 1996, p. 28). For, given an expressivist analysis of avowals, we can understand "*p*, but I don't believe that *p*" as *expressing* both a belief that *p* and a belief that *not-p,* despite the fact that there is no semantic-level contradiction. Here is Bar-On:

> If I utter the Moore sentence 'I'm thinking that there's water in the cup, but there's no water in the cup' the sentence I produce is not self-contradictory. However, for all that, when avowing 'I'm thinking that there's water in the cup' I may be expressing in the action sense…the same state that I express when saying 'There's water in the cup' (2010, pp. 51-52)
>
> …But going on to say under the same breath 'There's no water in the cup,' thereby [action-expressing] one's state of thinking that there's no water in the cup, would land one in an expressive conflict. (2010, p. 53)

On the Neo-Expressivist analysis, then, Moore-paradoxical thoughts and utterances are readily explained. I will appeal to this explanation in Chapter Three (see §3.4.4) for a purpose that need not yet concern us.

§2.6.2—Additional Features of Neo-Expressivism: Transparency to the World

Bar-On connects Neo-Expressivism to the so-called *transparency to the world* of avowals. This phenomenon has been famously captured by Gareth Evans, who writes:

> [I]n making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outward—upon the world. If someone asks me "Do you think there is going to be a third world war?" I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question "Will there be a third world war?" (1982, p. 225).[62]

[62] It has sometimes been complained, rightly I think, that Evans ought not to have said *must* here.

The first point of contact between Evans and Neo-Expressivists is their disavowal of introspective conceptions of avowing. However, the novel thought that Evans puts forward is that there is nevertheless an *extrospective* quality to what goes on in the course of avowing, at least where one's avowal has a worldly intentional content. While different accounts of exactly what goes on in extrospection vary,[63] a simple and valuable point of agreement can be found between Evans and Bar-On.[64] In Bar-On's words:

> If asked (or when considering) whether you believe p, you will normally directly attend to whether p is the case. We can think of this as a way of putting yourself in a position to give direct voice to your (first-order) belief, which is what the Neo-Expressivist account says you do when avowing. (2015, p. 142)

This point is important for Neo-Expressivists because it allows them to make better sense of what kind of cognition *does* transpire in the course of issuing an avowal, at least for avowals that do not count as "avowals proper", i.e., avowals where some deliberative effort does precede one's avowal, such that it is not perfectly spontaneous. Thus, Bar-On writes:

> What is important is that in neither case you need to discover what you believe. Instead, you simply give it voice, having considered (or reconsidered) whether things are as the proposition says. You pronounce on the truth of the proposition, though you are using a self-ascriptive expressive vehicle. (2015, p. 142)

In sum, the Neo-Expressivist take on Evans's insight accomplishes two things: it explains the transparency-to-the-world intuition in Neo-Expressivist-friendly terms, and it further enhances our understanding of how avowals can shade toward the more deliberate, studied end of the self-ascriptive spectrum without lapsing into introspective reports. There will be more to say about the

[63] Chiefly, these are debates about whether self-ascriptions produced via extrospection are, paradigmatically, exercises of rational agency (cf. Moran 2001; Keeling 2019a, 2019b), or whether they are epistemically reasonable precisely when they are produced without exercising one's rational agency (Shah & Velleman 2005; Parent 2017; Barz 2019).

[64] See also Falvey (2000, p. 81), who deploys the idea of transparency to the world to argue that "[d]oing justice to the idea that sincere avowals of belief involve the rational exercise of cognitive faculties does not require the postulation of a special cognitive faculty that informs each individual what she believes", since one can simply appeal to one's ordinary capacity for world-directed reasoning.

transparency-to-the-world phenomenon in Chapter Five.[65] For now, let us consider a series of objections for Neo-Expressivism.

§2.7.1—Objection to Neo-Expressivism: Explanatory Buck-Passing

In this section I respond to several objections facing Neo-Expressivism. We can begin with what is perhaps the most straightforward objection to Neo-Expressivism, namely, that it does not actually explain authority, because it does not explain *H*'s capacity to *take S*'s avowals as expressive in the required senses. This is crucial since, as the reader may have noticed, Neo-Expressivist explanations of authority frequently require us to view *H* as ceding proper *linguistic uptake* to *S*'s avowal. The question, then, is what explains *this* capacity of *H*'s. This has not yet been answered, and so it may seem that the explanatory "bulge under the carpet has only been moved elsewhere" (Byrne 2011c, p. 716; 2018, p. 72).[66]

I agree that we surely want an account of how agents are able to take expressive behaviour as expressive, and an account of how this goes in the avowing case, such that their recipients are justified in and psychologically motivated to presume the truth of avowals, and to take them as relatively indubitable. However, I do not believe that this concession undermines Neo-Expressivism. This is because the task of explaining how we take avowals to be expressive in various ways can be taken up *after* granting that the Neo-Expressivist analysis of avowals explains their authority. Indeed, I agree with Jacobsen that:

> In pre-philosophical innocence, we count an utterance of 'I'm disgusted' as an expression of disgust, an utterance of 'I want a drink' as an expression of a desire for a drink, and an utterance of 'I love you' as an expression of love. We are, in short, pre-philosophical expressive pluralists about self-ascriptions of mental states. (1997b, p. 133).

[65] This will be in connection with certain theories of self-knowledge. For more on the connection between (Neo-Expressive) avowing and transparency, see Falvey (2000).
[66] Owens (2007) complains, similarly, that Bar-On simply helps herself to the view that avowals "*show* one's occurrent mental states to the other", as does Brueckner (2011).

Similarly, though he is focusing on expressive behaviour generally rather than avowals specifically, Mitchell Green points out that "alleged imperceptibility of all psychological states is not in fact a bit of common sense; everyday discourse goes the other way" (2007, p. 90).[67] I take it that these observations are exactly right, and that it is from *here* that we have at least prima facie justification to offer Neo-Expressivism as an account of authority.

Of course, this leaves the Neo-Expressivist in an at least somewhat precarious position, since failure to supply a further explanation might leave these claims about our pre-theoretical views vulnerable to skepticism or dismissal. Fortunately, resources for saying more can be found in Green's *Self-Expression* (2007), which contains no discussion of avowals but is still a useful tome for studying self-expressive behaviour more broadly. Thus, I will close this subsection by first describing two forms of what Mitchell Green calls "*perception-enabling showing*" (2007, p. 85), and then applying them to the avowing case. In doing so, we will also see that one of Green's key ideas has already been anticipated by Jacobsen.[68]

Green draws our attention to two forms of perception-enabling showing, namely, perception-enabling showing "as *part-whole perception* and as *perceiving in*" (2007, p. 86). To take a mundane example of part-whole perception, Green gives the example of one's perceiving an apple by perceiving its face. Naturally, I do not perceive the entirety of the apple by perceiving its face. Still, I do perceive the apple, rather than some distinct object like *the face of*

[67] Here, also, is Wittgenstein: "It is possible to say 'I read timidity in this face', but, at any rate, the timidity does not seem to be merely associated, outwardly connected, with the face; rather, fear is there, alive, in the features…" (1953, §537). And again: "'We see emotion.'—As opposed to what?—We do not see facial contortions and make inferences from them (like a doctor framing a diagnosis) to joy, grief, boredom. We describe a face immediately as sad, radiant, bored, even when we are unable to give any other description of the features.—Grief, one would like to say, is personified in the face. This belongs to the concept of emotion" (1967, §225).

[68] While I draw on Green's and Jacobsen's ideas below, one can also see Bar-On (2004, pp. 273-274) for some remarks that should soften us up to the idea that avowals enable perception of the mental states self-ascribed. Notably, Bar-On also shows some sympathy for Green's views at (2004, p. 298).

a would-be apple. In other words, I perceive an apple *by perceiving a part of it*.[69] I perceive this because the face of an apple is a "characteristic component" of an apple, rather than a mere sign of it (2007, p. 87).

After qualifying that "a thing will, for purposes of perception, only be a characteristic component of an object *relative to an organism O in ecological situation E*," Green formalizes the preceding points as follows, where $\propto$ is a perceptible object, event, or process:

> **Part-Whole Perception**: relative to an organism O and ecological situation E, a *characteristic component of* $\propto$ is a part of $\propto$ that, when perceived in E without any other part of $\propto$ being perceived, enables O to perceive $\propto$.[70]

Besides part-whole perception, we have *perceiving in*. In this case, "instead of being shown a thing by sensing one of its characteristic components, we may instead be shown a thing A by sensing a distinct object B in such a way that we see (or in some way sense) A in B" (2007, p. 87). Drawing from Walton (1984, 1997), Green offers the examples of seeing an object in a mirror, telescope, or photograph (in the case of vision), and hearing something through a recording device (in the case of audition).

Green goes on to apply this distinction between forms of perception-enabling showing to the self-expression of emotions and other mental states through non-linguistic behaviour. I will not evaluate these applications, since our interest is in avowals specifically. Here, then, is how we might apply Green's distinction to the avowing case. First, consider avowals made in contexts where *H* has immediate perceptual access to *S* and *S*'s speech. The suggestion I want to

[69] As Green says, "we should be on guard against a confusion between seeing the apple and seeing the whole apple" (2007, p. 86).

[70] As regards the *relative to an organism O and ecological situation E* qualifier, one example is that the perceptibility of a state can be relative to a perceiver's conceptual repertoire—only then can one *perceive* $\propto$ *under a concept*, or *see that* $\propto$ in the Dretskean (1973) sense.

make now is that, when *S* avows that she is in ϕ, she enables part-whole perception of ϕ for *H*. *S* enables this for *H* because *S*'s avowals are characteristic components of her mental states.

Jacobsen explains how this can be the case. Being in a certain mental state standardly disposes one to various sorts of behaviour. Now, "only if a tendency or disposition to ϕ is a tendency or disposition to, *inter alia*, say that one ϕs, can saying that one ϕs count as an expression (or manifestation) of that disposition or tendency" (Jacobsen, ms). The idea, then, is that avowing is one manifestation of the broader dispositional cluster that constitutes being in a mental state. Insofar as this is so, avowing provides a means for *H* to perceive *S*'s mental state, because *S*'s avowal not only signifies but *manifests* part of the dispositional cluster that *is itself* the mental state avowed.

Green is skeptical that we can enable literal part-whole perception of (at least many) mental states. He writes, for example, that an assertion "expresses, and thus shows, a belief if it is sincere, but beliefs are not the sorts of things that can be perceived. Rather, a sincere assertion shows a belief by showing *that* we believe the content asserted, thereby enabling others to be aware of it…" (p. 25) However, if Jacobsen is right that avowals of beliefs manifest a portion of the dispositional profile that constitutes belief (at least for subjects who have learned to avow) then we might endorse a stronger conclusion than Green: we might say that perceiving the *avowal* is perceiving (a part of) a belief. Indeed, similar remarks apply to non-dispositional avowals, e.g., of pains and tickles: we do not experientially perceive someone's pains and tickles through her avowals, but we can see hear avowing manifestations and *thereby* perceive them.

Now consider cases in which *S*'s avowal is not immediately perceptible by *H*. These could be cases where *S*'s avowal is transcribed in a medium like a diary or audio/video recording that *H* then accesses. It is obvious that such transcriptions and recordings are not themselves

characteristic components of one's mental states or of one's avowals. Nevertheless, these media can enable *perceiving in* on *H*'s part. For example, the audio recording allows us to hear the avowal, and, I should say, to really hear *it*, albeit through a mediating apparatus. And since avowals manifest mental states, we can perceive mental states through media that do not themselves have mental states as characteristic components. The only outliers are avowals that we merely imagine *S* to have made, i.e., avowals made merely in thought that are never transcribed in any media for *H* to access. Here I think the correct thing to say is that we do not perceive them, but that we model our deferential practices onto them *because* they are, in point of their expressive character, exactly like those we *do* perceive when expressed.

This has only been a sketch of how a fuller account of the perceivability of mental states through avowals is possible, and how *H* might actually perceive them. A richer picture might draw on further aspects of Green's highly detailed discussion of self-expression, or it might take us in different directions. My aim has only been to show how things might go at a programmatic level, so that the above objection to Neo-Expressivism is at least partially addressed.

At the close of this subsection, let me preempt one further source of possible unease. One might argue that we have *once again* passed the explanatory buck, this time by failing to explain how agents learn to directly express their first-order mental states through their avowals in the first place. Here I only offer a few remarks in the direction of a fuller developmental story. The simplest cases will be those in which there is some natural, non-linguistic expressive behaviour onto which an avowing capacity can be scaffolded, as when we teach a child to replace winces and smiles with avowals of pain and joy.

Other cases may be slightly more complicated—for, as Coliva (2016, p. 147) claims, *beliefs* typically lack characteristic non-linguistic expressive behaviours (what sorts of non-

linguistic behaviour characteristically manifests one's belief that the earth is large?).[71] Here, however, we can follow Bar-On who, keeping with the example of belief, argues that "we can appeal to the first-order *linguistic* expressions of beliefs as candidates for replacements by self-ascriptions."[72] Paradigmatically, these will be first-order assertions. Now, as we have seen, avowals might not be easily understood as assertions (§2.4.1). Nevertheless, Bar-On's move is available because "the expressivist account in no way denies that there are ordinary assertoric or descriptive utterances; it only denies that self-ascriptions of mental states are always offered as assertoric or descriptive reports of one's mental states" (2004, p. 293).[73] So, by teaching an agent to replace her first-order assertions with avowals, she may learn a form of expressive behaviour that is not itself neatly classifiable as assertoric.

§2.7.2—Objection to Neo-Expressivism: No Authority for Avowals of Dispositional States

A second and related objection is really an objection to realism about authority as regards certain avowals, not just a Neo-Expressivist account of the authority of such avowals. The objection is that many mental states are broadly dispositional (e.g., standing beliefs, desires), and it is puzzling how we could have authority with respect to dispositions.

This objection can be made perspicuous through an analogy, due to Wright (1989, pp. 292-294). All parties to conversations about authority agree that we are *not* authoritative with respect to *all* of our mental features. Take, for example, a character trait like bravery: a self-ascription of bravery does not seem indubitable or presumptively true in any distinctive sense, for one might

[71] It is unlikely that they *always* lack characteristic non-linguistic expressions. To take an example from Ryle (2009, p. 118), a characteristic non-linguistic expression of a skater's belief that the ice is thin might be her wary skating. (Thanks to Victoria McGeer for this point).
[72] Jacobsen (ms) notes that "[a]lthough in its initial appearance Wittgenstein ties avowal-expressivism to speculations about language acquisition, it does not depend on the truth of those speculations".
[73] This description of how belief-ascriptions can be learned can meet an objection, by Coliva (2016, p. 152), against the Traditional Expressivist who, she claims, fails to explain how one could learn to avow belief, since beliefs typically lack characteristic non-linguistic behaviour.

reasonably demand that one *proves* one's bravery by way of demonstration. To be brave is to be disposed to perform courageous acts, and so one can test for the presence of this disposition by seeing what one does in a courage-demanding situation. Now, if it is precisely the dispositional character of bravery that renders these questions apt, one might wonder why we should think any differently "about psychological states which have no distinctive occurrent phenomenology and which have to answer, after the fashion of dispositions, to what one says and does in situations so far unconsidered?" (Ibid., p. 294).

My response is that Neo-Expressivism is especially well positioned to accommodate this concern. Recall Jacobsen's point, raised in the previous subsection, that:

> Only if a tendency or disposition to ϕ is a tendency or disposition to, *inter alia*, say that one ϕs, can saying that one ϕs count as an expression (or manifestation) of that disposition or tendency…*Expressions* of a state of a given type are among the behavioural manifestations of the state that, *inter alia*, serve to qualify (constitute) a subject as having a state of that type" (Jacobsen, ms.)

Now, consider what he subsequently says:

> An attribution of a disposition-like state will be true only if the subject of the state is also disposed to a wide and unsurveyable range of behaviour that she cannot be expected to foresee. This was supposed to be the source of trouble for authority. But if the subject's expression of that disposition-like state is sincere [in avowing] then she is, *a fortiori*, in the very state that disposes her toward the behaviour in that unsurveyable range. (Ibid.)

In short, Wright is correct that whether one is in a dispositional state is beholden to a range of behaviours that one may not be sure, say, by introspection, that one would perform until one sees oneself perform them. But since avowing is one way of manifesting the very state at issue, then whether one has also witnessed those behaviours is beside the point. One says something true about oneself insofar as one manifests the very state one's avowal self-ascribes, whether or not one knows how one would behave in the as-yet-unsurveyed range of behaviour that also partly constitutes one's dispositional mental state.

§2.7.3—Objection to Neo-Expressivism: The Contingency of Expressive Authority

As A. Minh Nguyen notes, several philosophers have tended to view authority as a so-called "mark of the mental".[74] Thus, Davidson declares that first person authority "is not an empirical discovery, but rather a criterion, among others, of what a mental state is" (1994, p. 234). Similarly, Shoemaker argues that "[authority] is of the essence of mind" (1990, p. 50) and that "it is constitutive or definitive of mental states, or of minds, or of the concepts of these" (1996, p. 25). Finally, even though I raised suspicions in Chapter One (§1.3.2) about whether this could adequately capture any idea of authority, it is useful to recall Wright's claim that "[w]holesale suspicion about my attitudinal avowals – where it is not a doubt about sincerity or understanding – jars with conceiving of me as an intentional subject at all" (2001, p. 325). Now, it is not obvious that these authors understand authority in exactly the same ways, nor is it obvious that they would specify authority in the ways that I have specified it in Chapter One. Nevertheless, Nguyen draws a problematic conclusion for Neo-Expressivism, which is that it fails to account for the *necessity* of authority. To voice his concern, he focuses on authority as a presumption of truth, which he describes thus:

> [It] is necessary that, for any person, if he sincerely ascribes the presence, or absence, of a particular intentional state to his present self, then there is a legitimate presumption that what he says is true. (2008, p. 104)

We need not accept this as a totally adequate specification of authority in order to see Nguyen's objection. If one likes, one can substitute something like Presumption of Truth_3 or Groundless-$\text{AT}_{1/2}$ in its stead.

To set up the objection, recall that Neo-Expressivists place a great deal of stock in the fact that avowals action-express the very mental states they self-ascribe. However, it is possible to

[74] To my knowledge, this phrase goes at least as far back as Rorty (1970).

imagine worlds in which agents *do not* typically (or at all) action-express the very mental states they self-ascribe. Thus, we can imagine a world in which its citizens have thoroughly internalized Freudian concerns about the unconscious, such that they are constantly skeptical about their capacity to express their first-order mental states by means of self-ascriptions (perhaps because they take it that self-ascriptions are typically self-deceived). With Jacobsen, who responds to Nguyen, let us call such a world "F-World" (2008, p. 659) and its denizens F-Worlders, owing to the role played by Freudian ideology in the example.

Here, then, is the objection: at F-World, authority is *not* necessary given Neo-Expressivism, because it is not legitimate for some hearer, *H*, to view *S*'s self-ascriptions as true (at F-World) for the reason that Neo-Expressivism would suggest. Again, this is because *S*'s self-ascriptions at F-World do not action-express the same mental states that they self-ascribe.

Jacobsen offers a reply to Nguyen. First, he concedes that "no one in F-World would simply *presume* that a self-ascription is true, even knowing that the speaker is sincere. If F-World is logically possible, then it is not necessary that there be any such presumption" (2008, p. 659). However, Jacobsen thinks that this concession does not really undermine the necessity of authority. This is because, while "[w]hether there is any legitimate presumption to this effect is a contingent matter…*it is not contingent that persons who do make sincere expressive self-ascriptions will make true self-ascriptions*" (2008, p. 660). Put differently, "the idea that authority is a mark of the mental does not imply that such subjects will actually make any authoritative self-ascriptions" (2008, p. 661). What is *necessary* about authority is only that *if* agents issue genuine avowals, a presumption of truth will therefore be warranted.

Perhaps Nguyen will not be satisfied, for perhaps his view is that we should hold onto the idea that there is no logically possible world at which agents capable of producing self-

ascriptions never make first-person authoritative ones. This is a claim about the self-ascriptions actually issued at a world, not just about those that *would* be authoritative if issued at a world. Now, if we need not understand authority this way, then Neo-Expressivists need not say more than Jacobsen already has. But perhaps Neo-Expressivists can say more.

For one thing, Nguyen also describes F-World as a world at which F-Worlders never learned to action-express the same mental states that their self-ascriptions semantically represent. However, he may now bear the burden of explaining how F-Worlders could acquire psychological vocabulary. After all, as we saw in §2.2.1, Wittgenstein supplies an account of psychological vocabulary acquisition that conduces to an expressivist view. Again:

> Here is one possibility: words are connected with the primitive, the natural, expressions of sensation and used in their place. A child has hurt himself and he cries: then adults talk to him and teach him exclamations and, later, sentences. They teach the child new pain-behaviour. (*P.I* §244)

If Nguyen wants to avoid both Wittgenstein's private language argument and the expressivist view that would require him to accept that at least *some* avowals express the very mental states they self-ascribe, he will need an account of psychological vocabulary acquisition that does not depend on inner ostension. Perhaps he can offer such account. But I imagine this will be no easy task. So he faces a significant burden in justifying his initial setup of F-World.

Admittedly, this reply is only partial. The reason is that Nguyen may redescribe F-World as a world at which, even though agents must initially learn to avow along the lines proposed by Wittgenstein, they inevitably lose this avowing capacity later on as they begin to internalize their Freudian doubts. But this is already a sizeable concession, since it means admitting that there *are* actual instances of authoritative-cum-expressive avowals at F-World, albeit among its younger citizens.

Now, if we are focusing only on adult populations, I do not have a knock-down argument against the possibility of F-World. But I am not confident that this is a problem either. For, rather than taking F-World as a counterexample to Neo-Expressivism, I think focusing on how different it is from our world can help us to explain why some philosophers might have thought, *erroneously*, that first-person authority is necessary in the sense at issue—that, necessarily, some actual self-ascriptions produced (by adults) at any given world are authoritative.

First, note that, as Nguyen himself sets up the scenario, F-Worlders can express their attitudes linguistically in *some* ways. Thus, an F-Worlder can "can use '*p*' to express his belief that *p*, and use 'I believe that *p*' to express his belief that he believes that *p*." The only problem is that "he cannot use the latter to express his belief that *p*" (2008, p. 118). This is a sensible admission, since it is hard to imagine any speech act that does not, in a minded creature, express some sort of mental state.

From here, we can begin to see just how bizarre adult human F-World psychology is. Imagine an adult F-Worlder, Daniella, who is in conversation with a child F-Worlder, Chezwick. Suppose that Chezwick points and says "there's a fire truck!". If Daniella grasps the force of Chezwick's utterance, she notices that Chezwick has expressed his belief that there is a firetruck. In light of this, why should Daniella think it impossible to teach Chezwick to replace his first-order assertion with an avowal that expresses the same belief? Indeed, if Wittgenstein is right, her recognition of this will be integral to her facilitating Chezwick's development of his psychological vocabulary. But then why, when she turns to thinking about her own thought and speech, must she think it impossible to express *her* first-order attitudes through avowals?

The skeptical reply will be that it is Daniella's further Freudian beliefs, that she has acquired over the course of her life, that prevent her from having this insight. Thus, she might

look at Chezwick and say "ah, to be young and capable of directly expressing one's first-order states by avowing them! It is a shame that, as we grow older and succumb to myriad forms of repression and self-deception, we lose Chezwick's innocent capacity!" The problem is that, having acknowledged that this is indeed a capacity we *lose*, Daniella should now have to have a story of why the loss of her innocent avowing capacity is inevitable, and it is hard to understand what beliefs Daniella would have to have in order to rationally sustain this belief. It cannot just be her belief that many of her attitudes are repressed, as Freudians argue. For this belief does not explain why such loss is inevitable, especially given her admission that issues of repression do not necessarily stand between Chezwick and his avowing capacity. The point is not that it is hard to imagine how Daniella could believe that her attitudes are sometimes repressed and so must be accessed by way of third-personal epistemic labour. Rather, the point is that it is hard to imagine how she could think that this is her inevitable and chronic condition. This is especially compelling once we consider that, in being able to express her *second*-order beliefs by way of her self-ascriptions, she need not first form third-order beliefs about herself, on pain of regress (McGeer 2015, p. 269). If she recognizes this about her self-ascriptions, she will need a theory of why her Freudian skeptical beliefs only cut off a direct expressive tie between self-ascriptions and *first-order* mental states. Her Freudian beliefs look arbitrary in their scope, at this point.

Still, I must reiterate that I am not attempting to show that Daniella's web of belief is (rationally) unrealizable. What I am trying to demonstrate is precisely that her web of belief, even if (rationally) realizable, is nevertheless extremely alien. I am trying to demonstrate this in order to make sense of why philosophers might have thought that there are surely no counter-examples to the claim that at least some actually-issued self-ascriptions made at a given world

are necessarily authoritative. My view is that this belief may stem from the difficulty of imagining coherent counter-examples, though there may be such counter-examples in the end.[75]

I want to close this subsection by making one further point, which is that the contingency of authority (as described by Nguyen) seems, in fact, to be rather platitudinous. As Matthew Parrott points out, "[d]iscerning what a speaker is expressing in a particular speech act requires a listener to have developed an auditory recognitional capacity that is sensitive to that type of expression", and even "listeners who have developed the necessary capacities will be liable to error" about what a given speech act expresses (2015, p. 2229). With this in mind, let us assume for the moment that Neo-Expressivism is false as an explanation of authority. Even so, Parrott's point remains cogent: for *H* to justifiably and routinely defer to *S*, *H* must surely have *some* sort of grip on what mental states *S*'s self-ascriptions express, even if these are just *self-beliefs* that *H* takes to be the products of reliable introspection (contra Neo-Expressivism). But is it *necessary* that *H* takes *S*'s self-ascriptions in any particular way? The very fact that *H* must develop the requisite auditory recognitional capacities seems to ensure that there is a measure of contingency here. This may be no objection if we are not supposing authority to be necessary in the sense at issue. But if it is an objection, then my point here is that Neo-Expressivism is not at a special disadvantage, since *any* account of authority that requires *H* to take *S*'s self-ascription in a certain way (as I think *any* account of authority must admit) will be vulnerable to it.[76]

---

[75] In the above, I have not challenged the claim that we necessarily have a privileged form of *knowledge* of our own mental states. That being said, it might be possible to retool some of the above considerations to attack this claim as well (*pace* Gertler 2000).

[76] I believe that a reply much like the one offered to Nguyen in this subsection can be applied to a similar objection raised by Coliva (2016, p. 158). Her concern is that avowals cannot directly manifest the mental states they ascribe, because we can imagine an alien who visits a world partly comprised of humans and partly comprised of zombies, but who all behave alike verbally and non-verbally. At this world, the alien would likely fail to distinguish between expressive and non-expressive verbal behaviours. Perhaps such a world is possible. But if it is, this does not mean that avowals cannot directly manifest mental states to their hearers, so long as the hearer is attuned to their expressive features and is in possession of no defeaters. It may be, however, that she lacks knowledge of her perceptual knowledge (though see Pritchard 2013 for a workaround). See also Williamson (1995, p. 537) for the following rejoinder to this sort of skeptical objection: "Since making the discrimination [between the zombie and

§2.7.4—Objection to Neo-Expressivism: Negative Avowals

Some philosophers have worried that Neo-Expressivism fails to explain the authority of *negative* avowals. A negative avowal is an avowal like "I am not hungry" or "I don't believe that magic is real." It is intuitive to think that at least some negative avowals are authoritative. But this seems to pose a problem for Neo-Expressivists, since negative avowals might strike one as avowals where one *lacks* any mental state to express (in the action sense).

The solution to this problem, Bar-On argues, is to recognize that negative avowals *do* action-express mental states, albeit ones that are "complementary" (2004, p. 334) to the absent states of mind. Thus, when an agent avows "I don't believe in magic" she can be understood in one of at least two ways. For some cases, she can be understood as expressing her belief *that magic is not real*. In other cases, she can be understood as avowing *agnosticism* about the reality of magic (Bar-On 2004, pp. 334-335). In both cases there is a mental state available for expression. A similar analysis follows for non-propositional mental states. For example, when one avows that one is not hungry, one speaks from a satiated state (2004, p. 334, fn. 31).[77]

Recently, however, Coliva has argued that there are cases that escape Bar-On's proposed fix, for "there seem to be cases in which the negative avowal is not based on enjoying a different, incompatible mental state" (2016, p. 155). She offers the following example:

> Suppose someone is being tortured and screams "I don't want this". It would be weird to say that this is not an avowal but a judgement based on introspection. Yet it would

human] is a matter of knowing that one's present situation is not the bad one…the claim that one cannot make it is tantamount to the sceptic's conclusion. The claim is not available to the sceptic as a premise."

[77] Christopher Campbell raises an interesting variant of the hunger case. "You ask me whether I'm hungry and I say 'No, I'm not'—after all, I haven't been feeling hungry; indeed I haven't been thinking about food at all (though it's been a couple of hours since I've eaten)." He then asks: "Must I nevertheless have been *feeling* 'satiation' in order to be able to 'avow' that I'm not hungry; must I be interpreted as positively avowing that feeling of satiation?" My response is that, in such a case, we can potentially understand the agent as engaging in the sort of 'reflection' that Bar-On describes, wherein she clears her mind so that her satiated feeling can "come to the surface" and thus serve as the complementary state that her avowal "I'm not hungry" action-expresses.

> be equally weird to hold that it is based on some other positive mental state like wanting something other than what one is being inflicted. (2016, p. 155, fn. 21)

I confess that I have a hard time seeing what is so troublesome about this case. I agree with Coliva that "I don't want this" does not express my desire for something other than what is being done *if* that means something like *wanting to feel a cool summer's breeze* instead of pain, or *wanting to be at home* instead of being in a torture chamber. Still, I do not see why we could not interpret the case as one in which the agent expresses her desire *for the torture to stop*.

Another point is worth making, even though I don't think it applies so well to this particular case. It may be observed that claims, e.g., about not wanting something are at least sometimes bad candidates for being authoritative, such that Neo-Expressivists need not accommodate them. This is because authority is a property of avowals, which are a distinctive class of mental state self-ascriptions, but at least some cases of saying that one is not in a mental state could turn out to be cases of *not avowing*. This point may not always be appreciated since, surely, I express *some sort* of mental state by saying that I am not in some mental state. Plausibly, however, what is expressed in at least some of these cases is a belief about the absence of some mental state. In these cases, if I am not actually avowing that belief, then it may be that I am not avowing any mental state.[78]

§2.7.5—Objection to Neo-Expressivism: No False Yet Sincere Expressive Avowals

Neo-Expressivists have sometimes disagreed about whether there can be sincere, expressive avowals that are false. For example, Jacobsen argues that an avowal, if sincere, expresses the very mental state it self-ascribes. On the other hand, Bar-On offers a case where, on her view, sincerity does not suffice for truth in avowing. The case involves a dental patient who, as the

[78] I am indebted, again to a conversation with Jacobsen for this suggestion.

dentist's drill approaches her mouth, avows "It hurts!". Her avowal is sincere, but it does not express pain since the drill has not actually touched her mouth yet. Coliva argues that such cases cannot possibly be genuine avowals:

> For, if it is constitutive of avowals that they are taken to be elicited from the mental state avowed, it should be said that there was no real avowal in the first place, when there is no corresponding mental state. Even more so if one holds, as Bar-On does, that the [avowal] is not merely… the symptom […] of the mental state but somehow embodies it. For, in the case of false (yet sincere) avowals, no mental state is present to elicit and be embodied in one's words. If there is none, then one's words may resemble avowals, but they are not. They are, at best, an *attempt* at avowing a mental state. (2016, p. 159)

I think that Coliva's interpretation of Bar-On goes somewhat adrift: for Bar-On, what is constitutive of avowals, properly understood, is *not* that they are necessarily issued from the very mental state avowed, but that they express first-order states (rather than or in addition to self-beliefs). It does not follow that we cannot understand false yet sincere avowals *as avowals*. The only issue is that what they action-express and semantically-express do not match in these cases.

However, there is another interesting problem in this vicinity that is worth addressing. Here, again, is Coliva:

> …it is true that *we* may take the failed avowal of pain as a symptom of the subject's fear of the dentist. However, if we construe this as meaning that she has managed to express a mental state of hers, though not the one actually avowed, this raises the question of how come the subject has used a self-ascription of pain, say, to give vent to a mental state of fear of which, clearly, she is not aware *as such*. Thus, the idea that there is a mental state M that gives rise to a self-ascription of a mental state N suggests that the subject has wrongly recognized her own mental state—that she has after all mis-taken M for N—contrary to what Bar-On wants to maintain. (2016, p. 159).

As we have seen, Neo-Expressivists want to avoid treating avowals as the upshots of recognizing one's mental states, but Coliva's take is that the only way to explain a mistaken avowal of the above sort is to posit an avower's misrecognizing her mental state.

In reply, here is an alternative take: what explains the subject's pain-avowal, despite its merely expressing a fear, is that her fear of the drill is (1) sufficiently intense so as to cause her to misspeak, where (2) her misspeaking has the particular content that it does because her fear has pain *as its intentional object*. In other words, what explains the mistaken avowal is that the agent is predicating pain-causing capacities of the drill, and is sufficiently fearful of being in pain that she 'skips ahead' to self-ascribing pain rather than the fear. Crucially, there is no introspective mistake here. She does not misrecognize her mental state; instead, her mistake is due to the cognitive duress that she suffers as a result of her outward-directed fear.[79]

Another reply involves reinterpreting the avowal's content. For perhaps, *despite* the choice of words the subject uses, we can *correctly* regard her as avowing a fear. Or, in a more nuanced fashion, we can regard her utterance as expressing a mixture of feigned pain and sincere fear (Jacobsen, ms.). The right interpretation will depend, no doubt, on the details of the case, but that is partly the idea: the project of interpreting another's utterances often affords a great deal of latitude on the interpreter's part.

§2.7.6—Objection to Neo-Expressivism: Avowing and Self-Knowledge[1]

Parrott (2015) correctly acknowledges that, for Neo-Expressivists, authoritative avowals directly action-express the very first-order mental states semantically-expressed. He represents this feature of Neo-Expressivism via the following principle:

[79] Coliva also takes issue with Bar-On's saying, of the dentist case, that "though the subject has successfully expressed *pain*, she has not succeeded in expressing *her* pain" (2004, p. 323). Thus, Coliva argues that: "[s]aying that the subject nevertheless succeeds in expressing a mental state, though not her mental state, is very confusing. Surely, she did not express someone else's mental state" (Coliva 2016, p. 159). However, this is a straightforward confusion about the *sense* of expression that Bar-On has in mind here. As Bar-On states in an earlier paper: "expressing regret, say, in the first two [i.e., action and causal] senses requires expressing *one's* regret, which is factive…Expressing regret in the semantic sense does not require expressing *one's* regret" (2000, p. 17). Bar-On is referring to the fact that the subject succeeds in *semantically* expressing pain, albeit not *her* pain because she does not also causally-express or action-express it.

> *Matching*: For any speaker a, a's assertion ''I am M'' (1) expresses her underlying psychological attitude M and (2) is true iff (a)M. (2015, p. 2223)

Now, as we have seen, there are some complications in Neo-Expressivist scholarship about whether avowals should be understood as assertions,[80] but set this complication aside. Two other issues with *Matching*, for our purposes, are (1) that Neo-Expressivism is intended to cover avowals of all stripes, not just avowals of attitudes, and (2) that Neo-Expressivists need not hold that *all* avowals are such that they action-express and semantically-express the same mental states (as we have seen with negative avowals and Bar-On's dentist case). So let us reformulate Parrott's thesis as follows (I also adopt *S* for speaker and ϕ for mental state):

> *Matching**: For any speaker *S Matching**: For any speaker *S*, *S*'s avowal "I am in ϕ" (1) generally expresses her underlying mental state ϕ and (2) is true iff *S* is in ϕ.

Parrott's concerns for Neo-Expressivism apply equally to *Matching* and *Matching**. So I will continue as if Parrott's target is *Matching**, even though he never actually articulates it. Now, Neo-Expressivism may indeed be committed to *Matching**.[81] However, Parrott argues that Neo-Expressivists face two big questions here: (1) what explains why *Matching** is true of avowals, and (2) what explains why hearers are sensitive to those cases where *Matching** holds?

At some points, Parrott is concerned with (2). For he notes that, while it may be true that listeners are attuned to those cases where *Matching* holds, "certainly this is what Neo-Expressivism should explain."[82] However, immediately after phrasing his concern this way, he asks why *Matching* is true "only when a speaker relates to her attitudes in a first-personal way?". This is clearly a way of phrasing (1). These are different questions, and so it might strike one that an answer to one need not bear on an answer to the other. However, Parrott's view seems to be

[80] Jacobsen (1996), Bar-On (2004).
[81] I say 'may' because Neo-Expressivists may argue that we merely *rationally take Matching** to be generally true.
[82] Parrott takes this worry from Brueckner (2011). See also Owens (2007) and Byrne (2018, p. 72).

that, for Neo-Expressivists, a hearer's justification to presume the truth of a speaker's avowal depends on her sensitivity to when and where *Matching** in fact holds. So, if *Matching** is false, then listeners won't be attuned to its truth. Accordingly, *Matching**'s falsity would undermine the justification agents have to defer to avowals, and so an answer to (1) is directly relevant to (2). This, I take it, is why Parrott ends up focusing primarily on (1).

I have already tried to speak to both questions in §2.7.1, and so my intention is not to rehearse all of this again. Instead, I want to address one particular issue that I did not take up in §2.7.1. The issue concerns a way of possibly explaining *Matching** that Parrott himself identifies. This is the possibility that agents themselves make *Matching** true by intentionally expressing their mental states through avowals that semantically represent them. However, Parrott notes that this strategy seems to require that the speaker has antecedent epistemic access to her attitude, for in order to intend to match the semantic content of one's avowal to the mental state one's avowal action-expresses, one must first know that one has the relevant mental state.[83] And this seems to contradict to the Neo-Expressivist's contention that avowals directly express their first-order objects without any antecedent recognition of one's mental state. Indeed, this accords with Bar-On's insistence that, in avowing, one "gives spontaneous expression to a present state of hers *by* performing some intentional act…that doesn't have expression as its intentional aim" (2010, p. 56). But is it possible, for at least a class of avowals, that Parrott's worry about self-knowledge problematic?

To see where I am going, note that many epistemologists have argued that a key difference between first- and third-personally relating to one's mental states is that in the third-personal case one is *alienated* from one's mental states, whereas one is not alienated from them in the

[83] See also Doyle (2015, p. 69).

first-personal case (Moran 2001; Finkelstein 2003; Boyle 2011). In the Neo-Expressivist discourse, Finkelstein invokes talk of alienation to illuminate the difference between consciousness *of* one's mental states and conscious*ly* expressing them. But sometimes the difference between alienated and non-alienated mentality is cashed out as a difference between kinds of self-knowledge. Keeping this in mind, Parrott himself has elsewhere written—focusing on avowals of *attitudes* like beliefs, desires, and intentions—that "[i]f my way of self-attributing beliefs [and other attitudes] rested entirely on third-personal ways of knowing, then it would mean that, from my own perspective, my belief that *p* might depend on something other than what I regard as adequate reasons for holding it" (2017, p. 10). The implication, if I understand him aright, is that third-personal self-knowledge would be alienated in a way that 'first-personal self-knowledge' is not. Here is one way to understand this idea: my self-ascription, based as it is on my mere recognition of my attitude, does not constitute a self-perspective from which I ratify or *commit* to my attitude.

Without actually giving a more detailed account of how we can have non-alienated self-knowledge of our attitudes, the basic idea I want to explore now is that some ways of knowing one's attitudes may enable one to express one's attitudes themselves through one's avowals, instead of *merely* expressing one's self-beliefs *about* them. For, in having a self-belief about my attitude that has the formal character of *committing* to that attitude, my subsequent avowal of that attitude can plausibly be seen as an act of expressing *my first-order attitude in a self-conscious way*. This, I suggest, is because self-consciously committing to a first-order attitude is sufficient (ceteris paribus) for being able to speak from the perspective of the attitude endorsed: it is sufficient for speaking from a first-order state in a dual-order way. If this is right, then even if you need self-knowledge in order to secure the truth of *Matching** in some cases, your self-

knowledge might have the sort of character that enables you to express your attitude itself in the avowing act. On this this line of thought, what is crucial is not that no self-knowledge precede one's avowal, but that one's self-knowledge is not the result of a *mere recognition* of one's attitude. In this way, we can avoid—as Bar-On and others want to—the idea that avowals are products of introspectively *detecting* one's mental states, where detecting *merely* means *discovering that you have some attitude*, without holding that self-knowledge is *always* an impediment to expressing the very attitude one's avowal semantically represents.[84] This will surely count as a somewhat revisionary Neo-Expressivist response, of course, since Neo-Expressivists often insist on the basic, epistemically unmediated spontaneity of avowals.

However, if we introduce self-knowledge into the picture, one might ask whether we should appeal to features of this knowledge in order to explain authority, or whether we should continue to cash out authority in terms of the expressive character of avowals. From my perspective, there is no need to abandon the Neo-Expressivist account of authority: accepting the existence of a distinctive (i.e., non-alienated, first-personal) form of self-knowledge hardly *forces* or even *encourages* us to abandon the Neo-Expressivist's non-epistemic explanation of authority. Instead, we can simply allow that self-knowledge is part of the *aetiology* of avowing, without treating it as important to the *authority* of avowing. By this I mean that *H*'s reason to presume the truth of *S*'s avowal, or to take it as relatively indubitable, may well still track *H*'s attunement to the fact that *S*'s avowal's action-expresses what it semantically-expresses, rather than tracking the epistemic credentials of *S*'s self-belief about the state she avows.[85]

[84] I do not go as far as Samoilova (2015), who thinks that we can directly express our mental states through avowals no matter what kind of self-knowledge precedes them.

[85] In Bar-On's words: "The [Neo-Expressivist] account needs to insist only that avowals' *distinctive security* derives from the fact that they serve to [action-express] the subjects' self-ascribed conditions, *rather than* from whatever epistemic security accrues to any self-judgments avowing subjects may action-express." (2004, p. 366)

Indeed, I believe we have positive reason to see things this way. For it is crucial that we make sense not only of why *H* takes *S*'s avowals to be authoritative, but why *H* does *not* take *all* of *S*'s self-ascriptions to be authoritative—i.e., those that *S* issues from a third-person point of view. To explain *H*'s competence for filtering these cases out, it is eminently plausible that we appeal to *H*'s tracking of a feature expressed by avowals that is not expressed by third-personal self-ascriptions. Neo-Expressivism offers just such an account: avowals and third-personal self-ascriptions express different states—one expresses first-order attitudes (and perhaps also self-beliefs, on the present line of thought) while the other expresses only self-beliefs.

But why not argue that *H* tracks differences in the security of *S*'s self-beliefs across the two cases, and argue in turn that this explains why *H* defers to some but not all of *S*'s self-ascriptions? I do not want to say that a view like this *cannot* be argued for, but I do find it implausible: we have no reason to credit *H* with a capacity to discriminate between different degrees of epistemic security between *S*'s self-beliefs in the first-personal avowing case versus the third-personal self-ascriptive report case. One reason is that it is unclear whether avowals can express a property like high epistemic security, and a second reason is that it is equally unclear whether *H* is attuned to the expression of such security when she understands *S*'s avowals. This is not to say that *H* cannot be, say, *entitled to assume* that *S*'s self-beliefs are especially epistemically secure when *S* issues an avowal. But it is not plausible that this is something *H* knows in virtue of its being expressed. And so it is not bound to be very helpful if we are trying to account for the fact that *H* does not treat *all* self-ascriptions as authoritative, since proceeding on the basis of this assumption would be a *general* habit of *H*'s.

Again, this entire line of thought is highly tentative and exploratory. As such, if one finds it contentious, one can still fall back on the broadly developmental-psychological account that I offer in §2.7.1 in order to answer Parrott's concerns about *Matching**.

§2.7.7—Objection to Neo-Expressivism: Avowing and Self-Knowledge$_2$

While many Neo-Expressivists would reject the (partial) explanation of *Matching** that I advanced in the previous subsection, it is striking that fewer Neo-Expressivists deny that avowals are or can be sites of self-knowledge. Thus, Jacobsen concedes that there is nothing "unintelligible" about according a "dual role" to avowals (1997a, p. 423), whereby they express both first-order states and second-order, knowledgeable self-beliefs about those states. Likewise, although Bar-On (2001) once rejected the view that avowing subjects thereby have self-knowledge of their mental states, she eventually came around to admitting that it would be "startling" (2004, p. 353) if it we do not actually have self-knowledge of the first-order mental states we express in avowing. How can this be? The basic point, again, is that Neo-Expressivists mostly worry about the idea that self-knowledge *precedes* the avowing act (this being the bone of contention in the previous subsection), but that they don't worry about self-knowledge *accompanying* avowals. Here, for example, is Bar-On:

> Spontaneously grunting or smiling, stomping one's foot, making a rude gesture, giving a hug, saying: 'What a mess' are typically not acts performed with a prior intention or with a specific goal or purpose in mind. Nonetheless, they are acts that meet at least one widely received test for intentional action—Elizabeth Anscombe's. They are voluntarily produced bits of behavior, where the person producing them knows what she's doing and that knowledge is *nonobservational*…I see no reason to single out acts of spontaneously producing self-ascriptions— avowals such as 'I'd like some tea' or 'I'm wondering whether I should go now'– as necessarily having belief-desire pairs as their reasons. (2010, p. 57).

Bar-On admits that avowals are intentional actions, even though "the reasons [for avowals] are not beliefs or thoughts we have about the expressed mental states but rather the states

themselves—I gave a hug because I felt happy to see you; I sighed because I felt exasperated" (Ibid., p. 57). In light of this admission, her point is that—for Anscombian reasons—avowals can still involve self-knowledge, even if they are not antecedently backed by self-knowledge (cf. Roessler 2015).[86] This is fortunate, since it does not require us to think of avowals as mere outbursts like what a parrot might make, in which the avowing agent does not understand what she says (cf. Boyle 2009, 2010). In allowing that agents understand what they avow, part of what they understand is that they are in the mental state avowed. The point is just that this understanding does not hinge on successful exercises of introspection, nor on possessing any non-introspective form of self-knowledge that precedes one's avowals.

One worry is that, if we allow that avowals are always accompanied by self-knowledge, the agent must take herself to be *justified* in avowing. After all, if she knows that she is expressing her mental state through her avowing act, she knows that she is making a truth-evaluable claim about herself. In reply, however, Neo-Expressivists can argue that "[t]he distinctive perspective of an avowing person…is not an epistemic perspective at all" (2010, 56), which is to say that the avowing subject need not understand herself "as being justified *or* unjustified" in avowing her mental state, even though she understands herself as avowing it.[87]

§2.7.8—Objection to Neo-Expressivism: Avowing and Self-Knowledge$_3$

Now, part of the story so far is that, if avowals are sites of self-knowledge, this must be a fairly distinctive kind of self-knowledge—knowledge that is non-observational, that does not involve having to detect one's mental states by means of some sort of introspective epistemic procedure, and so on. These ideas are commensurate with a common line of thought, which is that much of

[86] See also Bar-On (2012, p. 209).
[87] Perhaps this is the proper way to understand Wittgenstein's claim, as regards avowing, that "[t]o use a word without justification is not to use it without right" (*zu Unrecht*) (1953, §289).

our self-knowledge is *privileged* and *peculiar*. Self-knowledge is privileged, roughly, to the extent that it is more secure than anyone else's knowledge of one's mind. Self-knowledge is peculiar, roughly, because it is acquired by a means (if there be a means) that is available only to the subject herself. It is common to argue that we have such self-knowledge, and it is also common to argue that such self-knowledge is what explains authority. I have already argued that Neo-Expressivists do not need to accept this *epistemic strategy* for explaining authority, even if they accept that we have such knowledge. But if we really do possess such self-knowledge,[88] one might now join Matthew Chrisman in raising the following concern for Neo-Expressivism:

> …surely, once we've admitted that they [self-beliefs] exist and are distinctively secure, the security of these beliefs is something that needs to be explained; and it would be strange if whatever explains it isn't intimately related to what explains the security of avowals that express them…for the avowal expressivist even to engage the project of explaining the special security of these first-personal present-tense beliefs about our own mental states is, it seems to me, for her to give up on the primary advantages of the position in the debate with the introspectionist. (2009, pp. 8-9).

Now, while this passage may sound like a restatement of the thought that the epistemic strategy now threatens to overtake the non-epistemic Neo-Expressivist strategy for explaining authority, I take this passage to also express an additional concern. This is the concern that granting the existence of privileged and peculiar self-knowledge while also embracing Neo-Expressivism will eventually require us to say *something* about how these are related, seeing as the authority of avowals can seem like a mirror image of our privileged and peculiar self-knowledge, even if the relationship is not one where privileged and peculiar self-knowledge explains authority. In other words, *Chrisman's concern* for the Neo-Expressivist is that she has yet to answer:

(1) How do privileged and peculiar self-knowledge and authority relate, if not by one explaining the other?

[88] See the exchange between Vega-Encabo (2011) and Finkelstein (2011) for one sort of debate over whether Neo-Expressivists should understand avowals as involving any form of self-knowledge.

Moreover, I believe there is another question nearby, namely:

(2) If the privileged and peculiar self-knowledge agents have of their mental states does not explain authority, what *does* privileged and peculiar self-knowledge explain?

To be sure, one way to go is to deny that self-knowledge is privileged and peculiar. Thus, despite Jacobsen's admission that there is nothing unintelligible about according a "dual-expressive" role to avowals (1997a), according to which they express both first-order states and second-order self-beliefs (self-beliefs that amount to privileged and peculiar self-knowledge), he still seems to prefer a more deflationary perspective:

> Our problem is to explain how we are able to self-attribute mental states with groundless authority, not to explain how our ability to do that qualifies as knowledge. If we can successfully account for that ability and do so without portraying it as a species of knowledge, then we are free to conclude that what we have explained is not self-*knowledge*. (Jacobsen, ms.)

This reply may sound tempting, but it is not so easy as all that. The trouble is that many authors have argued that we have transcendental assurances in the existence of privileged and peculiar self-knowledge. If any of these views hit the mark, Neo-Expressivists may be able to hold onto their explanation of authority, but for all this we won't have any answers to questions (1)-(2) above. It is important, then, that we say more than Jacobsen does. It is important, in other words, that we evaluate transcendental accounts of privileged and peculiar self-knowledge.

§2.8.1—Preamble to Chapters Three Through Five

In Chapters Three and Four I will be considering several transcendental accounts of privileged and peculiar self-knowledge. Each account, in its own way, argues that privileged self-knowledge is in some sense indispensable to our rationality or our agency, at least where we have sufficiently developed human agents in mind. Importantly, however, each account will be somewhat narrow in its explanatory ambitions, for each will only concern—roughly—privileged

and peculiar self-knowledge of propositional attitudes. This means that they frequently leave aside any discussion of the cognitive indispensability of self-knowledge of our conscious experiences, sensations, emotions, moods, and so on.

It may seem disappointing, then, that whatever result these accounts may yield for addressing questions (1)-(2) above will only be partial, seeing as they will not deliver resources for answering those questions as regards all kinds of mental states to which we can bear privileged and peculiar self-knowledge or Neo-Expressive authority. There are justifications, however, for reining in the ambitiousness of what follows. First, it is not obvious that privileged and peculiar self-knowledge of different kinds of mental states has uniform roles in cognition (indeed, the accounts we will consider often explicitly deny this). For this reason, it cannot be simply assumed that questions (1)-(2) can be approached in a uniform fashion. The second reason is simply that the literature transcendental accounts of privileged and peculiar self-knowledge is vast, and the fact that it will take us two chapters just to evaluate a range of accounts addressing such knowledge of a subset of mental state types means that things would quickly spiral out of control if a more exhaustive approach were taken in this dissertation.

In Chapter Three we will reach a largely negative result: the transcendental accounts of privileged and peculiar self-knowledge will be met with a battery of criticisms. It is only in Chapter Four that I will develop my preferred account. In doing so, I will also try to respond to Chrisman's concern (question (1) above) as well as question (2), which I noted is a natural corollary to Chrisman's concern. Rather than conceding that privileged and peculiar self-knowledge explains our authority, I will argue that our Neo-Expressively understood authority and our privileged and peculiar self-knowledge (again, of a range of mental states) jointly contribute to explaining a *further* capacity of ours.

Chapter Three—Agentialist Self-Knowledge: Part One

§3.1.1—Introduction

Many philosophers think that there are significant epistemic asymmetries between how we know our own mental states, on the one hand, and how we know the mental states of other people and the wider external world, on the other. For starters, it is common orthodoxy to treat much of our self-knowledge as *privileged*, meaning—roughly—that "beliefs about one's mental states acquired through the usual route are more likely to amount to knowledge than beliefs about others' mental states (and, more generally, corresponding beliefs about one's environment)" (Byrne 2018, p. 5).[89] This claim about the privileged character of self-knowledge is often followed up by a claim about its *peculiarity*: that it is known in a manner available only to the agent herself (McKinsey 1991). For, if self-knowledge is achieved by ordinary perceptual, inferential, or testimonial means, then it is hard to see how it can enjoy a distinctively high-grade epistemic status.

Strictly speaking, however, privilege and peculiarity do not entail one another. In theory, ordinary routes to empirical knowledge might be more effective in the case of securing self-knowledge than in securing knowledge of other minds or the wider external world. Alternatively, one might have a peculiar route to self-knowledge that is not especially "likely to amount to knowledge". Moreover, even if privilege and peculiarity do typically or always come together as a matter of fact, epistemologists rarely assert that *all* of our self-knowledge is privileged and peculiar. Rather, it is *ordinarily* so, for suitably cognitively developed subjects. It remains plausible that we must sometimes gain self-knowledge through highly fallible, ordinary

[89] Examples of this sort of talk abound in the literature. As Smithies and Stoljar put it, "introspection is better than other ways of knowing about the world in certain epistemological regards…it is more reliable, or it is immune from certain types of ignorance and error" (2012, p. 6), where talk of introspection need only be understood as a placeholder for 'method by which we acquire self-knowledge' of a privileged and peculiar sort.

measures—talk therapy or self-interpretation, for instance. When we know our minds in these ways, we are thought to lack both privileged and peculiar self-knowledge. Some authors even go so far as to see that we are inevitably alienated from ourselves in these cases (Moran 2001). But these are said to be aberrations; in the ordinary case, we have privileged and peculiar self-knowledge.

In this chapter and the next my primary focus will be on accounts of self-knowledge that offer transcendental assurances of its privilege and peculiarity in at least some cases; they will purport to explain why we *must* have privileged and peculiar access to our mental states. These accounts do not treat privileged and peculiar self-knowledge indispensable for agents like us (e.g., Cassam 2015), or as potentially epiphenomenal (Wilson 2009), or as (at least sometimes) *maladaptive* (Flanagan 1992; Nguyen 2015[90]). Moreover, in referring to the indispensability of such knowledge, I am not here addressing its potential *value* from the first person point of view.[91] From the first-person point of view, having such self-knowledge makes one happy (Nguyen Ibid.), or one might value it because it conduces to a clearer self-conception (Moran 2001; Doyle 2018b). When I speak about the indispensability of privileged and peculiar self-knowledge, I am referring to its *cognitive* or *agential* indispensability: I am referring to the ways that such self-knowledge enables us to be the kinds of cognizers and agents we are, regardless of whether it is explicitly valued, for any reason, by any particular agent.

As I mentioned in §2.8.1, these transcendental accounts typically focus on privileged and peculiar self-knowledge of certain kinds of mental states—propositional attitudes like belief, desire, and intention, though also potentially hopes, wishes, and more. It is sometimes thought

[90] Nguyen sees self-knowledge as adaptive in *some* contexts.
[91] See Jongepier (2020) for a recent critical discussion of the value of a specific kind of privileged and peculiar self-knowledge.

that such accounts should or do restrict themselves even further, such that they only make claims about the experiential or occurrent manifestations of one's attitudes, these being mental acts like *judging that p,* for instance.[92] However, we will see that many authors take a much more ambitious view, opting to try and secure transcendental grounds for privileged and peculiar self-knowledge of our standing attitudes, rather than (just) their occurrent mental manifestations or associated mental acts (such as judgements).

Occasionally, in examining transcendental arguments for privileged and peculiar self-knowledge, proposals about how we acquire such knowledge will also be presented by their authors. But I will not tend to dwell on any of these proposals (since this 'how question' will be taken up in detail in Chapter Five). Accordingly, I will follow Ben Sorgiovanni (2019) in introducing some further taxonomical jargon: my chief focus will be on *non-substantive* as opposed to *substantive* accounts of privileged and peculiar self-knowledge. Non-substantive accounts are transcendental and so purport only to tell us *that* we have such self-knowledge. It is only once we consider substantive accounts—accounts that aim to explain *how* we have such self-knowledge—that we will be on a path toward a fuller account of privileged and peculiar self-knowledge of our propositional attitudes.

One more distinction is in order. All of the accounts considered in this chapter and the next can be referred to as *agentialist* or *rationalist* accounts of self-knowledge. These labels capture the fact that the transcendental arguments they present appeal to putative facts about the indispensability of privileged and peculiar self-knowledge to our agency, rationality, or both. There will be occasions where it is important to use one or the other label.[93] For example: we will see that Sydney Shoemaker's non-substantive account of privileged and peculiar self-

[92] See Gertler (2011a).
[93] In this I am following Gertler (2016).

knowledge is best construed as a rationalist account, since he transcendentally grounds privileged and peculiar self-knowledge in our rationality rather than in anything we *do* with our rationality. Contrariwise, the majority of non-substantive accounts of privileged and peculiar self-knowledge make claims about the indispensability of such self-knowledge to our agency—something that such self-knowledge allows us to do that we could otherwise not. So, I will tend to use the term 'agentialism' to refer to these non-substantive accounts, though I will occasionally flag my decision to use the term 'rationalism' when it is more fitting.

Here, then, is the agenda for what follows. In §3.2 I reconstruct Tyler Burge's (1996) highly influential agentialist account. On his account, privileged and peculiar self-knowledge of our propositional attitudes is transcendentally grounded in our capacity to engage in a certain kind of mind-directed reasoning that is itself indispensable to our rational agency. In §3.3 I offer up many criticisms of Burge's account. In §3.4 I consider a more recent agentialist account, one that I develop on behalf of what I take to be a growing consensus among philosophers who have recently argued in favour of a so-called "Taking Condition" on reasoning *generally*, rather than critical reasoning specifically (Boghossian 2014, p. 2). I consider several arguments for the Taking Condition and argue that they do not motivate a non-substantive account of privileged and peculiar self-knowledge. In §3.5 I turn to a rationalist account developed by Sydney Shoemaker (1994, 1996). On his account, basic facts about our rationality secure a transcendental ground for privileged and peculiar self-knowledge. I criticize this account in §3.6. In §3.7 I consider Matthew Parrott's Shoemaker-inspired rationalism, and I criticize it in §3.8. In §3.9 I consider one more agentialist account, due to Jared Peterson (2020), and I criticize it in §3.10. Having unearthed difficulties with each of these agentialist/rationalist accounts, I offer a brief preamble to Chapter Four in §3.11.

§3.2.1—Burgean Agentialism: Critical Reasoning[94]

In critical reasoning, agents aim to conform their attitudes to rational norms and to evaluate the quality of their reasoning itself. As Burge puts it, critical reasoning is reasoning "guided by an appreciation, use, and assessment of reasons and reasoning as such" (1996, 98).[95] We critically reason, Burge says, in the course of "…giving a proof, in thinking through a plan, in constructing a theory, in engaging in a debate," and so on (99). These are ubiquitous cognitive activities, at least among most adult human agents. No wonder, then, that Burge (2013) takes a capacity for critical reasoning as partly constitutive of fully developed human personhood.

Here is a toy example. Imagine that Evelyn is concerned about recent trends in her country's political discourse. In search of an explanation, she works her way into the following thought process. First, she notices that her belief *p* (*that her country is safe from fascist takeover*) squares poorly with her recently acquired and evidentially well-supported belief *q* (*that fascist rhetoric is becoming more popular in her country's political discourse*) and her further, equally evidentially well-supported belief *r* (*that fascist movements are quickly gaining traction in nearby countries*). She acknowledges that she is rationally required to believe in accordance with her evidence, and eventually concludes that she ought not to believe *p*. So long as her self-belief that she ought not to believe *p* is warranted, it exerts epistemic pressure on her to disbelieve *p*. In Burge's words, "justifiably finding one's reasons invalid or one's thoughts unjustified, is normally *in itself* a paradigmatic [epistemic] reason…to alter them" (1996, 110).[96] *Ceteris*

[94] With permission from *Res Philosophica*, much of the material in §3.2 is taken—with some small modifications—from Winokur (2021).
[95] Paul and Elder describe it, similarly, as "self-directed, self-disciplined, self-monitored, and self-corrective thinking" about one's own attitudes (2008, p. 2).
[96] The relevant reasons are *prima facie* reasons. For reasons of readability, I will not staple this qualifier to each remark about our critical self-beliefs.

paribus, Evelyn's *p*-belief will be excised from her psychology once she draws this conclusion.[97] In this way, critical reasoning enhances her first-order rationality.

§3.2.2—Burgean Agentialism: Privileged and Peculiar Self-Knowledge

Of course, the idea that higher-order reflection can improve our lower-level rationality is not a uniquely Burgean idea.[98] Burge's real contribution to our understanding of critical reasoning is an argument for what sort of self-knowledge one must have if critical reasoning is to be reasonable in the first place. His thesis is that the perspective *from which* we critically reason and the perspective *on which* we critically reason must enjoy an "immediate rationally necessary connection" (1996, 109-110). Roughly, this means having critical self-beliefs that are true whenever warranted, such that they count as self-knowledge whenever they are warranted. Thus, when Evelyn critically reasons and concludes that she ought not to believe that *p*, her self-belief that she ought not to believe *p* any longer is warranted only if she really ought not to believe that *p* any longer, all things considered.

I will elaborate on this picture in two broad steps. First, I will clarify the nature of this immediate rationally necessary connection between higher- and lower-order perspectives and explain why certain substantive accounts of self-knowledge—those that explain *how* we have it—cannot produce such a connection. After this, I will clarify why Burge thinks that critical reasoning must be conducted only when such a connection obtains.

To clarify the nature of this connection, Burge points the reader to cases where one has a warranted, higher-order critical belief, but where this critical belief is *not* related in an immediate

[97] Borgoni & Luthra (2017) defend the possibility of epistemic akrasia.
[98] One may describe Descartes's (1641) method of doubt as underpinning a particularly aggressive form of critical reasoning. An influential contemporary champion of the role of critical self-reflection is Sosa (2007), though the details of his view differ from Burge's. Moreover, many have credited McGinn (1982) as anticipating the importance of self-knowledge for something like critical reasoning and, more recently, Knappik (2020) has made a compelling case for the claim that Sellars anticipated something like Burge's account.

rationally necessary way to its object.[99] Paradigm cases are beliefs about *other* people's propositional attitudes. For example, I might believe that Pete believes that the stovetop is hot (I infer this from watching him hesitantly approach it). Next, I judge that he ought not to believe that, because the stove broke down last night. It can transpire, however, that "[I may be] in error about what [Pete's] beliefs are, or [Pete's] perspective may have different associated reasons or background information from mine" (Burge 1996, 109); perhaps, for example, Pete is only hesitant because he believes there is a spider on the stovetop. So he does not actually believe that the stovetop is hot, and my warranted belief that he believes this is false. This is a basic fact about my epistemic relationship to Pete: this dissociation between my beliefs and Pete's entails that there can always be a gap between my warranted beliefs about what he ought to believe and the actual facts about what Pete ought to believe. As such, what I believe about what Pete ought to believe lacks immediate rationally necessary consequences for what Pete ought to believe.

But now note that, if I ascribe a mental state to *myself* on the same basis, I will stand to my mind in a highly similar way as I stand to Pete's. By observing my behaviour, judging that I believe *p*, and inferring that I lack good reason to go on believing *p*, my self-belief can end up warranted yet false, even if I am better at knowing my mind this way than I am at knowing Pete's. Hence, my self-belief does not have immediate rationally necessary consequences for whether I ought to believe *p*. Burge's suggestion is that similar lessons apply to many other ways of forming self-beliefs: if my beliefs about my attitudes are based on Pete's testimony, or on inductive inferences about the likelihood of my believing various things given statistical generalizations about people my age, the same opportunities for dissociation between my warranted self-beliefs and their lower-level objects will obtain. These methods can be

[99] I will focus on critical reasoning about first-order attitudes, as opposed to second- or even higher-order attitudes.

characterized as *third-personal* because they are methods that anybody might use to acquire knowledge of another's mind. Third-personal methods cannot yield self-beliefs that are related to their objects in an immediate rationally necessary way. In order to critically reason, then, I must be able to take up a self-perspective that is essentially epistemically different from another agent's perspective on my attitudes.

In fact, Burge's critical target here encompasses more than third-personal methods for acquiring self-knowledge. Thus, he also considers and rejects the possibility that we acquire self-knowledge through "*sensed* inner goings-on" (104). Take, for example, the view that self-knowledge is delivered by the operations of a scanning mechanism in the brain (Armstrong 1968; Lycan 1996). Such a scanner, if real, could provide self-knowledge that is *peculiar* so long as it is not hooked up to anyone else's brain. We might also stipulate that it is especially reliable and so accounts for the *privileged* status of one's self-knowledge. Nevertheless, Burge rejects this account. This is because it is possible for an inner sense faculty to operate reliably but faultily, and so to deliver warranted false positives. Put differently, these accounts leave open the possibility of "*brute local error*" (Bar-On 2004, 98): errors that are not due to any psychological or epistemic failing on one's part. In the case at hand, one's inner scanner may simply misread its object, even if one is cognitively well-functioning and epistemically responsible. But this means that there is no immediate rationally necessary connection between the self-beliefs produced by one's inner scanner and one's first-order perspective, even if one's inner scanner usually produces warranted and true self-beliefs (Sorgiovanni 2019, 5). This is because, in these cases, one only *happens* to have a warranted yet true self-belief—it could have been otherwise, and so the connection between warrant and truth is not necessary. Let us say, then, that *first-personal self-knowledge* is self-knowledge that is not only privileged and peculiar, but that is

also connected to the mental states self-known in an immediate rationally necessary way. So understood, Burge's argument is that "inner sense" and other "observational" accounts of self-knowledge cannot produce first-personal self-knowledge, even if they can produce privileged and peculiar self-knowledge.

Of course, all of this is interesting only if Burge is right that critical reasoning really does require an immediate rationally necessary connection between one's self-beliefs and their objects. Here, however, one might be skeptical. After all, it can seem like we critically reason in situations where no such connection obtains. For example, I might find myself in some circumstance and come to believe correctly, through a warranted inference about what I typically desire in similar circumstances, that I desire to φ. I might then believe that I ought not to desire this. If I subsequently forfeit my desire on the basis of this belief, which certainly seems possible, my rationality will be enhanced. But because inferences from my observed behaviour to its mental causes can be warranted yet false, there is *not* an immediate rationally necessary connection between my first- and second-order perspectives in this case. The skeptic now asks: why is such reasoning not good enough to count as genuine critical reasoning?

In reply, I think we should follow Sorgiovanni's (2019) interpretation of Burge. On this interpretation, it is *a constitutive norm of critical reasoning* that there must be a necessary, immediate rational connection between one's higher and lower-order perspectives. As evidence for this interpretation, Sorgiovanni cites Burge's claim that "it is constitutive of critical reasoning that if the reasons or assumptions being reviewed are justifiably found wanting by the reviewer, it *rationally follows immediately* that there is a *prima facie* reason for changing or supplementing them" (1996, 109). His reading of Burge's use of 'constitutive' is that it refers to this constitutive norm. According to this norm, having a warranted self-belief that you ought to believe *p* requires

you to believe *p*. The reason why there should be such a norm is precisely that it is only when such a norm is fulfilled that critical reasoning is assured to enhance the rationality of one's attitudes.

To see this, we need only reflect on the fact that critical reasoning *unbound* by such a norm would allow for situations in which a self-belief was warranted yet false, whereupon one would be rationally required to change one's mind in a way that would *not* enhance one's lower-order rationality. It is because such possibilities are antithetical to the reasonability of critical reasoning that the whole enterprise is constitutively structured by a norm that requires us to avoid them. If critical reasoning is to add to the reasonability of our reasoning and attitudes, it must be guided by such a norm, even if only in the self-directed case.[100] This is why, even if critical reasoning that violates this norm is possible, it is not full-blooded critical reasoning. For unless one meets this norm, one's critical self-perspective will not be essentially different from that of someone else's perspective on one's mind, and another agent's beliefs about what attitudes one ought to have do not necessarily bear, all things considered, on what attitudes one ought to have.[101]

§3.2.3—Burgean Agentialism: Uniqueness of Warrant

The final piece of Burge's account is an *epistemic* account of self-knowledge, that is, an account of the warrant we have for our critical self-beliefs, such that they count as self-*knowledge*. Here is Burge:

---

[100] Plausibly, no such norm holds for critically reasoning about other minds (Sorgiovanni 2018, p. 11).

[101] An anonymous reviewer for *Res Philosophica* imagines a case in which one has excellent but misleading evidence that there is poison in one's cup. They ask why this evidence is not a reason "in and of itself" for one to not drink from the cup. The case is supposed to be unsettling because it involves warranted yet false evidential beliefs that still seem to serve as reasons in and of themselves to do something, contra what Burge allows for critical self-beliefs. My response is that, if it is true that one's misleading evidence really can serve as a reason in and of itself (in the relevant sense) to not drink from the cup, this is because one's evidential beliefs are part of the same (first-order) perspective as one's other beliefs about the cup and about what one should do with it. Contrariwise, it is because our first- and second-order perspectives have fundamentally different objects that, for Burge, following a stringent norm is needed to ensure that they are rationally responsive to one another in the right, rationality-enhancing way.

> if one lacked entitlement to judgments about one's attitudes, there could be no norms of reason governing how one ought check, weigh, overturn, confirm reasons or reasoning…If reflection provided no reason-endorsed judgments about the attitudes, the rational connection between the attitudes reflected upon and the reflection would be broken. So reasons could not apply to how the attitudes should be changed, suspended, or confirmed on the basis of reasoning depending on such reflection. But critical reasoning just is reasoning in which norms of reason apply to how attitudes should be affected partly on the basis of reasoning that derives from judgments about one's attitudes. So one must have an epistemic entitlement to one's judgments about one's attitudes. (1996, 101-102)

This argument is transcendental. Because (1) we are critical reasoners, and (2) critical reasoning is itself reasonable, it follows that (3) the self-beliefs that constitute our critical self-perspective must be warranted. We therefore have an "entitlement" to our critical self-beliefs. An entitlement is a species of epistemic warrant that "need not be part of the repertoire of the individual that has the entitlement" (Ibid., 94), meaning that it need not be articulable by its possessor.

In sum, Burge's picture is this. Critical reasoning requires us to have first-personal self-knowledge of our attitudes. First-personal knowledge is privileged and peculiar, but *also* such that it cannot be produced by a method that can yield warranted yet false self-beliefs. When we critically reason in accordance with the right method for acquiring self-beliefs, we are entitled to our self-beliefs and so have (privileged, peculiar, and first-personal) self-knowledge. This entitlement is transcendentally grounded by our status as critical reasoners. So, as critical reasoners, we necessarily possess first-personal (and hence privileged and peculiar) self-knowledge.

§3.3.1—Objection to Burgean Agentialism: Critical Reasoning Without Self-Knowledge

Let us refer to the foregoing account as *Burgean Agentialism*. Perhaps the most frequent objection to Burgean Agentialism is that critical reasoning does *not* require an agent to have privileged and peculiar self-knowledge, let alone what I described above as first-personal self-

knowledge. Indeed, the objection is that critical reasoning does not require self-knowledge of *any* sort. The reason, as Annalisa Coliva puts it, is that "there seems to be no bar to the possibility of conceiving of an agent who engages in belief revision and is at least capable of doing so in connection with devising and executing a practical plan and yet is unable to make the relevant psychological self-ascription[s]" (2016, p. 117).[102] To show this, she cites an example of Christopher Peacocke's:

> Suppose you come home, and see that no car is parked in the driveway. You infer that your spouse is not home yet … Later, you may suddenly remember that your spouse mentioned in the morning that the breaks of the car were faulty, and wonder whether she may have taken the car for repair. At this point, you suspend your original belief that she is not home yet. For you come to realise that the absence of the car is not necessarily good evidence that she is not home. If the car is being repaired, she would have returned by public transport. Then finally you may reach the belief that she is home after all, given your next thought that she would not have taken any risks with faulty breaks. (1998, p. 276)

According to Peacocke, the agent's reasoning in this case is best described as "second-tier" reasoning, for "[i]t involves thought about relations of support, evidence or consequence between contents, as opposed to first-tier thought, which is thought about the world where the thought does not involve any consideration of such relations between contents" (Peacocke 1996, p. 277). However, *ex hypothesi*, second-tier reasoning does not require psychological self-ascriptions, since grasping relations of epistemic support between propositional *contents* is not the same as grasping relations of epistemic support between one's propositional *attitudes*.

Of course, we can have different attitudes toward a given propositional content, and we do not reason from every attitude that has one and the same content. For this reason, proponents of the possibility of second-tier reasoning might have to add that second-tier reasoners "view the bare content [of their attitudes] in a believing [desiring, intending] way" or something similar

[102] Other proponents of this objection are Owens (2011), Cassam (2015, chapters 4 and 15), and Gertler (2016).

(Broome 2019, p. 41), where viewing a content in a believing (desiring, intending) way is not equivalent to self-ascribing an attitude. This may sound strange, but it needn't be. For example, first-order beliefs are contents toward which we harbour believing-attitudes even though, by definition, the contents of these beliefs do not include self-ascriptions. Similarly, the idea here is that second-tier reasoning requires being able to represent propositions and their epistemic support relations in a believing way.

Relatedly, second-tier reasoning presumably should not require conceptualizing propositional contents, or the relations between them, as *reasons for* believing, desire, intending, etc., since this would bring us dangerously close to reasoning with one's own mental states in view. Instead, the epistemic concepts deployed in second-tier reasoning must squarely concern indicators of what is *true* or *good* or *to-be-done*, perhaps via concepts like EVIDENCE or CONSEQUENCE, concepts that one might exercise without also exercising psychological concepts.

On the basis of her acceptance that second-tier reasoning is possible, Coliva argues that the burden is on the Burgean Agentialist to explain why being a second-tier reasoner is not sufficient for being a critical reasoner. Like other proponents of this objection, she does not deny that we sometimes do critically reason with our own attitudes explicitly in view. But this admission hardly conduces to a transcendental connection between critical reasoning and privileged and peculiar (or first-personal) self-knowledge. This is a bad result, for it opens up space for an alternative transcendental argument, one that only appeals to the indispensability of whatever conceptual and cognitive capacities are required for second-tier critical reasoning.

One reply on behalf of the Burgean Agentialist is to argue, in the traditional Anscombian way, that if one is *intentionally* revising one's attitudes through critical reasoning, then one cannot know what one is doing without knowing that one is reflecting on the reasonability of

one's attitudes. As Quassim Cassam points out, however, "…it's not clear why being able to revise your beliefs in this sense is in any sense a condition of being a rational agent" (2014, p. 215). It is not clear, that is, why an agent must intentionally set about to revise her attitudes in order to be a rational agent. Nor, relatedly, is it clear to me why Anscombe's point cannot be cashed out in second-tier terms. This is because it seems to me that there could be different descriptions under which an agent's critical reasoning counts as intentional. Thus, one might imagine an agent who understands herself as *trying to discover her wife's whereabouts* without thereby understanding herself as *trying to rationally adjust her attitudes concerning her wife's whereabouts*. If the former description under which the agent understands her own activity does not require her to think about her own mental states, then the Anscombian defense of Burgean Agentialism fails.

§3.3.2—Objection to Burgean Agentialism: Critical Self-Beliefs as Epistemically Inert

The previous objection segues into what may be an even more pressing objection for Burgean Agentialism, namely, that critical self-beliefs do not exert immediate rational pressure on their lower-level objects at all. One might arrive at a general worry along these lines if one wonders why Burge is so confident that our self-beliefs are justified, such that they can generate the sort of rational pressure required to render critical reasoning reasonable. For, in his transcendental argument for our entitlement to our critical self-beliefs, a crucial premise is that critical reasoning is reasonable. But perhaps it is not and, therefore, it is simply false that we are critical reasoners of the sort Burge thinks we are.[103]

[103] See Blackwood (2010) and Parent (2017, p. 254). See also Kornblith (2012, 2016), who does not focus on Burgean Agentialism specifically but on reflective models of rationality more generally.

One might also object from an *evidentialist* perspective about epistemic reasons. On this view, the only good epistemic reasons for holding our attitudes are evidence in their favour, at least with respect to belief.[104] Thus, to the extent that critical self-beliefs like "it is irrational for me to go on believing *p*" are not themselves evidence that *p* is false, it will be false that these self-beliefs serve as epistemic reasons to hold or forfeit them. But this is what critical self-beliefs need to be like if they are to play their intended role in critical reasoning. And now the worry is that they are not evidence for their objects and hence do not play this role.

No doubt, the Burgean Agentialist should deny this evidentialist move, for she already rejects the view that our *self*-beliefs, being beliefs after all, are only warranted to the extent that they are based on evidence. After all, as we have already seen, Burgean Agentialists argue that our self-beliefs are warranted by a transcendental *entitlement*, not by evidence of their truth. However, this reply does nothing to block a more modest (and more plausible) form of evidentialism. On a more modest evidentialist view, our critical self-beliefs fail to get a rational purchase on their first-order objects because they are not evidence in favour of their first-order objects, even if they are not themselves warranted by evidence.

To see the picture, consider Evelyn again. She reflects on the inconsistency between three of her first-order beliefs *p*, *q*, and *r*, and determines that the evidence for *p* and *q* outweighs the evidence for *r*. She then forms the self-belief that she ought not to believe *r* anymore. Burge's view is that her critical self-belief is an epistemic reason for her to disbelieve *r*. Our question is: is Evelyn's critical self-belief *evidence* against *r*? It does not seem to be; rather, the critical self-

[104] A somewhat related criticism is made by Gertler (2011a, pp. 181-182), who notes that it is not part of our ordinary conception of knowledge that the rational permissibility of having some mental state is governed by states of the agent (i.e., higher-order beliefs about their rationality) rather than states of the world.

belief seems to simply represent to Evelyn that her *first-order evidence* for *r* is insufficient. It is not itself evidence that *r* is false, even if it is warranted.

One might now reply that this objection presupposes that the only relevant form of evidence is *first-order* evidence, and that the objection fails so long as Evelyn's critical self-belief is *second-order* evidence that her first-order evidence in favour of *r* is insufficient. Higher-order evidence is evidence about the quality of one's first-order evidence, and can have direct effects on the reasonability of first-order beliefs that are based on first-order evidence. For instance, when a speaker is invited to give a talk in defense of some claim, *q*, at a prestigious university, this constitutes higher-order evidence that *q* has first-order evidence in favour of it. So this higher-order evidence confers some additional plausibility on the first-order evidence for *q*. This is because, in seeing that the speaker has been invited to defend *q*, one has evidence that the speaker has a certain degree of expertise (or is in an epistemic position to speak on behalf of those with the relevant expertise) with respect to *q*, and so one has reason to think that the expert's first-order evidence for *q* is good. Returning to Evelyn, we can now ask: why can't her critical self-belief (that she has insufficient evidence to believe *r*) be a piece of higher-order evidence about the poor quality of her first-order evidence for her first-order belief in *r*, thereby generating a legitimate epistemic reason for her to disbelieve *r*?

The problem is that Evelyn's critical self-belief is *not* higher-order evidence. This is because, even if the content of her critical self-belief is that there is first-order evidence against *r*, she already has access to this first-order evidence. It is because she already understands the first-order evidence that her self-belief does not introduce novel information into her epistemic situation. It merely codifies, in a self-conscious way, what her first-order epistemic situation is.

To better see what I mean, it may help to consider a case where Evelyn *does* come to acquire higher-order evidence about the poor quality of her first-order evidence for *r*. Imagine, then, that Evelyn's friend Thien approaches her and says that he has evidence that *r* is false. It is plausible, at least if Thien is (taken to be) her epistemic peer, that Evelyn now acquires higher-order evidence that her first-order evidence for *r* is less secure than she took it to be (at least to some degree, perhaps a small one). Why? Because Thien is a different epistemic agent from Evelyn who, therefore, may be in possession of evidence against *r* that she lacks. Thien is not *simply* recapitulating first-order considerations of which Evelyn is already aware. Rather, he is *also* making manifest to her the fact that *another reasonable agent* takes *r* to be problematic. This is why his testimony can generate higher-order evidence for her to disbelieve *r*, whereas her critical self-belief (based on nothing but her own scrutiny of her first-order evidence) cannot.

In sum, critical self-beliefs may fail to serve as genuine epistemic reasons for or against their first-order objects, at least where norms of evidence govern the reasonability of one's first-order attitudes, because one's critical self-beliefs are not themselves evidence, higher-order or otherwise, for those attitudes. If so, this undermines the importance and scope of Burgean Agentialism to the project of rational attitude revision, and thereby undermines the indispensability of privileged and peculiar self-knowledge as it figures into that project.

§3.3.3—Objection to Burgean Agentialism: Critical Self-Beliefs as Motivationally Inert

Yet another worry about the role played by critical self-beliefs comes to us from David Owens, who asks: "if you already have a non-reflective awareness of the reasons [whatever they are] which ought to motivate you, how does the judgement that you ought to be moved by them help to ensure that you are so moved?" (2000, p. 18). An answer: they do not. As such, and once

again, beliefs like *I ought not to believe that p* "look like an idle wheel in our motivational economy" (Ibid., p. 18).

Another way of putting the point is nicely articulated by Casey Doyle, who writes: "If I am rational, attending to new evidence [and, we might add, epistemic support relations between pieces of evidence] automatically results in revisions to my attitudes without reflection" (2018b, p. 16). If this is right, a further question now seems apt: "if my attitudes conform to my current assessment of the reasons I have, why attend to the attitudes instead of the reasons?" (2018b, p. 16). If one's attitudes are automatically adaptable to evidential considerations when one is rational, then in what sense is having an explicit self-perspective on one's attitudes an especially important part of the attitude-revising process, motivationally speaking?

In light of this concern, critical reasoning appears to play no indispensable motivational role in the process of rational attitude revision. It is, of course, possible to retreat to the position that critical reasoning is only intended to generate *rational* pressures on our first-order mental states, not motivational pressures to actually forfeit or maintain those states. But we have already challenged that view as well. Thus, from whichever angle (normative or psychological) we try to stake out a space for the importance of critical self-belief (amounting to privileged and peculiar self-knowledge or not), we find ourselves running into significant challenges.

§3.3.4—Objection to Burgean Agentialism: Easy First-Order Rationality

Yet another objection to Burgean Agentialism elaborates on the worry, broached in §3.3.2, that our first-order attitudes are not made more or less rational by critical self-beliefs. Thus, Stephen Blackwood points out that Burge's motivation for thinking that critical reasoning is indispensable to our rational identities issues from a questionable "supervisory" conception of

the cognitive functions of privileged and peculiar self-knowledge.[105] We see this conception play out when Burge argues that critical reasoning is essential, among other things, to the *maintenance* of one's first-order rational life, thereby implying that an agent who fails to critically reason risks collapse into wholesale first-order irrationality. However, it is controversial whether this sort of collapse is even possible. For it seems plausible that interpreting an agent *as* an agent *requires* us to attribute a certain degree of rationality to them. This is one version of Donald Davidson's principle of charity (1973).

The Burgean Agentialist might reply that charity actually requires us to ascribe critical reasoning capacities to agents *precisely in order to explain* their basic rationality. But, as Blackwood reminds us in this connection, there is another way to explain the agent's rationality, a way which trades on a metaphysical account of the propositional attitudes rather than on a psychological account of our critical reasoning capacities. Thus, according to mental holism, the identities of at least some of our propositional attitudes are determined in part by their logical relations to other attitudes. We can see this by way of simple examples. In believing that the cat is on the mat, it is hard to dispute that I have this belief only if I have some other set of beliefs about what cats and mats are. Notably, the logical relations between these beliefs are rational relations; they constrain the extent to which one's attitudes can fail to cohere with at least some other attitudes that one holds. Therefore, mental holism entails that, if one is an agent with propositional attitudes, some of those attitudes are rational.[106] Crucially, however, this measure of rationality *cannot* be understood as a product of critical reasoning since, *ex hypothesi*, an agent cannot even *have* attitudes about which she can critically reason unless some of them

[105] Blackwood (2010, pp. 184-188)
[106] Davidson (1999, p. 126) writes that there are "no beliefs without desires, no desires without beliefs, no intentions without both beliefs and desires."

already rationally cohere. Now notice the implications this has for the putatively constitutive relationship between critical reasoning and rational agency: it can never actually happen that, without critically reasoning, our first-order rationality might be totally compromised, and so it cannot be the case that critical reasoning is indispensable to rational agency.

Might the Burgean Agentialist reply by reiterating that we are interested in grounding a constitutive link between a certain level of *sufficiently developed* rational agency and critical reasoning? On this reply, since a capacity for critical reasoning is constitutive of sufficiently developed rational agency only, it is possible to argue that the relevant level of rational agency requires critical reasoning of a self-knowing sort, for it is only by this means that we can elevate our rationality beyond the level that is guaranteed by the truth of mental holism.

Notice, however, that this reply simply returns us to an earlier objection: critical reasoning might help here, but so too might second-tier reasoning. What we want and do not have is an argument for why critical reasoning requires something beyond second-tier reasoning. Moreover, even if critical reasoning *is* constitutive of suitably mature rational agency and does require us to have privileged and peculiar self-knowledge, the extent to which this result should interest us is unclear. For who among us are suitably mature rational agents, and if it is not all of us, how are we to decide what exactly is so cognitively indispensable about *this sort* of rational agency?

§3.4.1—Privileged and Peculiar Self-Knowledge and The Taking Condition[107]

In §3.2 we considered the Burgean Agentialist proposal that privileged self-knowledge is guaranteed by our status as critical reasoners. By the end of §3.3 we saw that this proposal is problematic in various respects. But what if a simpler route to more or less the same conclusion was right under our noses? What if a route to a transcendental ground for privileged and peculiar

[107] Much of the material in §3.4 is taken from Winokur (forthcoming), with permission from *Logos & Episteme*.

self-knowledge had to do not with *critical* reasoning, but with *reasoning in general*? Such a conclusion would certainly undermine those who argued, contra the Burgean Agentialist, that critical reasoning could take place in a merely second-tier way. For, if *all* reasoning requires privileged and peculiar self-knowledge, then no appeal to second-tier reasoning can ground an objection to Burgean Agentialism, because second-tier reasoning won't actually be possible.[108]

The argument I have in mind can be extracted, with some setup, from recent philosophical efforts to understand the psychological process or event of reasoning (i.e., drawing inferences). The guiding thought is that drawing an inference can be an agential phenomenon: it can be a psychological process or event that is predicable of the agent herself rather than of her sub-agential cognitive mechanisms.[109] To make sense of this, a growing cohort of philosophers argues that inference involves agency because drawing an inference requires one to (1) have a "take" on how one's premise(s) confer epistemic support on one's conclusion, and (2) to draw one's conclusion on the basis of this take.

Paul Boghossian formulates this condition as follows:

> **(Taking Condition)**: Reasoning from p to q necessarily involves the thinker *taking* p to support q and drawing q *because* of that fact. (2019, p. 110)[110]

If the Taking Condition (TC) is true, then agent-level inference (hereafter just *inference* or *reasoning*[111]) is not a wholly automatic process. Rather, the agent must have an intermediating conception of the quality of epistemic support between her premise(s) and conclusion.

---

[108] At the same time, the dialectical importance of Burgean Agentialism would be undermined.

[109] Philosophers often distinguish agent-level and sub-rational reasoning via Stanovich & West's (2000) distinction between system 2 (rational) and system 1 (sub-rational) processing (see also Kahneman 2011). In this chapter all talk of reasoning refers roughly to system 2, agent-level reasoning. I say 'roughly' because this is a rough and ready distinction with blurry lines, such that it may be more appropriate to follow Boghossian in focusing on "system 1.5 and up" reasoning (2014, p. 2).

[110] Boghossian's proposal adapts an earlier version due to Frege (1979). Hlobil (2019) points out that a similar condition can be found in Russell (1920).

[111] Some distinguish reasoning from inferring, such as Koziolek (2017). Like most others, however, I use these terms interchangeably.

Besides arguing for TC itself, its proponents must also clarify exactly what taking one's premise(s) to support one's conclusion amounts to. On this score, philosophers have variously argued that "takings" are beliefs (Kietzmann 2017; Koziolek 2017; Müller 2019), intuitions (Chudnoff 2013; Broome 2013), mental actions (Hlobil 2019), and *sui generis* mental states (Boghossian 2014). I will not be adding to this particular debate here, nor will I focus too much on whether TC is true.[112] Instead, I will ask a different question about TC, one that concerns the *contents* of takings. Specifically, I will ask: does an agent's taking her premise(s) to support her conclusion require that she have self-knowledge of the attitudes she bears toward her premise(s) and conclusion? Boghossian appears to think that it does. Focusing on theoretical inferences, he espouses the following:

> **Self-Awareness Condition**: Person-level reasoning [is] mental action that a person performs, in which he is either aware, or can become aware, of why he is moving from some beliefs [or perhaps, in practical reasoning, intentions and desires] to others.[113]

Indeed, as far as I can tell, his view seems to be that this Self-Awareness Condition (SAC) is either a direct upshot of TC or, stronger still, a different way of articulating TC.[114]

On the assumption that talk of self-awareness is interchangeable in this context with talk of self-knowledge or warranted self-belief (ditto for talk of self-consciousness, as used by some authors in discussions of TC), and on the assumption that TC indeed leads to or amounts to SAC, TC may underpin an interesting agentialist account of self-knowledge. The idea is that, because inference presupposes self-knowledge (as per SAC), at least one ordinary empirical route to

[112] Skeptics include Setiya (2013), Wright (2014), McHugh & Way (2016), Kiefer (2017), Rosa (2017), Koreň (2019), Siegel (2019), and Richard (2019).
[113] The label is Siegel's (2019, p. 16), though the quote is Boghossian's (2014, p. 16). Note that Siegel herself rejects this condition.
[114] This is, admittedly, less clear in his (2018), where he does not stress any self-awareness involved in takings.

knowledge—the inferential route—is closed off. This is because SAC entails that not all self-knowledge can be acquired inferentially, since self-knowledge is presupposed in inferring.

Similarly, it can be argued that TC cuts off a testimonial route to at least some self-knowledge of our propositional attitudes. For it is implausible that, simply by being told that *p* supports *q*, you will be able to *take p* to support *q*. This is because takings are ways of *seeing for yourself* that something follows from something else, and merely being told that something follows from something else is not sufficient for this. Indeed, if testimony were enough for this, we would be enabled to draw all kinds of inferences that we intuitively cannot. Simply by being told that the various steps of Gödel's incompleteness theorem follow from one another, for example, I would be able infer that they do. But this seems absurd (Hlobil 2018, p. 2591).

However, one might think that because testimony cannot transmit takings, takings cannot be self-beliefs and so cannot be ingredients in an account of self-knowledge (Hlobil 2018, 2019). After all, it seems to be a hallmark of belief that it can be transmitted via testimony, and so if takings cannot be so transmitted, there is a case against treating takings as self-beliefs. It may be possible to reply by arguing that takings are a non-belief-involving form of knowledge-apt self-awareness, such that they can qualify as self-knowledge, but this is not a path that I will explore here. Rather, I simply want to point out that, even if Hlobil is right that takings cannot be transmitted via testimony, this only shows that beliefs do not *suffice* for taking. To see where I am going, note that Hlobil himself seems to think of takings as akin to a kind of *insight* or *understanding*. One understands that *p* follows from *q*, and so does not *merely* have the belief that they do. But states like understanding can still include belief as an essential ingredient, at least insofar as understanding is a doxastic state (Grimm 2011). I have no theory of what elevates belief (or knowledge, for that matter) to understanding. Still, takings may be beliefs plus

whatever else entails understanding.[115] So it is at least possible that takings can be self-beliefs that, when warranted, count as self-knowledge. And so it remains possible that self-beliefs can count as *peculiar* self-knowledge, since they cannot be acquired by at least two standard empirical routes: inference and testimony.

So much for the suggestion that inferential beliefs must be known in a peculiar way if SAC is true. What does all of this say of privilege? Here, I suspect some SAC-motivated agentialists would argue as follows: if self-knowledge is required in order to infer, and if a capacity for inference is basic to our rational agency, then self-knowledge of at least some of our mental states may seem like a necessary rather than contingent feature of our psychological lives. In that event, any account of self-knowledge that treats it as a merely reliable rather than highly (perhaps maximally) secure form of knowledge will be misguided.

To be sure, this agentialist account may have its limits. First, note that SAC includes a crucial "or can become aware" qualifier. As such, SAC does not entail that we actually have self-knowledge in inferring, and instead entails that we are *in a position* to have it. Fortunately, this qualifier is dropped by several arguments for SAC. However, another possible limitation of this account is that SAC may only deliver privileged and peculiar self-knowledge *during the inferential process*, such that it says nothing about privileged and peculiar self-knowledge of one's standing attitudes. This being said, with some further machinery in place one might argue that SAC grounds a constitutive connection between rational agency and standing attitudes that are "available" for inferential application, whether or not they are occurrently embedded in an inference at any given time. On such a view, one's attitudes might be construed as self-known in

[115] Utterances like "I understand that *p*, but I don't believe *p*" also have an air of Moore-Paradoxicality, further suggesting (perhaps) that the irrationality of such utterances consists in the fact that understanding includes but does not reduce to belief.

a standing way, with privilege and peculiarity, because they are *access-conscious*, where access-consciousness is (roughly) a species of consciousness according to which some mental state is poised for deployment in an occurrent mental process.[116]

Despite the potential limits of this agentialist account, it is surely interesting and dialectically relevant if SAC is true. In what follows, however, I will argue that extant arguments for TC do not establish SAC. In other words, I will argue that SAC is *not* equivalent to (or an upshot of) TC. This means that, even if TC is true, no agentialist conclusion about privileged and peculiar self-knowledge follows. To reach this conclusion, I will evaluate many arguments for TC. These will be arguments that appeal to TC in order to illuminate: (1) the inference/association distinction, (2) the good/bad inference distinction, (3) a Moore-paradoxical phenomenon associated with inference, (4) inference as a mental act, and (5) inference as involving cognitive agency. I will argue that none of these arguments lead us from TC to SAC.

§3.4.2—The Taking Condition: Minimal Versus Robust

As aforementioned, SAC entails that second-tier reasoning (understood as an agent-level phenomenon, at least) is impossible, or that any agent who can engage in second-tier reasoning *can also become aware of her attitudes* (given that crucial qualifier contained in SAC). My aim is to see whether this is right. Consider, then, two schematic inferences:

> **(Minimal)**: *p*. *p provides sufficient epistemic support for q*. Therefore, *q*.
>
> **(Robust)**: I believe that *p*. *p provides sufficient epistemic support for q*. Therefore, I now believe that *q*.

[116] To see how this might go, see Shoemaker (2009) and Stoljar (2019). I discuss the access-consciousness/self-knowledge connection at greater length in Chapter Five (§5.7.1).

At least on the surface, the difference between (Minimal) and (Robust) is that the former does not involve thoughts about mental states while the latter does. Nevertheless, (Minimal) seems to involve some sort of appreciation—that is, taking—of epistemic support: it seems to involve what we might call a 'meta-propositional' as opposed to a 'meta-attitudinal' taking-attitude (Pettit 2016), just in the way that proponents of second-tier reasoning describe.

Opposed to the minimal view is what I will call *the robust view*. This is the view that I take authors like Boghossian to accept, given their acceptance of SAC. On this view, agent-level inference necessarily requires at least being in a position to appreciate epistemic support relations between one's own inferentially-embedded attitudes as such. Again, this qualifier (that is in SAC) allows proponents of the robust view to grant that agents do occasionally second-tier infer. What they must argue is only that any agent that can second-tier infer is also *in a position to know* her inferentially embedded attitudes. This is weaker than any claim about *actually having* self-knowledge of said attitudes. Nevertheless, as aforementioned, we will examine many arguments for the robust view that drop this qualifier and so aim to establish an agentialist argument that undermines realism about second-tier reasoning altogether.

Before evaluating the five aforementioned arguments for SAC, two crucial caveats are these. First, I will focus on *theoretical* reasoning in what follows and, moreover, on non-suppositional reasoning, such that one reasons from one's actual beliefs. Thus, I set discussion of practical reasoning to one side. One reason to do this concerns John Broome's (2014, 2019) claim that practical reasoning cannot involve taking-beliefs. He asks us to imagine a piece of practical reasoning in which "[y]ou intend to raise money for famine relief and believe that running a sponsored marathon is the best means of doing so." Subsequently, you "reason from these two premise attitudes to a conclusion attitude, which is the intention to run a sponsored

marathon" (2019, p. 38). What would it look like to *take* one's initial intention and belief, in the above case, to support one's subsequent conclusion attitude? According to Broome, no sense can be made of this idea, whether we understand this taking as minimally or robustly contentful. First, Broome (2019) takes up a minimal example, i.e., an example where one's taking-belief only registers the bare propositional contents of one's attitudes, which I call:

> **(Minimal-Practical-Taking)**: if you raise money for famine relief, and if running a sponsored marathon is the best means of raising money for famine relief, then you run a sponsored marathon.

This is supposed to be the content of a taking-belief that specifies the inferential connections between one's premise-belief, premise-intention, and conclusion. The problem is that, in practical reasoning, you might not actually have a taking-belief with this content, "at least not until you have completed your intention reasoning." This is because you "might doubt that you will take the best means to your end, or you might simply not have formed any belief about it" (2019, p. 40). So this sort of taking-belief can hardly be constitutive of practical reasoning.

Next, Broome considers a robust taking-belief, which I call:

> **(Robust-Practical-Taking)**: if you intend to raise money for famine relief and you believe that running a sponsored marathon is the best means of raising money for famine relief, then you intend to run a sponsored marathon.

Broome's primary concern about this taking-belief is that it overintellectualizes what is required for ordinary practical reasoning. With respect to the first concern he argues that, in reasoning, "you think about the contents of your attitudes; you do not think about the attitudes themselves" (2019, p. 41). With respect to the second point, he argues that children can reason without possessing the concepts of intention or belief. Again, he admits that children must register some sort of sensitivity to attitudinal differences; they must view certain contents in both *intending* and

believing ways if they are to practically reason. But this does not require applying the *concepts* of these propositional attitudes, even inchoately, on his view.

In light of this objection, notice that focusing solely on theoretical reasoning in what follows is highly useful. For, if it can be argued that theoretical reasoning requires *robustly* contentful taking-beliefs, then this charge of over-intellectualization in the practical case will lose its force. This is because, from the position that even theoretical reasoning requires robustly contentful taking-beliefs, someone like Broome can no longer argue that, *comparatively speaking*, theoretical reasoning requires less conceptual sophistication than practical reasoning.

My second caveat concerns the question of how "explicit" one's self-knowledge must be according to the robust view. I have been writing as if self-ascriptive, higher-order beliefs involving mental state concepts must be involved, but perhaps different versions of the robust view could take on different commitments here. For example, one version of the robust view might attribute "implicit" or "tacit" self-knowledge to agent-level reasoners, this being self-knowledge that somehow does not require explicit deployments of mental state concepts or the first-person pronoun. Alternatively, one could argue that, while inferring does require actively forming higher-order beliefs that conceptualize one's lower order mental states, these can lie in the "background" of our consciousness when we draw inferences (Müller 2019, p. 8). I will typically characterize the robust view as the view that inferring requires taking-attitudes with a conceptualized, self-ascriptive structure, whether consciously foregrounded or otherwise. This is partially to avoid repeatedly appending cumbersome qualifiers about the possibilities of tacit or backgrounded taking-attitudes to claims made in the ensuing discussion. But it is also because I

am somewhat skeptical of the idea of tacit self-knowledge[117] and so address the clearest (by my lights) version of the robust view.[118]

Here, then, is the plan for what follows. In the next several subsections I examine several arguments for TC. In each subsection, after describing the argument, I show that it is either silent about the robust view (i.e., the view that accepts SAC) or can be objected to by proponents of the minimal view (i.e., the view that rejects SAC).

§3.4.3—The Taking Condition, Reasoning, and Association

Inference is but one way for agents to form a mental state. A non-inferential cause of many mental states is *association*. An agent might associate by being subject to some "bizarre psychology experiment" where she is conditioned to think that the sun will one day explode every time she thinks that Donald Trump is the President of the United States (Quilty-Dunn & Mandelbaum 2018). Or we might imagine a more ordinary case of a habitual depressive who, whenever she thinks about how much fun she is having, also thinks that there is so much suffering in the world (Boghossian 2019). Both cases involve mental states that are in *some* way sensitive to one another. But they are not sensitive in an inferential way. This is why it is not enough to say, as Hilary Kornblith does, that inference is simply a matter of "transitions

[117] See, e.g., Coliva's (2016, Chapter 7) critical discussion of Shoemaker's (1996) appeal to tacit self-knowledge.
[118] A final preliminary is not a caveat for what follows but, rather, a comment about a different dialectical context in which the robust view might be motivated. I have in mind the possibility that the robust view can solve a problem for TC even if it does not motivate it. Many philosophers worry that TC gives rise to an infinite regress inspired by Carroll (1895). The regress worry is that, if taking-attitudes are themselves premises of inferences, they must also be taken by further-taking attitudes to support one's other premises, *ad infinitum*. There are various TC-friendly strategies for avoiding this regress—see, e.g., Kietzmann (2017), Müller (2019), and Marcus (2020). Kietzmann's particular strategy appeals to the robust view. Here I simply note that his strategy, according to which the regress is halted by conceiving of one's taking-attitudes as ontologically indistinct from their objects, might be made to work on a second-tier conception of taking-attitudes as well. I also believe—though I won't argue it here—that Marcus's strategy does not suffer from the defects that Kietzmann points out for other strategies, and that Marcus's account of inference does not seem to depend on the robust view.

involving the interaction among representational states on the basis of their content" (Kornblith 2012, p. 55), for the interaction needs to be of the right sort (Boghossian 2015).

Might one argue that association and inference can be distinguished by pointing out that, in the latter case, one's mental states stand in epistemic support relations, whereas no such support obtains in the associative case? On the face of it, this proposal can even dispense with TC: it can account for the inference/association distinction in terms of the presence or absence of epistemic support relations, without countenancing any appreciation of these relations on the agent's part. The problem with this is that epistemic support relations can obtain between attitudes even when those attitudes are not occurrently tokened in inferences.[119] Another problem is that attitudes can be related inferentially despite the *absence* of epistemic support relations between them. After all, as Boghossian says, this seems to be what happens with *bad* inferences.

In light of the above, Boghossian argues that:

> …something like a taking-based account seems not only natural, but forced: the depressive's thinking doesn't count as reasoning *not* because his first judgment doesn't *support* his second, but, rather, it would seem, because he doesn't *take* his first judgment to support his second. The first judgment simply *causes* the second one in him; he doesn't draw the second one because he takes it to be supported by the first. (2019, p. 112)

So, according to Boghossian, TC accounts for the difference between associative and inferential mental state transitions and the difference between good and bad inferences. For, while in both good and bad inferences the agent takes her premise(s) to support her conclusion, inference is only good when the epistemic support that she takes her premise(s) to provide is actual. Either way, the psychological relationship between the mental states at issue is not merely associative, because associations do not involve taking-beliefs.

[119] E.g., one might associate thoughts in a way that matches the pattern of a modus ponens inference without actually being a modus ponens inference (Quilty-Dunn & Mandelbaum, 2018, p. 13).

Supposing that Boghossian is right about TC's indispensability here,[120] what have we learned about the essential contents of taking-attitudes? I do not think that we have been led to the robust view. This is because proponents of the minimal view could argue that an agent who thinks that *p*, and that *p* supports *q*, infers *q* on the basis of appreciating (by her lights) an epistemic support relation between two world-directed propositional contents. In the good case, an agent takes *p* to support *q* and *p* does support *q*. In the bad case, she takes *p* to support *q* but *p* does not support *q*, say, because *p* is false or because there simply is no support relation between *p* and *q* even when both are true. In either case, the agent does not seem to be required to think about her *beliefs* in *p* and *q*. Moreover, we can agree with Boghossian that she is not merely associating, because a (second-tier) taking-belief is involved.

According to Nicholas Koziolek, however, this is too quick. He argues that we should prefer an account of the good/bad inference distinction that appeals to an agent's self-knowledge. Thus, consider how he understands bad inferences. To infer badly is, on his view, to associate while mistakenly taking it that you have actually formed a conclusion-belief on the basis of a premise-belief, whereas good inferences involve no such mistakes (2017, p. 18). Koziolek grants that this is a paradoxical-sounding view, since it means describing bad inferences as, in fact, just perverse associations. Still, I will grant his claim that these deserve to be called inferences "if only by a sort of courtesy" (Ibid., p. 19).

The important point for our purposes is that, on this account, one's second-order perspective on one's attitudes is doing explanatory work, since appealing to this perspective allows us to countenance two mistakes involved in bad inferences: (1) a false belief that one's beliefs have inferentially caused one's new attitude in a non-deviant way (they haven't, since

[120] Though see Siegel (2019) for a non-TC-based account of the association/inference distinction.

one's inference *does* involve deviant causation) and (2) a false belief that one has gained knowledge via her inference (one hasn't, since the cause of one's conclusion-belief is not a rational cause). Crucially, while one surely *lacks* a degree of self-knowledge in making these mistakes, one still possesses some self-knowledge. One has self-knowledge *of the beliefs involved* in the defective inference, even though one lacks self-knowledge of the true nature of the causal or logical relations between these beliefs (Ibid., p. 18).

This account allows us to make sense of bad inferences other than those where an agent (1) takes *p* to support *q* but (2) *p* does not support *q*, say, because *p* is false. This is especially interesting if Koziolek is right, as he argues, that inferences where *p* fails to support *q* merely because *p* turns out to be false, such that one's taking *p* to support *q* is mistaken, are *not* actually bad. After all, though we can grant that such inferences fail to yield knowledge, it is still the case that, *were p to be true*, one's inference could yield knowledge. For this reason, Koziolek argues that such inferences are actually good because they are "potentially productive of knowledge" (Ibid., p. 11).

In response, I think that proponents of the minimal view can countenance at least two species of bad inference, such that we need not rely on Koziolek's account. First, bad inferences can be inferences where one fails to accord the *proper* epistemic weight to *p* in concluding *q*, even if one takes *p* to support *q* to some degree and *p* really does support *q* to some degree.[121] In such a case, the agent's focus may strictly be on these world-directed evidential relations between propositions. Second, some bad inferences might have the following structure: (1) an agent takes *p* to support *q*, (2) *p* does not in fact support *q*, and (3) the inference is *not* "potentially productive of knowledge". For example, someone who affirms the consequent infers

[121] For a similar suggestion, not tethered to a second-tier conception of inference, see Siegel (2019, p. 29).

badly, but this need not mean that her conclusion is deviantly caused by an association (as Koziolek would have it).[122] Rather, while she genuinely infers, she does so in accordance with a bad first-order inference rule, one that is not potentially productive of knowledge. In other words, affirming the consequent need not amount merely to associating *p* and *q* via, say, conditioning that produces mental "jogging" that *looks like* affirming the consequent (Broome 2013, p. 226). A test for whether one merely associates in a way that looks like affirming the consequent, as opposed to actually affirming the consequent, is this: if an agent's mental activity satisfies the pattern of affirming the consequent for different values of *p* and *q*, this is evidence that she is not merely associating, since associations are content-based conditionings, whereas inference rules range over many possible contents.

Perhaps only Koziolek's account can explain the bad inferences he describes. But whether being able to do this is something that *all* reasoners must be able to do simply in virtue of being reasoners is a different matter. Proponents of the minimal view should admit that second-tier reasoners cannot infer badly in Koziolek's sense while arguing that they can nevertheless infer badly. Moreover, as we have seen, they can do so while accounting for the difference between association and inference. Two essential desiderata of an account of inference are hereby satisfied, but they appear to be satisfied along minimalist lines. The robust view has yet to follow.

§3.4.4—The Taking Condition, Reasoning, and Inferential Absurdity

Some philosophers have noted that inferences can figure into a version of Moore's Paradox. Thus, consider what Ulf Hlobil calls "Inferential Absurdity" (2019, p. 2), or INFA:

[122] Contra Koziolek, (2017, p. 19, fn. 29), who argues that there are no fallacious inferences without self-consciousness.

> INFA. It is irrational, and transparently so from the agent's own perspective, to infer B from A1, …, An and to believe also that these premises don't support B or to suspend judgment on whether they do.

The idea is that if an agent infers *q* from *p*, but also believes that *p* does not support *q* (or is ambivalent about whether this is so), she is manifestly irrational. Hlobil points out that INFA has the air of a Moorean paradox. The reason, he says, is that inferences are *acts*,[123] and acts are not content-bearing vehicles that can be in tension with *states* like beliefs (though of course inferences *operate on* content-bearing states like beliefs). So we face the question: "[h]ow can a doing that seems to have no content be in rational tension with a judgment or a belief?" (2019, p. 421).

Now, if TC is true, this absurdity is readily explained. For, as Christian Kietzmann puts it, "[i]f inference involves the thinker taking his premises to support his conclusion, this taking-attitude will clash with a belief or judgement that the premises do not support the conclusion" (2017, p. 295).

Once again, we can ask whether this explanation of INFA also motivates SAC and, hence, the robust view. On my view, the minimalist could accommodate INFA by explaining it in terms that do not require self-awareness, on the reasoner's part, of the tension between her taking-attitude and her further belief that there is no epistemic support relation between the premise(s) and conclusion of her inference (or her agnosticism about this epistemic support relation). Perhaps this can be done by supposing that, while an agent draws an inference from *p* to *q* while simultaneously believing either of:

(1) *p* may or may not support *q*, or;

(2) *p* does not support *q*

[123] Though note that in Hlobil (2014, 421, fn. 1) he does not assume that inferential acts are *intentional* acts. I take up the question of whether inferences are intentional acts in §3.4.5.

…the contents of her mental states will contradict with a minimally contentful taking attitude of believing that *p supports q*. In other words, the inference would stand in tension with one's beliefs like (1) or (2) because, in the final analysis, drawing an inference from *p* to *q* requires having a further minimally contentful belief to the effect that *p* supports *q*, and this belief contravenes (1) and (2). Hence, the puzzle of how inferences (being acts, processes, or events) can stand in rational tension with beliefs (being states) is explained. But because these states are minimally contentful, the robust view does not yet follow.

What about the fact, as Hlobil sees it, that inferential absurdities are "transparently so from the agent's own perspective"? One way to go is to read this as a claim about a second-tier rather than a self-conscious perspective. But even if we are thinking about this perspective as a self-conscious one, the minimalist can also argue that it is unclear whether we must understand Hlobil's insistence on the transparency, to the subject herself, of inferential absurdities as an essential feature of the phenomenon. Hlobil seems to stipulate that it is. But if the minimalist's reply above is roughly correct, then inferential absurdities seem to be one thing and their self-conscious appreciation another. What proponents of the robust view need, then, is an argument for two claims: (1) that this appreciation is both necessarily self-conscious (Hlobil does not argue for this) and (2) that this appreciation is necessarily available to reasoners. Unfortunately, (2) seems to implausibly assume that agent-level reasoners are vulnerable to inferentially absurd states of mind unless they self-consciously ward them off. It is implausible because warding against irrationality is not always a matter of having to actively prevent oneself from adopting clashing attitudes, for part of what it is to be an agent is to be by and large rational (Davidson 1973; McHugh & Way 2016, p. 322). But even if some such monitoring is required, why won't 'second-tier' monitoring do?

Perhaps it is worth closing this subsection by reminding ourselves that other Moore-paradoxical phenomena might also be explained without appealing to an agent's self-knowledge, *even when* the phenomena necessarily invoke an agent's capacity to *self-attribute* mental states. Take, for example, the original Moore's Paradox concerning thoughts and utterances like "*p*, but I don't believe that *p*". Such thoughts or utterances seem highly irrational despite the fact that their conjuncts do not formally contradict one another (just as INFA begins with the thought that inferential acts and certain beliefs can stand in rational tension without formally contradicting one another). As we saw in §2.6.1, Neo-Expressivists have appealed to the expressive function of avowals in order to offer a "smooth account" of Moore's paradox (Jacobsen 1996, p. 28), one that does not depend on facts about the agent's self-knowledge.

To reiterate, the expressivist account of Moore's paradox is this. An avowal of "*p*, but I don't believe that *p*" *expresses* both a first-order belief that *p* and either (1) a first-order belief that not-*p*, or (2) a first-order agnostic attitude toward *p*. Because of this, the rational tension consists in a first-order "expressive conflict" (Bar-On 2004, p. 217). All the while, the utterance contains no contradiction at the level of its semantic meaning. Moore's paradox is dissolved when we recognize that expressive force and semantic meaning can come apart. But notice that the explanation does not invoke any claims about the speaker's self-knowledge. Rather, it invokes claims about a rational tension at the level of her first-order attitudes.

Crucially, these Moorean thoughts and utterances have an irreducibly *de se* conjunct. After all, if the target utterance or thought was merely "*p*, but not *p*" rather than "*p*, but I don't believe *p*", the irrationality manifest in such an utterance or thought would be wholly visible at the semantic level and would hardly constitute a puzzle. My point in raising the expressivist analysis of Moore's Paradox, then, is this. Because it is possible to give an account of it that does not

depend on the subject's self-knowledge of her attitudes (again, the tension exists as a first-order expressive conflict), despite the fact that the paradox cannot even be set up without granting that the subject self-ascribes her belief, this should make us doubly confident that INFA can be explained without appeals to self-knowledge (privileged and peculiar or otherwise). This is because there is at least an apparent possibility of construing the semantic contents of the components of INFA (the inference's taking-attitude and additional incompatible belief) in terms that do not make explicit reference to one's attitudes, unlike the original Moorean Paradox.

§3.4.5—The Taking Condition, Reasoning, and Practical Knowledge

Proponents of TC frequently argue that inference is *active*, and that its activeness explains why inference is attributable to the agent herself. The connection between taking and inferring can now be put as follows: "[a]ppreciating the support relation between premises and conclusion and drawing the conclusion on account of that appreciation seem to be things persons actively engage in. Inference will then count as something persons do because it involves the person-level activity of taking" (Kietzmann 2017, p. 295).

Although I am confident that this thought motivates many philosophers to embrace TC, its implications for characterizing the essential contents of taking-attitudes are not straightforward. One might think that there are straightforward implications if one thinks that there is a constitutive connection between *mental action* and *practical knowledge* of what one is doing. One could take one's cue here from Elizabeth Anscombe, who famously argued that non-observational, non-testimonial, and non-inferential knowledge of what one is doing is constitutive of acting intentionally (Anscombe 1963). The idea now is that, if the activeness of inference is the activeness of intentional (mental) action, then TC and SAC could fall out as a consequence. For, if inference requires practical knowledge of what one is doing, one's taking-

attitudes might be the very site of such knowledge—taking-attitudes might be the form of practical knowledge-in-inferring. Put differently: in inferring, one's practical knowledge of what one is doing takes the form of a taking-attitude to the effect that one's conclusion derives epistemic support from one's premise-beliefs.

In reply, one might question whether there really is a constitutive connection between intentional action and practical knowledge, however frequently these may come together as a matter of fact (Piñeros Glasscock 2019). One might also wonder whether there can be another description under which one knows what one is doing, in inferring, that does not presuppose knowledge of one's mental states.[124] I will not pursue these thoughts here. Instead, I will begin by raising the possibility that inference is not a species of action at all. Good evidence for this consists, as Kieran Setiya (2013) and Casey Doyle (2015) argue, in the grammar of inference-talk. We do not say, for example, that we are in the middle of drawing an inference from *p* to *q*, even though we may be in the middle of considering whether *p* is evidence for *q* (Doyle 2015, p. 105).[125] This suggests that the term 'inferring' and its cognates do not have the grammar of ordinary process verbs, in that they lack intelligible progressive aspects.[126]

But let us suppose, for the sake of argument, that inference is a kind of action.[127] Another concern is that it does not seem to be a *voluntary* act whether one infers *q* from *p*, since one cannot simply *decide* to infer *q* from *p*; the alternative suggests an implausibly strong form of doxastic voluntarism that I will not bother to argue against here. But it is natural to think that

[124] Thanks to Dorit Bar-On for this suggestion.
[125] See also Quilty-Dunn & Mandelbaum (2018, p. 14).
[126] By analogy: believing also consists in *taking* a proposition to be true, despite being states rather than actions (Koziolek 2018). Perhaps *judgements* are mental actions, but even these do not seem to be *intentional* ones (*pace* O'Brien 2005, 2007).
[127] Perhaps Doyle's argument from the non-processual nature of inference to its non-actional nature is too quick if there can be instantaneous mental acts. Indeed, this strikes me as one way of reading Hlobil (2019).

intentional actions are standardly voluntary actions, and so if inferences are not voluntary this may be another source of pressure against the claim that they are intentional.

It might be argued that inferences are intentional despite not being voluntary. On this argument, inferences involve *intentions-in* the act, analogously to how I can intentionally albeit reflexively (and so, in one sense at least, non-voluntarily) raise my arm as a basketball is being hurled at my head.[128] Such actions are not intentional because preceded by a decision that one voluntarily pursues, but because they have an appropriate means-end telos. However, it may seem that inferences are *never* such that one can simply decide to perform them, whereas even actions like raising my arm to block a basketball can *sometimes* be. Moreover, even when an action is reflexive and hence not voluntary, it is usually something that can be overridden: I can lower my arm quickly after it reflexively raises. But this is not possible with inference. With inference, one understands that there is an epistemic support relation between some set of propositions and draws a conclusion without the possibility of choosing to start or stop (cf. Marcus 2020). A plausible explanation of this fact is that inferences are not intentional actions.

It might be replied that inferences *can* in fact be voluntary, and so the argument that inferences are not intentional because not voluntary is a bad one. For example, David Hunter (ms.) focuses on cases where an agent's evidence for or against *p* is strong enough to license an inference in either direction. He concludes that it is up to you (i.e., is voluntary) whether you infer *p* or ~*p* from the evidence. But this at most shows that *some* inferences are voluntary. For, in cases where the evidence strongly favours only one conclusion *p*, Hunter agrees that one cannot voluntarily infer *p*. So, if voluntariness is a sign of intentional action, it will only be a sign that some corner cases of inference are intentional actions.

[128] The example comes from Parent (2017, p. 186).

More importantly, Hunter himself concedes that inferences are not intentional actions even when they are voluntary. For he follows John Hyman (2015) in arguing that the voluntary and the intentional have different scopes, insofar as voluntary actions depend on a lack of coercion and lack of ignorance, whereas intentional actions involve doing something to satisfy a desire. On this view, inferences do not aim at desire satisfaction even if they occur without coercion or ignorance, and so are not intentional actions. But since the Anscombian move depends on a connection between intentional action and practical knowledge, this is a bad result for the view that inferences (and the attitudes that figure into them) are self-known because they are intentional actions. I conclude that there is no obvious move from Anscombe's practical knowledge thesis to the robust view (i.e., SAC).

§3.4.6—The Taking Condition, Reasoning, and Cognitive Agency

Even if one agrees that inferences are not intentional actions, one might think that inferences are nevertheless active in *some* sense, and that TC can help us to understand this activeness. In other words, TC might still help us to understand how inference is a site of cognitive agency.

To see how we might proceed, note first that there exist fairly uncontroversial characterizations of the activeness of *mental states* that do not trade on the notion of an intentional action. Thus, consider Joseph Raz's (1997) claim that our "beliefs are a product and an aspect of our active nature because they are responsive to reasons" (1997, p. 222) or Tim Scanlon's claim that we have "judgement-sensitive" attitudes (1998, p. 20). The domain of judgement-sensitive attitudes has contestable boundaries, but the basic idea is that judgement-sensitive states are inherently adaptive to reasons and reasoning. No doubt, those attitudes that figure into our inferences will be among them.

Now, whatever philosophers typically mean by 'cognitive agency', Doyle doubts that they are merely thinking about the judgement-sensitivity of many of our attitudes and the reasonings in which they figure. This is because the existence of cognitive agency is typically treated as controversial, whereas the judgement sensitivity of many of our attitudes is not. A thicker conception of cognitive agency will require, on Doyle's view, that we do better to explain what makes reasoning "*attributable* to the agent herself" (2015, p. 106). And to explain *this*, we must concede (so Doyle argues) that the agent's attitudes are not only rationally sensitive to other states that figure into her inferences, but that they are also "possessed in virtue of the agent's own assessment of their credentials" (Ibid., p. 106). With this point in mind, Doyle offers the following conception of inference: "I bring it about that I believe that *p* on the ground that *q* when I believe that *p* because I take it that this is what I should believe". In this way, "my sense of how things should be with my beliefs is explanatory of my believing as I do" (Ibid., p. 107). Notice that these descriptions of the agent's taking-beliefs involve self-ascriptions. They characterize cognitive agency in terms of an agent's *de se* sense of how things should be with her own beliefs; it is by taking it that I ought to believe *q* on the basis of *p* that I come to believe *q*.

One reply is simply that Doyle has not hereby ruled out a minimalist alternative to the picture he presents. Perhaps, then, the minimalist could say that inference involves cognitive agency to the extent that we are the ones who bring our mental states into cognitive contact via second-tier taking attitudes in inferences. On this picture, my belief that *q* is responsive to my reason *p* because I take *p* to provide epistemic support for *q*, minimally understood.[129] If I am rational, we can imagine that my inference proceeds accordingly.

[129] I am not arguing that mental states are *only* sensitive to each other in inferential episodes.

To better see why this minimalist alternative actually makes a good deal of sense, notice that it would be strange, if not outright disquieting, if a reasoner's awareness of the proposition *p* as providing rational support for the proposition *q* could never generate motivation to believe *q*. This reiterates a point that was made against Burgean Agentialism in §3.3.3: "if you already have a non-reflective awareness of the reasons which ought to motivate you, how does the judgement that you ought to be moved by them help to ensure that you are so moved?" (Owens 2000, p. 18). An answer, once again, is that they do not. As such, beliefs like *q is what I should believe in light of p* "look like an idle wheel in our motivational economy" (Ibid., p. 18).[130] If this is right, an agent's sense of how things should be with her beliefs is *not* explanatory of her believing as she does. Rather, an agent's awareness *that p supports q* is explanatory. We can grant that the only reason why one would (rationally) take it that *q* is what one should believe on the basis of *p* is that one takes *p* to support *q*. And it is surely reasonable to believe that one ought to believe *q* on the basis of *p* if one takes *p* to support *q*. But then the question resurfaces: what additional motivational role, in coming to believe *q*, is played by this further (self-)belief?[131] As I understand Owens's point, neither my sense of what I ought to believe nor the norms that constrain my sense of what I ought to believe must figure into the contents of my inference itself if I am to be motivated to draw an inference.

I am not hereby denying that we sometimes place a great deal of stock in ensuring, from a self-conscious perspective, that we believe in accordance with our sense of how we ought to believe. I am only denying that such an aim is *constitutive* of reasoning, and that it is indispensable to our identities as rational agents. Granted, some proponents of TC do understand reasoning, constitutively, as a matter of "figuring out what follows or is supported by other

[130] In §3.3.3 we also saw that Doyle himself (2018b) has more recently made this point.
[131] Owens is originally responding to Burge's account of the importance of Burgean "critical reasoning" (see §3.2).

things one believes" (Boghossian 2014, p. 5), where presumably this is one's aim *de dicto*.[132] However, my point is that this is an implausible characterization of the constitutive aim of reasoning. Thus, I agree with Conor McHugh and Jonathan Way that "[i]t is surely more plausible that the aim of reasoning is something like finding out what is true, rather than finding out facts about what one's own beliefs support" (2016, p. 325).[133,134]

Now, McHugh and Way also seem to think that this latter aim requires no taking-beliefs, not even second-tier ones.[135] However, I cannot be sure whether this is because they reject the indispensability of taking-beliefs even as a second-tier matter, or if it is because they never considered any such class of attitudes, say, because they assume that TC entails SAC and hence is objectionable. At any rate, *if* they intend to deny even the indispensability of second-tier taking-beliefs to the project of finding out what is true, I am less optimistic about this move: *some amount* of seeing what supports what seems indispensable to aiming at the truth, even if one's aim is not to see what follows from what. This is what the minimal view of TC captures.

In all, I conclude that the robust view of TC is presently undermotivated. In other words, SAC does not follow from TC. However, I have hardly denied that TC is plausible in its minimalist reading. Indeed, I believe that those who would reject TC altogether incur significant burdens that I am not at all confident they will be able to discharge in the end. They will have to explain: the difference between inference and association, the difference between good and bad inferences, Inferential Absurdity, and the activeness (however minimal) of inference.[136]

---

[132] Shah & Velleman (2005) espouse a similar view.
[133] Consider also Malmgren: "the [reasoning] agent needn't think of herself as settling, or trying to settle, what to believe…She just has to try to do it." (2019, p. 206).
[134] McHugh & Way propose that the aim of inference is "fitting attitudes". But while this aim references one's attitudes, they argue that "agents can be sensitive to fittingness-preservation in reasoning without representing their reasoning as fittingness-preserving" (2018, p. 180).
[135] See also Rosa (2017, p. 12).
[136] Does this mean that TC is also true of practical reasoning, and if it is, does it mean that an analogue of SAC is true of practical reasoning? This might be so, given Broome's argument that second-tier taking attitudes are

§3.5.1—Shoemaker: Self-Knowledge and Self-Blindness

So far I have considered arguments that attempt to tie reasoning capacities of various sorts to privileged and peculiar self-knowledge of whatever attitudes are involved (or potentially involved) in those forms of reasoning. In this section I will turn away from arguments that tie such self-knowledge to our reasoning capacities, and consider instead an influential argument to the effect that *simply being rational* (along with meeting some other conditions, like having the concepts required to self-ascribe one's mental states) is sufficient for having privileged and peculiar self-knowledge. Accordingly, this account will be a *rationalist* rather than *agentialist* account of privileged and peculiar self-knowledge, as defined in §3.1.1.

This account comes to us from Sydney Shoemaker, who asks us to try and imagine a *self-blind* agent. On his definition, a self-blind agent is an agent who can only know her mind third-personally, say, by drawing inferences about the mental causes of her behaviour, or by soliciting testimony from other people about what she believes, desires, and so on (1996a, pp. 30-31). I will take us through Shoemaker's discussion of the self-blind agent by focusing on an agent who is putatively self-blind with respect to his beliefs.

Consider George, an ordinarily rational agent who possesses mental state concepts: he can infer things based on evidence, forfeit his beliefs when he deems them problematic, reason practically about what he ought to intend, believe, desire, and so on. Imagine now that he is in process of coming to know that he believes that Asad is a good person. As a self-blind agent, he

unintelligible in the practical case, and given that TC is nevertheless plausible, such that we have to accept the robust view for the practical case. Rather than pursue this argument, which would demand a response to Broome's over-intellectualization response, I will address concerns about over-intellectualization in §4.6 when developing my own agentialist account, thereby (potentially) motivating the robust view of practical reasoning.

might gain this knowledge by observing himself praising Asad's actions and character. Because he notices himself praising Asad, George self-ascribes the belief that Asad is a good person.

Apparently, however, we cannot really imagine George. Shoemaker's first argument for this turns on a certain understanding of the following situation involving a putatively self-blind individual:

> Now it seems possible that the total evidence available to a man at a given time should support the proposition that it is raining, while the total "third-person" evidence available to him should support the proposition that he does not believe that it is raining. This could happen even if the third-person evidence included the fact that he had just said "It is raining"; for the rest of the third-person evidence might support the proposition that in circumstances like these he is likely to lie! (1996a, p. 35).

The point of raising this situation ties back to Moore's Paradox. The thought is that, if George is conceivable, it will be perfectly rational for him to utter "it's raining, but I don't believe it" in a situation like the above. Again, this is because he could have behavioural evidence about himself, on the one hand, and evidence about the world, on the other, that support each conjunct of his utterance. However, it is generally taken as a basic datum that such utterances are deeply irrational to produce, even though they contain no formal contradiction. And so we are faced with a theoretical choice: deny that George's utterance is Moore-paradoxical, or deny that he is self-blind. Shoemaker thinks that we should deny the latter. Thus, George must have some way of knowing his own mind that does not require him to gain evidence, in a third-personal manner, about himself. For it is only if he has some such way of knowing his mind that it will not be reasonable for him to assert what Moore's Paradox tells us he cannot reasonably assert (perhaps, on occasions where he knows his mind third-personally, he will be able to reasonably produce such utterances, but this cannot be his *general* condition if Moore's Paradox is a real phenomenon).

One upshot of this argument, then, is that George must have peculiar (i.e., non-third-personal) self-knowledge. More than this, Shoemaker thinks that his self-knowledge depends on nothing but his having various first-order attitudes "plus a certain degree of rationality, intelligence, and conceptual capacity" (1996a, p. 34). This is because the impossibility of self-blindness turns on nothing more than what is generally true of a rational subject—in this case, being someone who would be *irrational* in making the relevant kind of Moore-paradoxical assertion. If this is right, then perhaps we have the makings of an argument for privileged access as well. For, if agents need not do anything to achieve their self-knowledge, and instead have it whenever they are rational, intelligent, and in possession of the relevant concepts, then their risk of *failing* to acquire self-knowledge is slim to none so long as they meet these criteria. And surely this is a way of showing that such self-knowledge is exceptionally reliable among other forms of empirical knowledge where, clearly, simply being rational does not suffice for its possession.

But perhaps all of this has been too quick. For, as Shoemaker himself points out, there may be ways in which George could learn to recognize the "logical impropriety" of Moore-paradoxical utterances even without privileged and peculiar self-knowledge (Shoemaker 1996a, p. 34). He could learn this, for example, on the basis of realizing that those who believe that *p* should be disposed to express it by uttering "*p*" in appropriate contexts, which is why following up any such utterance with "but I don't believe that *p*" would indicate that one doesn't have the belief one's first-order assertion expresses. Moreover, because George, being rational and conceptually equipped, would recognize the connection between asserting and believing, he would also recognize that it is appropriate to follow up his assertions by self-ascribing beliefs. But none of this seems to depend on privileged and peculiar self-knowledge. Again, it just seems

to depend on his understanding his psychological concepts and some basic facts about the nature of assertion. If all of this is right, then George's uttering "*p*, but I don't believe that *p*" *would* be irrational (even by his own lights), given his knowledge of its logical impropriety, even if he is self-blind.

From here, Shoemaker provides a different argument against the possibility of self-blindness. On this argument, because even the self-blind agent would not produce Moore-paradoxical utterances and would be (*ex hypothesi*) in every other sense just like a normal rational agent, "there would be nothing in his behavior, verbal or otherwise, that would give away the fact that he lacks self-acquaintance [i.e., privileged and peculiar self-knowledge]" (1996a, p. 36). Put differently, Shoemaker's argument seems to be that, because we could not recognize anything in George that would indicate a lack of privileged and peculiar self-knowledge, he therefore surely has it. For, unless one can do something to show that one is self-blind, we have no reason to countenance it as a possible form of rational agency.[137]

§3.6.1—Objections to Shoemaker

Shoemaker's claim is that the behavioural indistinguishability of a self-blind person from a non-self-blind person is tantamount to denying her existence. But this argument may make an illicit leap. For we might just as easily ask how we can be sure that a given agent has anything *more* than third-personal self-knowledge if her behaviour is indistinguishable from a possible person who has (or could conceivably have) privileged and peculiar self-knowledge (Siewert 2003, p. 134). As Karsten Stueber puts the point: "[a]ll that Shoemaker proves is that the rational person behaves as if she has [privileged and peculiar] self-knowledge in avoiding Moore-paradoxical

[137] I share this reading with Siewert (2003, p. 133) and Parrott (2017, pp. 5-6).

utterances. But this does not prove that she behaves in that manner because she has [privileged and peculiar] self-knowledge" (2002, p. 278).

But perhaps Shoemaker is simply taking a different dialectical starting point from the skeptic. Perhaps, having noticed that the putatively self-blind subject is behaviourally indistinguishable from the so-called non-self-blind subject, he is simply making an intuitive choice to reject self-blind agency, rather than giving into the skeptical option that George does not, or at least cannot be shown to have, privileged and peculiar self-knowledge. The idea here is that Shoemaker's choice may be dialectically innocent *unless* we are already in the grips of skepticism about privileged and peculiar knowledge. More than this, his choice may be dialectically *favoured*, since so many epistemologists share the intuition that we do have privileged and peculiar self-knowledge. If one has this perspective, one might concur with Setiya that "the impossibility of self-blindness is not a doctrine to be argued for, but a datum in the study of self-knowledge to be taken for granted and explained" (2011, p. 180).

But is the impossibility of self-blindness a datum to be explained, or is its *appearance* such a datum? If only the appearance of self-blindness's impossibility is our datum, then what entitles us to conclude that we must have privileged and peculiar self-knowledge? Perhaps there are dialectical contexts where theorizers are entitled to assume the reality of ϕ over and above the mere appearance of ϕ, say because there is some defensible principle to the effect that being aware of skeptical *possibilities* does not, all on its own, suffice to motivate genuine skeptical *hypotheses*. Here one might think of a common response offered to twin-earth skeptical scenarios about knowledge of semantic content. The skeptical scenario involves an agent who has been removed from her native world and placed on an alien world where she interacts with substances that are experientially indistinguishable from those with which she interacted on her home world.

In such a situation, certain semantic externalist theories dictate that she will eventually come to have 'switched concepts', and they take this to show that we cannot know the contents of our thoughts purely by introspection. One common response to this scenario is to point out that the mere possibility of one's having undergone concept-switch is insufficient to motivate skepticism about knowledge of content, for the possibility of such a switch is not really a *relevant alternative* to the possibility that one has not undergone a covert switch onto an alien world.

Let us grant that the relevant alternatives strategy for responding to unmotivated skeptical problems is a good one. Even still, it is far from obvious that it applies here. This is because there are ways of being a skeptic about privileged and peculiar self-knowledge that do not depend on fanciful scenarios like covert world-switches.[138] Thus, Peter Carruthers (2011) and Quassim Cassam (2015) argue that self-knowledge is the product of *sub-personal* inferential procedures. Sub-personal inferences are not fanciful skeptical posits; they are the bread and butter of explanatorily fruitful projects in the cognitive sciences. Carruthers's and Cassam's account can explain the putative absence of inferential procedures, from the first-person point of view, in possessing self-knowledge, all while denying that there is anything especially privileged or peculiar about self-knowledge. This is because the view allows for agents who (or whose sub-agential cognitive mechanisms) systematically fail to perform these inferences, or to succumb to all kinds of ordinary inferential errors.[139] Perhaps Carruthers and Cassam have not actually supplied the true explanation of how we acquire self-knowledge. But the point is only to raise possibilities that are (1) empirically respectable, and (2) incongenial to denying the possibility of

[138] See Falvey & Owens (1994) and Ludlow (1995), however, for plausible cases that do not involve being transported into an alien environment unawares.
[139] Similarly, Knappik (2015, p. 194) argues that *any* form of what he calls *interpretivism*, which we can think of as a broadly *detectivist* or *recognitional* model of self-knowledge, can accommodate the idea that self-knowledge—phenomenologically speaking—is not epistemically mediated.

self-blindness, all while (3) respecting the appearance of the impossibility of self-blindness. This means that Shoemaker's dismissal of the possibility of self-blindness is not (at least not obviously) dialectically innocent. I conclude, therefore, that Shoemaker's argument against the possibility of self-blindness does not go the distance, for he does not get us beyond the mere appearance of its impossibility.

§3.7.1—Parrott: Self-Knowledge and the First-Person Perspective

Matthew Parrott has recently offered a different argument against the possibility of self-blindness. He sets up his argument by distinguishing himself from Burge. He writes: "Burge overemphasizes the importance of critical reflection" to securing a basic connection between rationality and self-knowledge. "It rather seems to me":

> …that a rational agent who has the capacity to consciously self-ascribe beliefs requires first-person access to them, whether or not she ever subjects them to deliberative or critical evaluation. This is because, from a rational agent's point of view, one's beliefs depend on her having adequate reasons for them. This is fundamental to the nature of the first-person perspective of a rational believer…[As] a rational agent, when I *attribute* a belief to myself, I am attributing an attitude that I at least tacitly conceptualize as being appropriately grounded in reasons for believing, even in cases where I cannot articulate what those reasons are. (2017, p. 9)

While Parrott speaks of beliefs here, we can reasonably imagine that he might make a similar point about other attitudes that can be based on reasons, such as intentions and desires. But why should it be that, because self-ascriptions are ordinarily taken by the self-ascriber to be based on good reasons, she must have "first-personal" access to her attitudes?

The reason is that "[i]f my way of self-attributing beliefs rested entirely on third-personal ways of knowing, then it would mean that, from my own perspective, my belief that *p* might depend on something other than what I regard as adequate reasons for holding it" (2017, p. 10). This is because, if I had to *discover* my belief, then I would have to concede that this belief could

have been determined by something other than what I take to be good epistemic reasons for holding it. This, in turn, is because "behavioral evidence in favor of the proposition that I *believe* that *p* is typically not evidence for the truth of the proposition *p*" (2017, p. 10).

Parrott offers an example. Suppose you are walking down the street and find yourself frequently glancing over your shoulder. This can count as good evidence that you believe your neighbourhood is unsafe. But it is not good evidence that your neighbourhood is, in fact, unsafe. For this reason, by self-ascribing a belief that your neighbourhood is unsafe in this way, you "leave open the possibility that what [you] believe is not determined by what [you] think [you] ought to believe" (2017, p. 10). But from the first-person perspective, you must take this possibility to be closed off. You must take it that your beliefs are based on good epistemic reasons. The same point would seem to apply to my being told what my attitude is by somebody else: being told that I believe *p* would not ensure that, from my own perspective, my belief was determined by my own judgement about the reasons in its favour.[140] If our only form of access to our attitudes is third-personal, then we cannot epistemically rationally take up this perspective on ourselves.

Parrott does not couple this claim with anything quite as strong as Shoemaker's claim that self-knowledge simply supervenes on our rationality and our capacity for taking up the first-person perspective (I consider his account of how we have self-knowledge in Chapter Five), and so we might wonder whether something more is required. Nor does he say anything about why such self-knowledge must also be privileged, rather than simply peculiar. Nevertheless, he does think that this argument tells against the possibility of self-blindness, and so he concludes that

[140] A somewhat similar line of thought can be found in Gallois (1997).

we have a different route—one that does not run into Shoemaker's troubles—for establishing a necessary connection between rational agency and peculiar self-knowledge.

§3.8.1—Objections to Parrott

One concern for Parrott is that, from a third-person perspective, one could view one's attitudes as by and large justified by good epistemic reasons, even if one cannot be sure that they have been *determined* by one's appreciation of these reasons. This follows if, being by and large rational, we are resourceful enough to discover justifying reasons for our attitudes in most situations. The question, then, is why this is not good enough to constitute a rational self-perspective.

Parrott anticipates this sort of objection. He writes:

> But couldn't I have some reasons, perhaps even excellent ones, for thinking that the beliefs I ascribe even from a third-person perspective are based on good reasons? For instance, especially since I don't remember the basis for much of what I believe, I might reasonably just take myself to have a general reason for thinking that all my beliefs are based on good reasons. (2017, p. 11).

But he replies that this misses the point. The problem now, he says, is that even if I can take myself to have good reasons for my attitudes (from a third-person point of view), no judgements that I make about the reasonability of my attitudes will have an immediate rationally necessary connection to those attitudes (in the sense described by Burge in §3.2). This is because a "third-personal mode of self-ascription leaves open the possibility that, if I were to reconsider the question of whether *p*, I might come to a conclusion that *diverges* from what I actually believe." In other words, if I self-ascribe a belief on a third-personal basis, I may go wrong about what my actual belief is, and so my self-perspective will dissociate from my first-order belief. When this happens, my self-belief cannot have immediate, rationally necessary consequences for what I ought to believe (think of Burge's argument from §3.2). So my third-person self-perspective cannot be the perspective of a suitably rational agent.

The problem is that this reply simply returns us to Burgean Agentialism. For, if the point is just that it is only from the first-person perspective that one's judgements about what one ought (or ought not) believe, desire, or intend are able to immediately dictate what one ought to believe, desire, or intend, then we have just returned to thinking of the ground of privileged and peculiar self-knowledge in terms of critical reasoning. If this is all there is to it, then Parrott's proposal suffers from all the same defects as Burge's.

However, §3.7.1 began with Parrott claiming that he wants to distance himself from Burge, at least to some extent. Perhaps, then, the idea is simply that it is only from the first-person perspective that self-beliefs about what one ought to believe, desire, or intend are able to immediately determine what one ought to believe, desire, or intend, whether or not these self-beliefs are ever involved in critical reasoning per se. However, all the standard objections to Burgean Agentialism apply here too, e.g., objections about the motivational and rational efficacy of such self-beliefs.

Perhaps, instead, the argument is simply that there is nothing for the first-person perspective to be other than a privileged and peculiar self-knowing one. But this too is undermotivated. For example, Pascal Engel writes that "perception is not necessarily reflexive. Whether or not we take it to necessarily imply awareness, everyone agrees that our perceptual beliefs are not necessarily reflexive in the sense of having second-order beliefs" (2010). If Engel is right, surely this does not mean that the perspective of a perceiving creature is not first-personal. It is a perspective from which the perceiver acts and is engaged with her environment; it is not a *third*-personal perspective from which she detachedly contemplates herself and her perception.

Similarly, Hilary Kornblith writes:

> I have said that it is possible to give an account of unreflective processes from the first-person perspective. It may seem, however, that this is not obviously so. After all, insofar as they are unreflective, doesn't it follow that there really is no first-person perspective on them at all?
>
> …It doesn't follow because, to take a single example, there is something it is like to have a visual sensation, even if one is not reflecting on the sensation itself…Similarly, there is no contradiction in asking what processes of belief acquisition, or reasoning, or acting are like from the first-person perspective, even in cases in which they are not themselves reflected upon. (2012, p. 157)

Finally, as we have seen throughout this chapter, it may be possible to make sense of the perspective of a reasoner in second-tier terms, this being the perspective of one who determines one's attitudes in accordance with one's thoughts about evidence, the good, or whatever, without simultaneously thinking about one's attitudes as such. This may still be a first-person perspective, since it is a perspective from which one's own evaluations of reasons are rationally and psychologically efficacious. I do not see, then, how Parrott's account of the necessary relationship between the first-person perspective and privileged and peculiar self-knowledge can both differ from Burge's problematic account and be plausible in its own right.

§3.9.1—Peterson: Epistemic Control and Self-Knowledge

One final account to be considered in this chapter, due to Jared Peterson, argues that privileged and peculiar self-knowledge greatly increases our capacity for "epistemic control". Note below that when Peterson refers to privileged access, he has in mind self-knowledge that is privileged *because peculiar*:

> …privileged access is required if we are to possess a type of robust epistemic control over facts about our minds. The epistemic control in question is a matter of being able to keep private or disclose particular facts about one's mind to others. If, for example, a defense lawyer has privileged access to the fact that she believes her client is guilty, this is a fact that, in the typical case, she will be able to refrain from disclosing to her client, a client that will lack the type of epistemic security she has to it. Her client will lack this type of epistemic security because he is not an epistemic authority with respect to the fact in question. (2020, p. 4)

> Certainly, a poker player would have no chance of bluffing her opponent if the latter possessed the type of epistemic access to her mind the bluffer possesses.…a middle school math teacher would have a very hard time accomplishing the goal of motivating a struggling student if the latter had epistemically secure knowledge of the fact that her teacher believed she was not a strong student. (Ibid, p. 5)…An estranged lover might want a former partner to know in a highly epistemically secure manner that she still loves him. Disclosing this fact would be of paramount importance in such a case. (Ibid., p. 5)

The basic idea is that we would lack the capacity to keep our mental states private or to disclose them in ways that are sufficiently trustworthy if it were not the case that we tended to have more secure epistemic access to our minds than the minds of others. Notably, Peterson does not argue that this capacity for epistemic control is indispensable to our agency, rationality, or whatever; he contents himself with the weaker conclusion that it is instrumentally valuable. My objections to his account will not turn on this point.

§3.10.1—Objections to Peterson

I am less interested in the second part of Peterson's account, the one which claims that privileged and peculiar self-knowledge undergirds the special trust that hearers cede to us when we choose to avow our mental states rather than keep them quiet. This is because we have already seen that Neo-Expressivism provides an alternative account, wherein we are especially trustworthy because we alone can directly *express* our mental states to others. The estranged lover may express her love itself by avowing it, whatever the nature of her self-knowledge, thereby garnering renewed affection from the beloved. This being said, the first part of Peterson's account is worth considering, since what we keep private is precisely what we do *not* (Neo-Expressively) avow.

My view is that the epistemic control involved in keeping one's mental states private does not require (and is not necessarily greatly enhanced by) privileged and peculiar self-knowledge. Here is a counter-argument to Peterson's claim:

(1) I discover, third-personally, that I believe *p*.
(2) I have greater agency over my own behaviour than others do.
(3) I use my agency over my own behaviour to prevent myself from self-ascribing my belief that *p*, or from behaving in ways that indicate that I believe *p*.

By way of this process, we can see how I can keep my mental states private.

Later in his article, Peterson considers a worry that is applicable to my counter-argument: if both myself and another person wanted to know my mind and I had to use the same methods as the other person to know my mind, it would often happen (or be at risk of happening) that I would fail to acquire knowledge of my mind before the other person did. In that event, my superior control over my own behaviour would make no difference, for this sort of control would be exercised too late.

A first point to make is that, at least sometimes, it *is* too late to keep one's mental states private. Despite herself, the poker player exposes her intention to bluff in a tense game, or the lawyer accidentally exposes his lack of confidence in her client's innocence. This is a simple consequence of the fact that mental states are not intrinsically private Cartesian atoms: they can often be ascertained by others even when we try to conceal them.

A second, stronger point is this. Even if we know both ourselves and others third-personally, such that nobody has a privileged position with respect to my mind, there is a familiar Rylean argument according to which we are better at knowing ourselves than others are because we typically have more evidence about our mental states than others do. This is because we have more knowledge of our own behaviours than others do, given that we are around ourselves all the time whereas others usually are not around us all the time. If this is right, then

we can explain why I am usually in a position to acquire self-knowledge before you acquire knowledge of my mind. And once I have this knowledge, I can control my behaviour to (at least often) conceal my mental states.

What if the Rylean response fails?[141] Here I think Peterson faces one final skeptical reply. The skeptic could argue that we do not typically find ourselves in situations where others learn about our mental states before we do, but only because we are more motivated to know ourselves than others are and so set out to third-personally acquire self-knowledge more often than we set out to acquire knowledge of other minds. This is because we often *value* the sort of epistemic control Peterson describes. So we set about to gain knowledge of our mental states as often as we can in order to exercise control over how we disseminate these facts about ourselves.

But don't we also value knowing the minds of others? We do, but because we also value epistemic control, and because having epistemic control requires us to prevent others from knowing our minds before we know them, one could argue that we often prioritize acquiring self-knowledge. Moreover, even if we are not better at inferring our own mental states than we are at inferring any other person's mental states (this being the rejection of Ryleanism), we still have more time with ourselves than we do others in which to set about drawing these inferences, and so we will likely end up having a good deal of epistemic control. Perhaps this control will not be possessed by agents at nearby possible worlds whose motivations are sufficiently differently structured, such that they tend not to prefer having epistemic control over the mental states they reveal to others. But at such a world, it is also unclear whether this must really a problem for them.

[141] Peterson dismisses Ryleanism (2020, p. 6), but he does not argue for this dismissal. Perhaps he finds it phenomenologically implausible. In that event, the Rylean might rebound to a "Sellarsian" behaviourist view of the sort described by Manning (2014).

§3.11.1—Preamble to Chapter Four

One might be led, from what has so far transpired, to think that privileged and peculiar self-knowledge plays no particularly deep role, if any, in shaping us as the kinds of agents and rational beings that we are. This is not actually my view, however. Thus, in Chapter Four, I will turn to the positive project of offering my preferred argument for the indispensability of privileged self-knowledge, though I will have to begin with two more problematic accounts, since I intend to fold some key elements of them into my own.

Chapter Four—Agentialist Self-Knowledge: Part Two

§4.1.1—Introduction

In this chapter I evaluate a final set of agentialist accounts. As before, I primarily engage with these accounts as *non-substantive* accounts of self-knowledge: accounts of why we *must* have privileged and peculiar self-knowledge, rather than *how* we have it. One exception is the agentialist account discussed in §4.7, which I engage as both non-substantive and substantive.

I begin with Akeel Bilgrami's (2006a, 2012[142]) agentialist account in §4.2, according to which the privileged and peculiar features of self-knowledge are to be explained as *conceptually entailed* by facts about free and responsible agency. I then raise objections to his account in §4.3. In §4.4 I consider Annalisa Coliva's (2009, 2012, 2016, 2019) agentialist account, which is in many ways similar to Bilgrami's while also making some (by my lights) improvements. Alongside her account, I also discuss one more argument, due to Paul Boghossian, concerning a possible connection between inference and self-knowledge (see §3.4). By criticizing both Coliva and Boghossian, I place myself in a position to develop my own agentialist account. This agentialist account, which I develop in §4.6, transcendentally grounds privileged and peculiar self-knowledge in various aspects of our social-epistemic agency. In §4.7 I explicate one more agentialist account (McGeer & Pettit 2002; McGeer 2008; Vierkant 2012a, 2012b, 2013). Finding this account plausible in the main, I then integrate it with my own agentialist account. In §4.8 I set us up for Chapter Five.

§4.2.1—Bilgrami's Agentialism: Privileged Self-Knowledge, Freedom, and Resentment

Bilgrami writes that if "…self-knowledge, being *knowledge* after all, was just another narrow epistemological theme, I don't think we could account for our intuitions about privileged access"

[142] Both sources are successors to his (1998).

(2012, p. 277). What does this mean? Roughly, the thought is that treating self-knowledge as privileged and peculiar is liable to sound like an "outdated Cartesian dogma" (Ibid., p. 277) when we consider the ubiquity of self-deception, self-repression, and our general epistemic fallibility, unless we understand self-knowledge as it relates to other aspects of our personhood. For Bilgrami, these aspects are richly normative: they concern themes of value, freedom, and responsibility. Out of reflections on the interrelations of these themes, Bilgrami promises to develop an agentialist account of privileged and peculiar self-knowledge.

In my view, the most perspicuous distillation of Bilgrami's account comes not from him, but from his commentator Krista Lawlor, who represents it as follows:

(1) Truly agential mental life is responsible
(2) Responsible mental life is subject to reactive attitudes, such as praise and blame
(3) Attitudes subject to reactive attitudes are necessarily self-known
(4) So truly agential mental life is necessarily self-known (2008a, p. 477)

The idea, as we will see, is that the *necessity* of self-knowledge, as concluded at (4), entails its privilege and peculiarity. Moreover, its necessity is something to be gleaned from its connection to the aforementioned normative notions, not from anything 'purely epistemic'. Before we evaluate the argument, let me clarify a few of its key terms and elaborate on some of the basic machinery at play in it.

First, since the argument only applies to the self-knowledge *of agents*, self-knowledge is not supposed to be necessitated *outside of conditions of agency*. Loosely specified, these are whatever conditions—psychological and otherwise—ensure that we are able to exercise our agency freely and responsibly, such as having the right conceptual repertoire and rational faculties. Sometimes, these conditions can be compromised, and some minded creatures (lower animals, infant humans) may never meet them. In these cases, the argument will not apply.

Second, as regards premise (2), not all mental states are, for Bilgrami, the appropriate objects of reactive attitudes of praise or blame. To take a simple example, there is no appropriate sense in which my pains or tickles are praiseworthy or blameworthy. Thus, like many other agentialists, Bilgrami restricts his account of self-knowledge to intentional, propositional attitudes like beliefs, desires, and intentions.[143] Momentarily, however, we will see that the argument is restricted even further, such that it applies primarily to beliefs, desires, and intentions of a certain kind.

Third, still focusing on premise (2), for an attitude to figure into a "responsible mental life" is for it to figure as a possible or actual ingredient in a rationalizing explanation of freely undertaken intentional actions. The idea, to clarify, is not merely that we are responsible for our attitudes when they figure into a specific action-explanation at some time T_1; rather, we are responsible for our mental states inasmuch as we *can* freely act on them at a given time.

Fourth and finally, now focusing on premise (3), saying that "attitudes subject to reactive attitudes are necessarily self-known" means that we do not know them by way of merely highly reliable epistemic procedures. This is because it is always possible for an agent to fail in executing an epistemic procedure for acquiring self-knowledge, even when her agency is not compromised in some way, and so no such process yields self-knowledge necessary. For Bilgrami, then, privileged and peculiar self-knowledge is maximally privileged because necessary, and it is peculiar because it cannot be arrived at by any fallible route.

Let us now begin our reconstruction of the argument by focusing on each step, beginning with premise (1): truly agential life is responsible. What supports this premise? For Bilgrami, the relevant conceptual ingredient here is that of *freedom*. In brief, the idea is that it is because

[143] Most of his discussion centers around beliefs and desires, though he occasionally makes clear that his arguments ought also to apply to intentions (2006a, p. 208).

agency is only agency to the extent that it is free, and because *freedom* entails responsibility, agency transitively entails responsibility. Now, because our beliefs, desires, and intentions figure essentially into explanations of free action, they too are implicated in our free agency, and so we are responsible for them as well.

Consider now premise (2): responsible mental life is subject to reactive attitudes, such as praise and blame. is Bilgrami's adoption of a broadly Strawsonian (1974) conception of agency, freedom, and the reactive attitudes. Specifically, he draws on Strawson's compatibilist argument concerning free will and determinism. Prior to Strawson, much of the debate about whether freedom could be reconciled with our embeddedness in a world of causes turned on whether we could carve out a principled distinction between coercive and non-coercive causes of behaviour. Strawson moved away from these increasingly insular and arcane disputes about metaphysical criteria for distinguishing the coercive from the non-coercive. Instead, he focused on the indispensability of our reactive-attitudinal practices, thus arguing that our (fitting) reactive attitudes are themselves a guide to detecting free and responsible actions. This might seem question-begging, for it might seem to presuppose the aptness of the reactive attitudes. However, Bilgrami argues we can justify our practices of praise and blame "*from within our attitudes and evaluations*" themselves (2006a, p. 62). Thus, we arrive at premise (2).

Having argued that agential mental life is responsible, and that responsible agential mental life is subject to reactive attitudes, Bilgrami argues that (3): attitudes subject to reactive attitudes are necessarily self-known. This is because being appropriately subject to a reactive attitude means being required to accept praise or blame for it, and because one cannot do this if one is ignorant of one's attitude.

Finally, from (1)-(3), we get (4): truly agential mental life is necessarily self-known. Indeed, this is a conceptually necessary conclusion, since the argument is based on conceptual relations between agency, responsibility, freedom, and self-knowledge. As aforementioned, self-knowledge of our attitudes is privileged because it is maximally reliable qua *necessary*, and it is peculiar because no third-personal route to acquiring self-knowledge can deliver self-knowledge as a matter of necessity.

§4.2.2—Bilgrami's Agentialism: Clarifications and Further Features

I have said that, for Bilgrami, it is propositional attitudes specifically that are appropriately subject to praise or blame, and so it might seem that *all* propositional attitudes are necessarily self-known (again, in 'conditions of agency'). However, this is not really Bilgrami's view. Instead, he argues that there is a "deep ambiguity" (2012, p. 264) in our concepts of belief, desire, and intention, and that clarifying this ambiguity limits the scope of his agentialist account.[144] The ambiguity, in his terminology is between propositional attitudes as *dispositions* and as *commitments*. As we will see, the idea that we can be appropriately praised or blamed for our attitudes is supposed to be a necessary feature of *commitments* only, whereas reactive attitudes toward dispositions are only sometimes apt.

To better grasp the disposition/commitment distinction, imagine that you frequently watched horror movies as a kid. Now an adult, you often find yourself looking over your shoulder when in a dark hallway. In so doing, you arguably manifest a belief of sorts, say, the belief that a ghost might be behind you. You may even be aware of your belief and admit that it is not based on good evidence. But you may find that it is hard to dislodge from your psyche all the same. Such a belief is properly viewed as a dispositional belief—it is a tendency to behave in

[144] It is worth noting that not everybody finds this ambiguity so compelling (cf. De Bruin et al. 2015).

ways commensurate with the truth of *p*, but it is not a belief that is inherently rationally constrained (though there may be cases where one's dispositional attitudes are not judged by us to be irrational—more about this in §4.4.1).

Now imagine an alternative scenario, in which you believe in ghosts as a result of having formed a view on the matter through sustained research. You thus accept various epistemic obligations to forfeit your belief in the face of competing evidence. In that event, your belief is a commitment: it is a judgement-sensitive attitude that is sensitive to rational pressures, and so is not a mere tendency to behave in certain ways that are commensurate with the truth of *p*. Similarly, you might have a desire to smoke, but this may be either a mere tendency or, alternatively, a desire that is sensitive to the (by your lights) reasons in favour of smoking. In the former case, your desire is a disposition. In the latter case, it is a commitment.

One of the primary distinguishing features of commitments is that they are essentially normative in ways that dispositions are not. As Bilgrami puts it, with commitments "I *ought* to believe various other things, even if I don't actually believe them, just as my desires commit me to do various things, even if I don't do them" (2012, p. 265). To form beliefs or desires as commitments is to incur further obligations. It is to incur obligations to reason and act in ways that accord with the truth of *p* or the goodness of φ-ing (and perhaps, as regards intentions, the to-be-doneness of φ-ing, though Bilgrami does not discuss intentions very much). This is not true of our dispositions: because they are not our commitments, we are not obligated to do anything in accordance with them.

Bilgrami argues that there is only a conceptually necessary connection between commitments and self-knowledge (given, as always, conditions of agency). The reason is that, because commitments engender obligations, failing to live up to one's commitments renders one

blameworthy in a way that requires one to be able to accept criticism for failing to live up to them. But one can only accept such criticisms if one is aware of one's commitments as such. In other words, having commitments entails having a *second-order* disposition to accept praise or blame for failing to or succeeding in living up to them. However, it is still the case that one's *first-order* state is one's commitment proper. To say that commitments are irreducibly normative, then, is to say that even if one fails to act in ways commensurate with one's first-order commitment, the correct conclusion to draw is not that one lacks that commitment, but that one has failed to live up to it in one respect or another.

Contrariwise, since we have just seen that our dispositions are not necessarily attitudes for which we are responsible, they are not necessarily self-known. Nevertheless, Bilgrami admits that our dispositional attitudes themselves *can sometimes* be the appropriate objects of reactive attitudes, so long as they "potentially go into the production of…an action" that is "free and accountable" (2012, pp. 267-268).[145] In other words, it is not just commitments that are self-known in virtue of their role in (or readiness to figure into) free and responsible action-explanations, but it *is* just commitments that are *necessarily* self-known in conditions of agency. So Bilgrami allows that we can be praised or blamed for our dispositional attitudes on certain occasions. When this is so, we must have self-knowledge of them as well.

One might wonder about Bilgrami's conception of the *functions* of self-knowledge. One point that Bilgrami makes is that our commitments:

> …are relevant to revisions of belief that are not brute but are revisions of belief that are *mediated by reactive attitudes of criticism that we might have toward our first-order beliefs*…To lack this deliberative, evaluative perspective is to lack a 'first person point of view' (2006a, pp. 180-181).

[145] See also Bilgrami (2006a, p. 206).

This is essentially an appeal to Burge's account of critical reasoning (see §3.2). But this won't satisfy anyone who is skeptical of that account for any of the reasons offered in Chapter Three (see §3.3). Fortunately, Bilgrami does not take the functional import of self-knowledge to be exhausted by its role in critical reasoning. Instead:

> Once the underlying issues of agency and the first person point of view are brought out into the open, it has become clear that self-knowledge and its special character depend not just on considerations of rationality, not just on processes of revision that reveal a rational sensitivity to incoming information, nor on exemplifications of rationality…it depends on being the subject of reactive attitudes whether of one's own or (justifiably) others. (2006a, p. 182)

Similarly:

> The point…is that it is not just *self*-reactive attitudes that presuppose self-knowledge but *any justifiable* reactive attitudes—even those that *others* might have toward one's actions and intentional states or that one may have toward *others'* intentional states and actions. (2006a, p.184)

These passages allow us to see that, for Bilgrami, self-knowledge makes possible the interpersonal practice of holding one another accountable for our actions and attitudes.

In summary, Bilgrami argues that free agents necessarily possess self-knowledge because, where such agency is present, whatever mental states figure into its exercises are such that we are responsible for them and, therefore, must be able to reckon with fitting criticisms or praises of them. Such self-knowledge is privileged and peculiar in virtue of its conceptual necessity, given conditions of agency—always, for commitments, and sometimes, for dispositions. Given this picture, we can see why self-knowledge is not just another "narrow epistemological theme" (2012, p. 277). It is, rather, a theme in a much wider vision of free, responsible agency.

§4.3.1—Objection to Bilgrami: The Case of Emotions

Alessandra Tanesini (2008, p. 241) argues that *emotions* are often the appropriate objects of reactive attitudes, even though these are completely left out from Bilgrami's account. Now,

Bilgrami might deny that emotions are the legitimate objects of reactive attitudes. But it is hard to deny that we often criticize people for their anger (to take just one example), and that we do so with justification. Alternatively, Bilgrami could argue that emotions figure into free and responsible action and so are necessarily self-known. In this case, however, there seem to be obvious counter-examples, such as when we are spiteful or jealous but (1) do not realize that these emotions motivate our actions and (2) are nevertheless blameworthy for our spitefulness or jealousy (Ibid., p. 242). On either strategy, it seems that Bilgrami has a problem accommodating emotions. If emotions cannot be adequately accommodated within the conceptual web that Bilgrami has built, despite the fact that they seemingly should be (or should be pushed out for principled reasons), this gives us reason to think that the conceptual web Bilgrami has constructed for us is not nearly as tightly woven as he has claimed. In Tanesini's words, it shows that there can be states "within agency" that are not self-known, even as those states contribute to our actions in just the sorts of ways that propositional attitudes do.[146] How, then, can we argue that Bilgrami has established conceptual truths tying self-knowledge to free, responsible agency?

§4.3.2—Objection to Bilgrami: Too Much Self-Knowledge?

Now recall Bilgrami's claim that, while dispositional attitudes are not necessarily self-known in conditions of agency, they are self-known so long as they happen to figure into a free and responsible action. Coliva has claimed, in response, that this "is oxymoronic to some extent":

> For if a propositional attitude is something that happens to one or that one finds oneself saddled with, like a sudden urge to smoke a cigarette, in what sense can a reactive attitude against *it* be legitimate? One cannot be blamed for that urge but, at most, for realising it and yet indulging in it…If, in contrast, one thought that the very urge could be legitimately blameworthy, then the distinction between propositional attitudes as dispositions and as commitments, based just on an appeal to justifiable reactive

[146] An account of responsibility without (self-)awareness can be found in Sher (2009). I have yet to explore how exactly Sher's arguments might impact Bilgrami's.

> attitudes, would founder. For they would just be treated on a par vis-à-vis the justifiable reactive attitudes they would elicit. (2016, p. 187)

There are two objections here. The first serves as a sort of inverse objection to Tanesini's. Recall: Tanesini worries that there are emotions for which we *are* praiseworthy or blameworthy despite their *not* being self-known, contra Bilgrami. Coliva, on the other hand, argues that dispositional propositional attitudes are *not* plausibly blameworthy or praiseworthy, no matter what sort of action-explanations they (actually or potentially) figure into. The second objection arises if we insist, contra the first objection, that we are praiseworthy and blameworthy for our dispositional attitudes. For it now becomes unclear what the difference between dispositions and commitments is supposed to be. After all, commitments are supposed to be attitudes for which we are responsible in virtue of the obligations they engender, and in virtue of their contributing to our free agency. But if we can be blamed for our very dispositions themselves, the risk seems to be that they will also inherit these commissive properties, even if they only possess these properties in certain situations. At this point, the disposition/commitment distinction becomes problematically blurry.

Coliva, who happens to be sympathetic to a version of the disposition/commitment distinction, thinks that we should reconstrue our understanding of these attitude types. I am going to present her version of this distinction now. After doing this, and after modifying her conception in a few respects, I will present Coliva's own agentialist account of privileged and peculiar self-knowledge. I will then critique her account, thereby paving the way for my own.

§4.4.1—Coliva's Agentialism: Privileged Self-Knowledge and Rational Responsibility

Coliva's agentialist account closely resembles Bilgrami's in several respects, not least of which being her embracement of a distinction between dispositional and commissive propositional

attitudes. However, she also refines this distinction in important ways. These refinements illuminate a *certain kind* of responsibility we bear for our commitments but not our dispositions, and so helps to clarify why only commitments are ever self-known in so-called conditions of agency.

Let us begin with her take on the commitment/disposition distinction. Coliva begins by suggesting that the following properties are essential to dispositional propositional attitudes:

> (a) these mental states are not the result of a conscious deliberation, like a judgement, on a subject's part, based on considering and, in particular, on *assessing (or even being able to assess) evidence in favour of* P (or of P is worth pursuing, it would be good if P happened, etc.);
>
> (b) these mental states are not within one's direct control, being rather something one finds oneself saddled with;
>
> (c) hence, these mental states are not something one will be held rationally responsible for. (2016, p. 28)

Ted Parent (2019, p. 325) helpfully catalogues some other useful descriptors from Coliva's subsequent discussion of dispositions. Thus, dispositions are also (with citations corresponding to Coliva 2016):

> (d) often credited to a-conceptual animals to make sense of their behavior (p. 28),
>
> (e) not always self-known (p. 29),
>
> (f) often self-known only through a process of self-interpretation (e.g., on the therapist's couch, when explaining one's behavior) (p. 29),
>
> (g) often predictable, either through inference or simulation, as having an influence under specific types of circumstance (even though they are not under one's direct control, as per (b)) (p. 30),
>
> (h) characterized in functionalist terms (p. 255, n. 28), i.e., they are explanatory mediators between sensory inputs and behavioral outputs (cf. p. 256),
>
> (i) not intrinsically normative (p. 255, n. 28; p. 256).

As Parent points out, there is surely some conceptual overlap between Coliva's core properties (a)-(c) and the additional properties (d)-(i). Nevertheless, I also agree with Parent that "the additions provide further clarity" (2019, p. 325), even if there is room for disagreement about some of them.

Where might one disagree with Coliva? One complication is sure to arise in thinking about animal minds and other potentially 'a-conceptual' creatures, such as infant humans. For, on some views, it is not possible to ascribe determinate propositional attitudes of any such kinds to these creatures, insofar as it is not possible to tell a story of how they fix the contents of these attitudes.[147] Similarly, one might think that descriptors like the above give us reason to think that we are dealing not with potentially dispositional attitudes like belief, but with a sui generis sort of mental state like what Tamar Gendler (2008) calls *aliefs*. Aliefs are defined as habitual, automatic, associative and arational behaviours (2008, p. 641). This can sound a lot like states we are simply saddled with, and that we attribute to agents to make sense of their behaviour without any intervening rational agency. But there are good reasons to resist reducing dispositional attitudes (chiefly, beliefs) to aliefs. First, because dispositional beliefs can be, e.g., insensitive to evidence or evidential assessment, we can take it that it would be better to excise them from our psychology (for example, as with a dispositional sexist belief). But if dispositional beliefs are really aliefs, then they are arational, and it cannot rightly be challenged as a cognitive failure (Borgoni 2018a, p. 214). I believe, however, that many attitudes we might characterize as dispositional are such that we can take them as *irrational* to hold, even if we are not rationally responsible for forming them (e.g., implicit biases). Moreover, since aliefs are associative and habitual, we should

[147] This is a broadly Davidsonian move (cf. Davidson 1984).

not expect them to have such robust, holistic effects on behaviour and thought. Aliefs are atomized in their effects—they are often indexed to certain visual experiences, for example (Gendler, ibid.). But with a dispositional belief—e.g., a dispositional sexist belief—we can imagine it having quite robust, wide ranging effects on an agent's psychology and behaviour (Borgoni 2015, p. 214).

Crucially, though, while Coliva refers to dispositional attitudes as those that are often ascribed to a-conceptual creatures and creatures that have to interpret their attitudes in order to know them,[148] this does not mean that dispositions *cannot* be responsive to evidence, such that we are inevitably bound to resent their existence whenever we become aware of them. Rather, the claim is only that they are not dependent on *assessments* of evidence. Thus, I might be taken to believe that *p* because I respond reliably to the fact that *p* in some behavioural respect, even if I have never actively evaluated the evidence for *p* and hence never *judged* that it is true on any such basis. In some such cases, my behaviour may be perfectly rational, and so there is no implication that dispositional attitudes are always aberrations on our otherwise rational natures. This is not something that Bilgrami always makes clear, though Coliva is more insistent about this.

Perhaps the biggest difference between Bilgrami's conception of dispositional attitudes and Coliva's, however, is her claim that dispositions are not intrinsically normative because we are not *rationally* responsible for them. To see what this means, and why it is important for distancing Coliva's account of the disposition/commitment distinction from Bilgrami's, we must turn to her discussion of commitments. "What is essential to commitments", she begins, are the following properties:

[148] See Lawlor (2008b) for more about how this might go in the case of belief.

> (a') that they are the result of an action—the mental action of *judging* that P is the case (or worth pursuing/having)—on the subject's part, on the basis of considering and hence of *assessing* evidence for P (is worth pursuing/having) ["or at least of being disposed to do so, if required" (2019, p. 4)];
>
> (b') that these mental states are (at least) *normatively constrained* —that is, they must respond to the principles governing theoretical and practical reasoning;
>
> (c') and, in particular, they are so constrained (also) *from the subject 's own point of view*;
>
> (d') that they are mental states for which the subject is held *rationally responsible*. (2016, pp. 31-32).

As with Colivan dispositions, Parent (2019) catalogues some further properties of commitments that Coliva (2016) describes in her subsequent discussion of them. Thus, commitments are also:

> (e') in the subject's control (p. 31), although this does not mean one can adopt any commitment by an arbitrary act of will (p. 33),
>
> (f') [typically, and ought to be[149]] co-instantiated with "dispositional elements," e.g., a belief-commitment comes with dispositions to use the belief as a premise in reasoning, to allow challenges to the belief, to produce evidence in response to challenges, yet to possibly withdraw the belief if the challenges are evidentially powerful enough, and to adjust other relevant commitments in light of the withdrawal…(p. 33)
>
> (g') had by some subjects even without an attempt to pursue the action recommended (if any) by its content (p. 259). E.g., It is possible for one to have a commitment to the ethical belief "I should donate to charity," even if one never actually donates to charity. (It is a case of "failing to live up to one's commitments.")
>
> (h') are first-personal rather than third personal in nature. Minimally, this means that if one has incompatible social (third-person) commitments, as when one makes conflicting promises to different parties, this does not suffice for incompatible first personal commitments (p. 260)[150]

[149] Coliva (2016, p. 33)

[150] Coliva (2019a) adds that, if one has a commitment to *p*, then forming a commitment to ~*p* automatically excises one's commitment to *p* from one's psychology. So, it is constitutive of commitments that they cannot be contradictorily held, though one can simultaneously commit to *p* while *dispositionally* believing ~*p*.

Once again, Parent takes these additions to provide further clarity even if there is some conceptual overlap between them and the initial properties (a')-(d'). One final clarificatory point, however, is that the distinction between commitments and dispositions is not a distinction between occurrent and standing attitudes, at least for Coliva (I am not sure about Bilgrami's view on this, though I suspect he would agree). Both sorts of attitudes can be standing or occurrent, short-lived or long-lived. This might generate some confusion, since if both attitudes can be standing then they might be thought to involve dispositional components, and so one might wonder anew what separates them. But this confusion, while understandable, hinges on a failure to distinguish between the *types* of dispositional features that constitute either sort of attitude (and to forget that commitments also have *normative* properties that dispositions lack—normative properties that constitute one's commitments over time). After all, commitments are individuated by their judgement-sensitivity, their being constrained by norms of reason, and their being constrained from the agent's point of view. These dispositional properties involve a readiness to defend or retract them in the face of recalcitrant evidence that one has assessed or is prepared to assess. Dispositions are not necessarily so individuated, even though they are not necessarily irrational as a result.[151]

Now we can say something about the idea of rational responsibility. To be rationally responsible for one's attitudes is to be responsible for adjusting them to rational pressures, and for accepting *epistemic* criticisms of them if one fails to live up to them. For example, I can be condemned for my inconsistency if I fail to act on my commitments, and I can be criticized for failing to adjust my attitudes as reason demands. This sort of responsibility is not necessarily

[151] While not all dispositional attitudes (or dispositional beliefs specifically) need be recalcitrant to countervailing reasons, some of them will be. This is a problem if one follows Helton (2018), who argues that any attitude that is recalcitrant to countervailing reasons is not a belief. See Viedge (2018), however, for the opposite view.

morally salient. For Coliva, this is by design: she wants to avoid the Bilgramian thought that responsibility for one's commitments means being morally reproachable or resented for them (2016, p. 32, fn. 14). This seems to be because this sort of reproachability can extend beyond commitments (say, to emotions), and so makes a poor candidate for helping to individuate commitments as the kinds of states they are. Dispositional attitudes are *never* those for which we are rationally responsible, whereas commitments always are, and so a stark contrast is available.

With this distinction drawn between dispositions and commitments, Coliva produces an argument that is at least structurally similar to Bilgrami's, in that it draws on putative conceptual truths in order to establish the privilege and peculiarity of self-knowledge. In Coliva's words: "…commitments are such that one should accept criticism (or be self-critical) for not living up to them. The latter feature obviously requires knowledge of them *qua* the mental states they are" (2016, p. 190). Keeping in mind that the relevant sort of criticism is rational rather than moral, I propose to distill her position in what I call the *Rational Responsibility Self-Knowledge* principle (hereafter RRSK):

> **RRSK**: In conditions of agency, we necessarily have self-knowledge of our commitments, because we are rationally responsible for them in such conditions

Nothing like RRSK corresponds to our dispositional attitudes, because we are never rationally responsible for these. After all, they are not based on our own judgements. Therefore, while our dispositional attitudes can be self-known, this will be by third-personal methods like inference and testimony. So, whereas Bilgrami argues that dispositional self-knowledge is sometimes necessitated (as when our dispositions contribute to free and responsible action), Coliva's account has no such implication. It is only ever commissive self-knowledge that is necessitated.

§4.4.2—Why Commit to Commitments?

For a moment I want to bracket Coliva's agentialist account of self-knowledge in order to focus in more closely on her conception of commitments. I do this both to motivate the general value of a commitment/disposition distinction a bit further (this subsection) and to prepare myself to set my preferred conception of commitments apart from hers (the next subsection).

Some philosophers have denied that we should distinguish between dispositional and commissive propositional attitudes (Bar-On 2009; Zimmerman 2019). They sometimes—unflatteringly—describe those who do draw this distinction as espousing "mental-mental dualism" (Bar-On 2009, p. 62). The criticism behind the mental-mental dualism accusation seems to be that the mental states of very young children and animals—those who, plausibly, cannot possess commitments—therefore occupy a "lower class" of mindedness (Ibid., p. 62). I fail to see why this should be so: we can acknowledge that dispositional attitudes have robust explanatory roles in explaining behaviour and can exhibit rational hallmarks like responsivity to evidence (keeping in mind, admittedly, that such responsivity is not based on *assessments* of evidence). Somewhat similarly, Coliva admits that in the human case: "…there is no denying that we can end up identifying deeply with mental states of ours we find out through inference and self-interpretation; nonetheless, it is because they are initially alienated from us that we have to go through the whole procedure, and we wouldn't have to do that if they were just the result of deliberation [as our commitments are]" (2015, p. 251).

At any rate, it is eminently plausible that there are creatures who cannot actively evaluate evidence and conform their attitudes to rational norms via deliberation, and so it is not plausible to attribute commitments to them. If anything, then, having a univocal conception of the propositional attitudes makes it harder to attribute attitudes to such creatures at all. This may be a

consequence to embrace depending on one's views about animal and infant human minds. But setting those cases aside, Bilgrami and Coliva are not attempting to provide a highly technical distinction between attitude types; rather, they are trying to clarify one that is—if only inchoately—already at play in our general folk psychology. Some of our mental states are those we stand behind and hold ourselves accountable for. Some have more subterranean influences on us. They may each have functional profiles characteristic of a given kind of attitude (belief, desire, and intention), but the differences between them are not to be dismissed.

Why else might one prefer this bipartite conception of the propositional attitudes? Well, this conception has explanatory potential in a variety of contexts (Coliva 2019b). For example, the distinction might play a key role in explaining self-deception (see §5.7.2). Relatedly, it might help to explain why *some* utterances or thoughts with a Moore-paradoxical form are not really Moore-paradoxical. Thus, we sometimes find that when agents say things like "My husband is faithful, but I guess I just don't believe it!", the correct thing to do is not blame them for their blatant inconsistency. Granted, they are pointing to an inconsistency among their attitudes, but the inconsistency is not one between two attitudes that they fully endorse. We can understand them as at once committed to their husband's faithfulness while also recognizing that they have a countervailing dispositional attitude—a belief rooted in a deep insecurity, perhaps. Such an agent expresses a rational tension in her attitudes, but it is not one that she should be epistemically criticized for asserting. Her assertion is perfectly cogent. It is not one that we would demand she retract, even if we subsequently agree that she must surely set her psychological house in order.

§4.4.3—A Dispositional Component of Commitments

In this subsection I want to propose a modification (or perhaps clarification) of Coliva's account of commitments. To begin, recall Bilgrami's claim that one can have commitments even if one is

"not disposed to do *anything*" to live up to them (Bilgrami 2006, p. 227, emphasis mine), so long as one has a second-order disposition to accept criticism for failing to live up to them. At a glance, Coliva makes a similar claim in asking us to:

> …consider the case where you have the commitment to help the poor, while not giving to charity (or doing anything else which would help them). On a resolute notion…there would be no real commitment. However, this seems fast. For, as long as you are prepared to be self-critical or to accept criticism for acting the way you did, then you could still count as having the commitment. (2019a, p. 9)

Once again, such preparedness involves a *second-order* self-belief that one has about one's commitment, rather than anything psychologically first-order.[152] I want to argue, however, that there is at least one necessary first-order psychological component of commitments. This is a disposition to use one's commitments as premises in (theoretical or practical) reasoning. One way to press the point is to note how strange it would be to hold someone accountable for a commitment that was totally inferentially isolated, such that the agent who had the commitment was not at all disposed to think in its terms. Indeed, being able to accept criticism for failing to live up to one's commitments seems to depend on one's appreciating that one's actions do not align with one's commitments, and this knowledge strikes me as itself inferential in nature. It is inferential because I have to draw an inference from facts about my actions, to facts about my mental states, to facts about my blameworthiness.[153] Not only this, but to be *rationally* responsible for my commitments, as Coliva argues I am, my commitments surely have to be at

[152] Christopher Campbell points out that Bilgrami's reference to a disposition to accept criticism *for one's commitment* differs from one's disposition to accept criticism for *acting as one did*. This is interesting if the latter is not second-order in the way that the first is. Turning to Coliva, I cannot be certain of her view, but given how she elsewhere describes our rational responsibility for our commitments themselves and the second-order disposition this entails to accept criticism for failing to live up to our commitments (if and when we do so fail), I assume that she has something distinctively second-order in mind here.

[153] Perhaps, in some such inferences, my commitment won't actually be a premise, but my commitment will nevertheless count as inferentially engaged so long as my inference from facts about my actions to facts about my blameworthiness proceeds *in light of* my commitments. Thanks to Christopher Campbell for pressing this point.

least somewhat responsive to counter-evidence and the like, or to my other commitments, and it is plausible that this responsiveness often depends on my inferential competencies.

Granted, as we saw, Coliva herself says that commitments:

> (b') …are (at least) *normatively constrained* —that is, they must respond to the principles governing theoretical and practical reasoning

…and so one might think that she is already on board with the previous suggestion. This may be so, but I clarify it here in order to prevent any misunderstandings that might arise from reflection on statements like Coliva's and Bilgrami's above, where it can seem as though only second-order dispositions are taken to be constitutive of commitments.

In arguing this, I am prepared to accept that just about any other dispositional elements that are typical or ideal of commitments can be absent without one's thereby lacking a commitment. Note also that I am not hereby suggesting that we should be disposed to draw *every* inference that our commitments could potentially figure into. No doubt, some sound inferences from our commitments will outstrip one's cognitive powers, and other inferences will be too far removed from the cognitive contexts in which one finds oneself. Moreover, it seems possible for situations to arise in which an agent can be asked for a reason for her commitment, while being unable to give any (Moran 2012, p. 215). This suggests that she lacks access to inferences that take her commitment as a conclusion. But this does not mean that her commitment is totally inferentially isolated. Nor does it mean that she lacks the commitments that support the commitment at issue, for it may simply be that their inferential availability is masked at the time. Still, she should at least be able to recognize the applicability of the demand. Moreover, she should still be able to consider at least some of what potentially *follows from* her commitment. I thus conceive of

commitments as (a')-(h') above, but also as (i'): being constituted by a disposition, however broad, to infer in light of them.[154]

Highlighting this feature of commitments can also help to clarify how a broadly Colivan conception of commitments differs from an earlier conception of them, due to Brandom (1994). On Brandom's conception, being committed to *p* entails being committed to various other matters *q*, *r*, *s*, whether or not one has ever considered these other matters. This has obvious and immediate implications for self-knowledge: not all of an agent's commitments are self-known if true attributions of commitments such as these are possible (Levine 2009, p. 100), even in conditions of agency. On the account of self-knowledge that I intend to develop, just as it is for Bilgrami and Coliva, this sort of result cannot occur. And now we can see one more reason why the (independently plausible, on my view) point about the inferential availability of commitments is important. For one cannot, on my usage, be committed to a proposition *p* if one has never so much as considered it or would not be prepared to acknowledge it if asked: one cannot be disposed to use *p* in an inference if one has never so much as considered its content as true.[155] After all, if drawing an inference from *p* to *q* requires *taking p* to support *q*, this requires that the agent have cognitive contact with that proposition (see §3.4). But even if the Taking Condition (TC) is false, the main point is still plausible: drawing an inference plausibly requires one's having *some* sort of cognitive contact with *p*, even if this does not take the form of a taking-attitude about the logical relations between *p* and whatever one infers from it. So, on my

[154] Before moving on, I want to flag a concern about the possibility of commissive desire. It has been suggested to me that desires are not rationally sensitive to our other attitudes in the way that other attitudes are. For comparison: if I have two incompatible intentions, then I will surely forfeit one, but this is not true of desires, even putatively commissive ones, since I may find that I am equally committed to the goodness of φ-ing and ψ-ing even if I can only intend to φ *or* ψ. On my view, however, this only shows that desires are less *intra*-attitudinally rationally sensitive than intentions. Inter-attitudinally speaking, commissive desires may still be highly rationally sensitive to other attitudes. For example, if I believe that φ-ing is not good to do, then I will not *commissively* desire to φ, since my commitment will not be based on judgement to the effect that φ-ing is good to do.

[155] Though often, were one to consider *p*, one will *thereafter* be disposed to draw inferences from it.

view, commitments are less easily ascribed to agents than they are on the Brandomian conception.

§4.4.4—Rational Responsibility and the Self-Awareness Condition on Inference (Redux)

I have spent some time refining my preferred understanding of commitments, one that is largely in line with Coliva's. One final thing to do before evaluating Coliva's agentialist account is to contrast it with a structurally similar agentialist account that can be derived from Paul Boghossian's remarks about the Taking Condition on inference (§3.4). My reason for doing this here is that, in criticizing Boghossian's agentialist account, we will have drawn out further features of Coliva's agentialism that I will criticize in turn.

We already saw in §3.4 that Boghossian is a proponent of:

> **(Taking Condition)**: Inferring from *p* to *q* necessarily involves the thinker *taking p* to support *q* and drawing *q because* of that fact. (2019, p. 110)

However, in §3.4 I set aside one argument of his for how TC might lead to an agentialist account of self-knowledge. That argument, which I will now discuss, begins from his claim that inferences are not only epistemically evaluable in themselves, but are such that *those who draw them* are epistemically evaluable in their doing so (2019, p. 113). Boghossian argues that TC can purportedly explain this: we are epistemically evaluable for our inferences because *our* taking *p* to epistemically support *q* leads us to conclude *q*. In Markos Valaris' words: "if you make a bad inference, we can legitimately criticise *you* as having been hasty, irresponsible, biased, and so on" (2017, p. 2010), where such vices are reflected in one's epistemically shoddy taking-attitudes. Indeed, the idea seems to be that we are *rationally responsible* for our inferences and can be criticized for inferring badly, just as Coliva argues that we are rationally responsible for our commitments.

If Boghossian is right about this, then perhaps drawing an inference also requires self-knowledge of one's inferential mental states (again, in just the way that Coliva thinks is true of having commitments). For, if we are rationally responsible for our inferences, then we should be able to accept criticism for inferring badly, and this plausibly requires us to understand ourselves as psychologically active with respect to those mental states that figure into our inferences (think of Coliva's RRSK, modified to refer to inferences). On this reading of Boghossian, self-knowledge is essential to inference inasmuch as we are rationally responsible for our inferences. Moreover, such self-knowledge may be privileged and peculiar, for the reasons given in §3.4.1.

In the next subsection, however, I will develop a criticism of Boghossian's agentialist account by considering a further feature of Coliva's agentialism. Then, in the subsequent subsection, I will criticize Coliva's agentialist account.

§4.5.1—Objection to Boghossian's Agentialism

To evaluate Boghossian's agentialist account, I want to draw our attention to Coliva's view of how we acquire the *concepts* required to self-ascribe our commitments. Here is the story in full:

> Take a subject who is able to judge that P, give evidence in favour of it and withdraw from it if required and, therefore, has the first-order belief that P based on judgement. Suppose you ask her "Do you believe that P?" and she is unable to answer. You conclude that she does not have the concept of belief. In that case, you would simply train her to the use of that verb by *drilling* her into using the expression "I believe that P". You teach her to *substitute one form of behaviour*—one kind of expression of her mind, that is, the outright assertion of "P" accompanied by the ability to give reasons for it, which manifests her first-order belief (based on judgement)—*with another*, that is, the assertion of "I believe that P"…
>
> …"I believe that P" and "I desire that Q" are taught *blindly*, as alternative expressions of one's mind: they are *ingrained* as alternative ways of expressing one's first-order beliefs and desires (based on judgement) other than asserting that P or that Q would be good to have. Hence, in this account, there is no inner epistemology, just a substitution of one form of behaviour with another. (2016, pp. 191-192)

I will call this *the blind drilling account* (of propositional attitude concept-acquisition). I will refer to *the subject* of blind drilling as someone who is merely at the beginning stages of blind drilling, such that, *ex hypothesi*, she does not yet possess propositional attitude concepts and so cannot yet self-ascribe her commitments.

Notice that the subject of blind drilling is trained to self-ascribe attitudes that are generated by reasoning capacities that she already possesses—capacities she has to form first-order, judgement-sensitive mental states through inferences *prior* to learning to self-ascribe them. She is described as being able to supply justifications for or otherwise forfeit her attitudes in the face of countervailing (first-order) reasons. It is only once the subject of blind drilling has been trained to replace her merely first-order expressions of these attitudes with second-order self-ascriptions of them that she counts as a self-knowing reasoner.[156] But if rational responsibility and self-knowledge go hand in hand, as Coliva's RRSK principle says, then the subject of blind drilling is not rationally responsible for her inferences, and yet can draw them anyway.

Though Anna-Sara Malmgren (2019) does not concern herself with Coliva's blind drilling story, she seems to embrace a similar possibility about the gap between our capacity for reasoning and rational responsibility:

> Small children can engage in effectual deliberation, at least of the primitive sort…But small children and most animals lack the kind of responsibility that grounds rational evaluability…
>
> …One might suggest that what makes the difference is the capacity for second-tier thought. But it's not entirely clear why, by itself, it would. The ability to (say) make judgments to the effect that such-and-such is a reason to ϕ, that this evidence outweighs that, or that a given claim implies another, might make one *better* at deliberation—in that it makes one better at conforming to the governing norms. But how is that, in turn, supposed to explain fundamental responsibility? (Ibid., p. 207)

[156] And indeed, this process is supposed to make her a self-knower, one who has self-beliefs, and not just someone who can use a self-ascriptive semantic vehicle to express her first-order commitment. For it is by acquiring this newfound *conceptual* competence that the agent can make explicit to herself the fact that it is *her* perspective on the world that she is articulating (cf. Coliva 2016, pp. 192-193).

Setting aside Malmgren's talk of "effectual deliberation" of a "primitive sort", the important bit in our dialectical context is her insistence that even having a capacity for second-tier inference (that is, inference that takes propositions and epistemic support relations as its objects, rather than one's own mental states) does not account for one's "fundamental responsibility" for one's inferentially embedded attitudes or for one's inferences themselves. If this is how we should read Malmgren,[157] then Coliva's subject of blind drilling is not rationally responsible for her attitudes either, and so the two authors are in agreement.

My first point, then, is that Coliva's and Malmgren's view seems like a relevant alternative to Boghossian's. All three authors seem to take up a similar conception of rational responsibility, but Coliva and Malmgren deny that such responsibility follows directly from our capacity for inference, because they (but not Boghossian) countenance less sophisticated forms of inference (or tiers of cognitive agency) that do not require us to treat the subject as rationally responsible for her inferences. Moreover, both Coliva and Malmgren agree that "[w]hat makes the difference" to whether we are rationally responsible for our inferences (or their mental state products) "are our *introspective* and *self-reflective* capacities" (Malmgren Ibid., p. 207). If this is right, then Boghossian's view is in danger. The reason is straightforward: if rational responsibility requires self-knowledge, and there can be inferences for which one is not rationally responsible, then inferring does not require self-knowledge, whether it is privileged and peculiar or otherwise.[158]

[157] Perhaps she only means that second-tier thought cannot explain rational1 responsibility, but that we still cannot have one without the other. Thanks to Christopher Campbell for raising this point.
[158] It is possible, of course, to appeal to other features of inference to block the possibility of second-tier inference. But I have already evaluated these arguments in §3.4.

§4.5.2—Coliva's Agentialism: So What?

Interestingly, however, Coliva's blind drilling story might be problematic in its own right, even if her conclusion is correct that inference without rational responsibility is possible. The path to this conclusion will be somewhat complicated: I will begin by flagging some apparent instabilities in the relationship between her blind drilling story and her agentialist thesis that we necessarily have self-knowledge of attitudes for which we are rationally responsible. I will then consider how we might dissolve the problem. In doing so, I will lead us to the conclusion that rational responsibility, as Coliva understands it, is a superfluous property of commissive mentality.

To begin, consider Coliva's blind drilling story once more. The subject of blind drilling is said to possess judgement-sensitive attitudes that are based on reasoning; she is also said to be able to revise these attitudes in the face of countervailing reasons and to justify them in the face of epistemic demands. Eventually, she is trained to substitute first-order linguistic assertions of them (e.g., "*p*" or "φ-ing is good") with self-ascriptive expressions of them. Now, these first-order attitudes are quite clearly *not* dispositional propositional attitudes, precisely because they are based on judgement and are sensitive to counter-claims (see §4.4.1). At the same time, however, they may not be commitments, seeing as the subject of blind drilling is not yet rationally responsible for them (if she were, RRSK would dictate that she already has self-knowledge of them). Perhaps, then, these are *sui generis* attitudes, i.e., attitudes that are neither dispositions nor commitments. So that they have a name, and because they clearly have *some* commissive properties, let us call them *pre*mitments.

Now, one worry is that premissive attitudes are enough like commitments that their possessors should be construed as rationally responsible for them. After all, the subject of blind

drilling (call her Maya) is described as being cognitively sophisticated enough to argue interpersonally, since she can give reasons for (or forfeit) her premitments in conversation with her interlocutors. If this discursive intelligence suffices for her to be criticizable for her attitudes, then mustn't she have self-knowledge of them? This is a worry, at least, if one thinks that rational responsibility entails self-knowledge, as per RRSK. The point of this objection is that Coliva's blind drilling story of how we acquire the concepts needed to have self-knowledge fails if Maya is already rationally responsible for her attitudes, because we must then say that she already has self-knowledge.

Here is a possible reply. Coliva defines RRSK by describing rational responsibility in specific terms—that is, in terms of the appropriateness of and ability to accept epistemic criticism for one's attitudes. This acceptance must be, on her view, *de se*: it must amount to accepting *that my attitude* is criticizable or *that I am criticizable* for my attitude. Now, Coliva might argue that Maya is not rationally responsible for her attitude in this sense, although she is rationally responsible for it in some *other* sense that does not require attributing self-knowledge to her. For example, she might argue that Maya is 'weakly' rationally responsible for her belief that *p*, in the sense that she ought to (and can) revise it if she discovers that ~*p*, despite the fact that she is not cognizant of the fact that she believes *p*.[159]

In fact, Coliva argues that "in order to have beliefs (and other propositional attitudes) based on judgement, a subject will already have to have the *ability* to differentiate between, for instance, *believing p* and *p*'s *being the case*, by being sensitive to the fact that her point of view may be challenged—thus responding with reasons in favour of it—or indeed proved wrong—thus abandoning it." (2016, p. 192). This talk of abilities and sensitivities, rather than self-

[159] Thanks to Coliva for this suggestion.

knowledge, might suggest some content for the proposed notion of 'weak rational responsibility'.[160] Thus, imagine that Maya asserts *p*, and that her assertion expresses her premissive belief which (again, *ex hypothesi*) she lacks self-knowledge of. On the present suggestion, she is still weakly rationally responsible for her premitment because she is expected to forfeit it if she encounters sufficiently good reasons to believe *~p*.

Perhaps this proposal will work. But we are also now in a position to appreciate a certain "so what?" objection. For it may be true that once Maya has self-knowledge, such that her premitments become commitments, she becomes *strongly* rationally responsible for them in the sense of being epistemically criticizable for having them. But if Maya's discursive transactions with her interlocutors were already such as to produce rational, reasons-responsive changes of mind, whereof she was already able to provide reasons for her interlocutors to change their minds and to defend the propositions she believes, desires, or intends, it is hard to see what importance there is, cognitively speaking, to Maya's ascendency to strongly rationally responsible agency. In short, the question is: why does it matter that Maya can grow to eventually accept second-order criticisms about her commitments *as such* if she can, all the same, exercise her weak rational responsibility by adjusting her attitudes on the basis of *first*-order reasons? Whether we change our attitudes on the basis of accepting *that our commitments are problematic*, or whether we change them because we directly appreciate the reasons for or against *their contents*, the end result seems to be the same.

One might respond by claiming that I have simply shifted the goalposts for agentialist accounts of self-knowledge. After all, if Coliva has established that we are strongly rationally

[160] Notice also her careful reference to "attitudes based on judgement" rather than to commitments; this strikes me as evidence that she would indeed countenance premitments as a class of attitudes distinct from (because developmentally prior to) commitments.

responsible for our commitments, and that this explains the privilege and peculiarity with which we know them, then this is already a remarkable achievement. Perhaps this is fair. However, it now puts me in a position to present what I think is an even stronger challenge to the stability of her blind drilling story. This challenge will emerge from a part of my own agentialist account of privileged and peculiar commissive self-knowledge, an account that will also show us—if I am right—that self-knowledge of one's commitments is privileged and peculiar even if one is not rationally responsible for them (or for one's inferences, *pace* Boghossian).

§4.6.1—Social Agentialism: The Setup

Like Bilgrami and Coliva, I believe that our commitments are known with privilege and peculiarity. But I do not think that their accounts do sufficient justice to this connection. To move forward, I develop a "Social Agentialist" account of such self-knowledge.

Social, rational agents such as us (sufficiently cognitively developed) human beings are in a position to cognize a variety of social-epistemic facts. To know a social-epistemic fact is, on my definition, to know something about one's epistemic position (or *possible* or *likely* epistemic position) in a larger society of minds. According to Social Agentialism, self-knowledge is indispensable to one's capacity to *self-locate in social-epistemic space*, and it is facts about how one is able to social-epistemically self-locate that explain why one's commitments must be known with privilege and peculiarity. When an agent self-locates in social-epistemic space, she grasps how her mental states stand relative to actual or possible others. I will argue that privileged and peculiar self-knowledge of our commitments enables us to self-locate in three kinds of social-epistemic spaces, these being:

(1) *Interpersonal reasoning spaces*

(2) *Collaborative group actions and structures*

(3) *Linguistic interpretation spaces*

As regards (1), I will argue that there is, in fact, a deep relationship between reasoning/inference, on the one hand, and privileged and peculiar self-knowledge, on the other. However, unlike those arguments considered in Chapter Three (and Boghossian's in this chapter), my argument will focus on social forms of reasoning. In developing this argument I will hold off on explaining why commissive self-knowledge must be privileged and peculiar. Instead, my focus will mostly be on simply establishing that commissive self-knowledge of some sort is necessary. The privilege of such self-knowledge will be made clear once we consider my account of (2), and its peculiarity will be made clear once we consider my account of (3).

As we proceed, the reader will notice that my accounts of (1)-(3) also frequently invoke claims about the knowledge (or at least beliefs) we have about *other* minds. This suggests a question: why does social-epistemic self-locating require privileged and peculiar *self*-knowledge only? Here I simply ask the reader to bear with me—the asymmetries between self- and other-knowledge will be accounted for as I develop my arguments.

§4.6.2—Social-Epistemic Self-Locating in Interpersonal Argumentation

To begin, let us focus on social-epistemic self-locating in interpersonal reasoning contexts. While not all contexts in which we interpersonally justify ourselves to others are argumentative (one might respond to an interlocutor's simple curiosity), I will take interpersonal argumentation as a paradigm case. My claim will be that interpersonal argumentation requires self-knowledge of one's commitments.

Imagine two agents arguing about *p*. On my view, arguing with an interlocutor about *p* requires having a sense for one's *discursive position* relative to one's interlocutor. In other words, interpersonal argumentation requires that the participants appreciate that their attitudes

with respect to *p* diverge. Moreover, this divergence must be grasped *de dicto*: it is not enough, on my view, that they merely appreciate that the *propositions p* and ~*p* conflict with one another. This claim is in straightforward conflict with Coliva's own view, for we saw that she was willing to attribute interpersonal argumentative powers to Maya (our token subject of blind drilling) without attributing self-knowledge to her. Let me now explain why I disagree with Coliva's assessment of Maya's cognitive situation.

I argue that Maya's cognitive condition cannot be as Coliva describes, because it is a condition for the possibility of discursive exchanges like those Maya undertakes that she be cognizant of a divergence of perspectives. Why should this be? Here is a first pass: if Maya merely recognizes that *p* and ~*p* are incompatible (where Maya believes *p* and her interlocutor believes ~*p*), without also recognizing *who* takes either proposition to be true, we cannot understand how she can be intelligibly viewed as *responding to* her interlocutor when she interpersonally defends *p*. Failing to attribute Maya's cognizance of this dispute *as a dispute about beliefs* deprives us of an intentional description for her discursive behaviour. In the absence of such a description, she cannot be viewed as intentionally exchanging epistemic reasons with her interlocutor.

As a second pass, we can think about Maya's situation in terms of the importance of cognitions about higher-order evidence in interpersonal argumentation. Agents who enter into a debate will begin by asserting their position: *p* or ~*p*. But for an argument to even get off the ground—for there to be a motivation on the part of the agents to take it up—it is not enough that either agent simply considers the proposition, offered up by one's interlocutor, which happens to contradict a proposition that one believes. Rather, the agents need to appreciate the proposition as at least *potentially* having some countervailing evidential value. Otherwise, there is no reason

to take the countervailing proposition seriously as potentially uprooting one's belief. Now the question is: what is required in order for one to take the proposition, offered up by one's interlocutor, as potentially having countervailing evidential force against the proposition one believes?

At the very least, I believe one must have the concept of belief. Consider Maya again: if she does not grasp that the countervailing proposition *is believed by another agent*, then simply considering that proposition should not have any effect on her subsequent epistemic situation or discursive behaviour. After all, she will simply be entertaining a proposition that runs contrary to her own belief and to the evidence she has for it, and so it should not strike her as a live option at all. However, an agent who believes *p* often will and should take ~*p* seriously when an interlocutor asserts it: it is surely epistemically reasonable, in many situations, to take a countervailing assertion seriously, even if only to investigate whether that assertion is epistemically well-founded. But why, from Maya's perspective, should she take ~*p* seriously if, from her own point of view, it is false?

The answer has to do with the fact that, when Maya encounters the claim ~*p*, she does not comprehend the proposition ~*p* in an entirely neutral way. That is, she does not merely encounter it as a proposition that, by her lights, is false given the evidence she has. Instead, she comprehends it as a salient alternative—a proposition *worth disputing or considering*. To make sense of this, I argue that she must comprehend her interlocutor's assertion "~*p*" as a providing evidence of the falsity of her own belief that *p*. Maya, who believes *p*, can and should consider ~*p* as (potentially) epistemically relevant to her because she appreciates that the assertion of it is made by another agent who has a potentially different set of background information from herself. The reason why Maya can and should bother to take ~*p* seriously, even when she already

believes the opposite and has reasons for the opposite belief, is that ~*p* may be supported by information that only her interlocutor has so far considered. But this appreciation is second-order: it involves representing a divergence in the first-order epistemic situations of herself and her interlocutor.[161] In short, it is because Maya can appreciate the possibility of divergent perspectives with potentially different associated background information that she can be understood as intentionally and intelligently engaging in a debate with her interlocutor. If this is right, then Coliva's blind drilling scenario is unintelligible: Maya cannot at once reason with her interlocutor and lack a perspective on her discursive position relative to her interlocutor.[162] Crucially, moreover, the perspective Maya must have here is on her *commitments*. After all, it is her commitments that are, on both Coliva's account and mine, the kinds of judgement-sensitive attitudes that one takes to be rational and, therefore, to take as worthy of discursive application. Insofar as one's dispositions are not based on one's judgement and assessment of the evidence, they are not the kinds of attitudes that are transacted in interpersonal reasoning.

I also take this argument to address Ladislav Koreň's claim that young children who reason interpersonally merely exercise a "practical competence" in using certain kinds of dialectical devices ('no', 'but', 'so'), all without actually conceptualizing their attitudes as in conflict with their interlocutor's attitudes, thereby manifesting a "sensitivity" to rational connections between claims without any genuinely "metarepresentational" achievement (2019, p. 5). This deflationary view fails because it still cannot explain why agents ought to and do bother to take their interlocutors' assertions seriously: Maya could be sensitive to the rational connection—really,

[161] Here I understand metacognition as involving cognitive states that take lower-order mental states as their objects, rather than as involving higher-order *processes* that act on lower-order processes (cf. Proust 2013).
[162] Zimmerman (2019, fn. 2) points out independent reasons to be skeptical of the blind drilling scenario. Also note that, in rejecting the blind drilling account, I am not rejecting wholesale the idea that psychological vocabulary is taught in such a way that one learns to replace first-order expressive utterances with second-order self-ascriptive ones (see Chapter Two). I only deny that this training takes place in a communicative nexus where a developed capacity for interpersonal reasoning is already in place.

disconnection—between *p* and *~p* and, for that reason, *reject ~p*. But she could not do so after *taking ~p seriously* unless she was aware of the fact that *~p* had been asserted by an agent with a potentially different set of background information about *p* than she possesses. And this, I have been arguing, requires that she know her own commitment to *p*.

If it helps to address the deflationist, note that concerns about "overintellectualizing" the metarepresentational capacities of young children such as Maya may be empirically undermotivated in our context. After all, we know that around the age of three children start to grasp terms like 'know' and 'guess' (Johnson & Wellman 1980), as well as terms like 'think' and 'remember' (Shatz, Wellman & Silber 1983) and, finally, other mentalistic language like 'hide', 'trick', 'say', and 'tell' at this age (Lohmann and Tomasello 2003). It has even been argued that mastery of these latter terms, despite not directly referring to mental states, could play a role in enabling children to solve false-belief tasks (Turnbull et al. 2008), the solving of which is often taken as evidence of metarepresentational cognition.[163]

Even more to the point, many three-year olds *can* explicitly self-attribute mental states like beliefs and desires. This has sometimes been denied, owing to many documented false-belief tests in which the test subjects in this age range fail to do so (even where some more tacit signs of false belief awareness are present). For example, Frank Esken declares that

> The only thing which seems to be more or less uncontroversial about the development of self-consciousness is that at around 4–5 years of age humans develop…the ability to entertain higher-order thoughts of the form 'I remember that *p* ' or 'He believes that I believe *p* ' in an explicit way (i.e. bound to the usage of mental predicates such as 'believe' or 'perceive'). (2012, p. 134)

[163] Thus, even if metarepresentational thoughts / utterances involving the 'belief' predicate are relatively delayed, e.g., reliably produced only from the ages of 4-5, this is not evidence that agents cannot metarepresent beliefs at earlier ages (see also Onishi & Baillargeon 2005). Note, however, that I make an even stronger claim below.

However, others have hypothesized that many children below the ages of four or five fail to attribute false beliefs only because of the high information-processing loads that go into cognizing many of the questions and scenarios that are read out to the test subjects (Sullivan et al., 1994). Children, asked to attribute mental states to other agents after hearing a certain story, often fail to do so before the ages of four or five, as Esken says. However, Sullivan et al. adjusted a classic false-belief test (cf. Wimmer & Perner 1983) by reducing the complexity of (1) narrative structure, (2) the number of characters and episodes, (3) the length of test-questions, and (4) the complexity of the linguistic structures used to express the narratives (1994, p. 401). Their finding was that over 40% of pre-schoolers (i.e., three-year olds) could attribute false beliefs—even false *second-order* beliefs—to other agents in these conditions. This suggest that even the youngest of agents (who we might consider to be potentially capable of interpersonal reasoning) can attribute attitudes to themselves and others.[164,165]

Now to consider some further objections. First, it might be thought that there is a simpler agentialist argument that leads to the same the same result. On this account, it is simply the possibility of asserting itself, in a discursive context or otherwise, that requires one to have self-knowledge. The idea is that assertion—being a kind of intentional act—is subject to a certain

[164] Perhaps we can imagine young agents who appreciate, in some sense, their caregivers' states of mind and form similar states of mind as a result, out of a natural desire to be like their caregivers, without (yet) being able to reason with them, and indeed without fully comprehending the caregivers' *sentences* that express those attitudes (Doyle 2017). There are also early ontogenetic precursors to the argumentative exchange of reasons, as when young children simply express "no!" in response to a speaker's claim, without yet having learned to evaluate propositions as true or false (Moll 2013, p. 343).

[165] It is plausible, to be sure, that there are some representations of mental states or actions that are not metarepresentational. For instance, Butterfill (2016) argues that some tracking of perceptions and beliefs are like this. The cases he describes are cases where, under conditions of heavy cognitive load, the agent is nevertheless able to track an agent's beliefs indexed to salient objects in a shared environment. Inasmuch as interpersonal *reasoning* goes beyond immediately (perceptually) available objects, this is no counter-example to the claim that interpersonal reasoning requires a more robust form of metarepresentation. Similarly, Gómez (2005, pp. 74-75) argues that certain lower primates can represent the *attendings* of others, where attendings are represented as relations between the attending agent and a perceived object of attention, rather than as an internal state. Once again, however, these representations are bound to immediately available features of the shared environment, and are limited to attendings to perceived objects rather than to whatever mental states an interlocutor might express in an argument.

"why?" question just as any other intentional action is. For assertions in particular, the question might be "so, you really believe that *p*?" (Marcus 2016, p. 384) or "why do you believe that *p*?" (Boyle 2009, p. 151). Being unable to answer a question such as these means that one is not intentionally asserting at all. And for this reason, it follows that there is a conceptual tie between intentionally asserting and having knowledge of the belief your assertion expresses (Marcus 2016, p. 384).[166] If this is right, then what is all the fuss about having to have a grip on a divergence of perspectives between oneself and one's interlocutor?

My response is that self-knowledge must be possessed prior to the act of assertion. This is because, to be an interpersonal reasoner, the participants need to comprehend or at least begin to investigate their differences of epistemic position, since this is what gives a *point* to their *subsequent* assertoric exchanges with one another. Assertions are made, in argumentative contexts, in response to the recognition, prediction, or imagining of disagreement. These recognitions are what *motivate* discursive assertions, and so they cannot be *generated by* discursive assertions. Self-knowledge precedes the assertoric act, and so cannot be transcendentally grounded solely by paying attention to assertions themselves.

It might now be argued that self-knowledge is not necessary for interpersonal reasoning because merely *self-believing* that one believes *p*, even when one does not actually believe *p*, suffices to explain why an agent is motivated to argue with her interlocutor's assertion ~*p*. Similarly, it might be thought that one does not actually need to *know* one's interlocutor's mind, but to simply have a belief about her mind.

In response, I suggest that we must consider again the nature of the kinds of mental states at issue, namely, commitments. In §4.4.3 I argued that commitments are essentially constituted

[166] See also Marcus & Schwenkler (2018).

by a disposition to draw inferences from them. What I now want to argue is that this point has crucial implications for the possibility (or lack thereof) of false self-beliefs about one's commitments. My claim here is that, while having a commitment to *p* requires being able to infer from *p*, having a self-belief that one believes *p* suffices to dispose one to infer from *p* as well, which means that one's self-belief suffices for believing *p*, given conditions of agency. For example, if I believe that I believe the clocks are changing, I will be disposed to infer that I should change my clocks before the week's end. But this is just the same inference as I will draw if I believe that the clocks are changing, and so suffices for me to (commissively) believe that the clocks are changing. So there is a crucial difference between my self-beliefs and my other-beliefs; my self-beliefs suffice for me to have the commitments my self-beliefs are about, whereas my other-beliefs are not sufficient for you to have the commitments I take you to have.

Behind my argument here is the thought that the self-beliefs we have about our commitments have a distinctive form: self-beliefs about commitments have the formal character of *endorsing* or *ratifying* their first-order objects. It is with respect to self-beliefs of *this* kind that having them is sufficient for having the relevant first-order commitment, given conditions of agency. This is because, in endorsing one's commitment from a second-order perspective, one *self-consciously sees from the perspective of one's commitment*. I cash this out, at least partly, in terms of one's ability to see some range of the logical *implications* of one's commitment. And this is what assures that the sorts of inferences one will draw from one's self-belief about one's commitment are, broadly, the same or at least partly the same as those that constitute one's commitment (I say partly because having self-beliefs about our commitments enables us to draw a broader range of inferences from our commitments than would otherwise be possible, e.g., any chains of reasoning that necessarily involve self-ascriptions).

I conclude, then, that interpersonal argumentation is based on an agent's self-knowledge of her commitments. And while I have often focused on commissive beliefs, I do believe that the argument extends to *any* attitudes about which one might have interpersonal disputes—one's commissive desires and intentions that one may undertake to defend or forfeit in light of objections. In order to interpersonally argue with others or justify one's commitments to others, one must exercise a social-epistemic self-locating capacity: one must have a sense of one's discursive position relative to one's possible or actual interlocutors.[167] And this is why I reject Coliva's blind drilling story: commissive self-knowledge is on the scene earlier than she contends. Indeed, it is on the scene whether or not we add that interpersonal reasoners are "rationally responsible" for their attitudes. However, as aforementioned, I have yet to explain why commissive knowledge must be privileged and peculiar. In §4.6.3, after discussing another social-cognitive context in which self-knowledge seems to figure necessarily, I will address the question of what makes commissive self-knowledge privileged. I will then argue for its peculiarity in §4.6.4.

§4.6.3—Social-Epistemic Self-Locating in Collaborative Group Action

I will now argue that self-knowledge is also necessary for certain forms of more overtly *collaborative* rather than *discursive* or *argumentative* social actions.[168] Consider two scenarios:

[167] My argument in this subsection might explain why at least some reasoners—those capable of interpersonal reasoning—must take their premises to support their conclusions in a "robust" way, i.e., in a way that involves self-knowledge of one's premise- and conclusion-*attitudes* rather than just knowledge of their contents (see §3.4). Alternatively, it may even explain why *all* reasoners are self-knowers, if it turns out to be true that our capacities for non-interpersonal (i.e., personal) reasoning is ontogenetically and phylogenetically rooted in our acquisition of interpersonal reasoning capacities (see Mercier 2009 for the ontogeny; see Sperber & Mercier 2011, 2012, 2017 and Smith & Wald 2019 for the phylogeny). Either way, it is not the Taking Condition itself that yields this result, since I have already argued that arguments for the Taking Condition itself do not motivate the robust view over the second-tier view.

[168] I understand that the collaboration/discourse contrast is somewhat artificial, since interpersonal argumentation can be a collaborative project in the pursuit of collective knowledge.

(1) Tariq attends a protest out of anger at an injustice committed by his local government. He finds himself to be one among many at the protest site. Several police officers approach him and tell him to go home. Tariq knows that he is not alone in his intention to protest, and stands his ground.

(2) Grace believes that rising tuition fees on campus are a dire issue for today's youth. She forms a campus coalition dedicated to addressing this issue. Others join the coalition. Because she knows she shares core beliefs and desires with members of her coalition, she can delegate different roles to different members. At a meeting, Grace and her colleagues confront a complex organizational challenge, one that cannot be met merely by Grace's φ-ing. However, Grace knows that, while she intends to φ, her colleague intends to Ψ, and she reasons that her coalition's both φ-ing and Ψ-ing will suffice to overcome the challenge. Knowing this, she φs.

In both of these cases we have agents whose actions within a group structure are rationalized, in part, by their awareness that they share in or diverge from the attitudes of others.

Take Tariq first. I grant that it may be possible for a protest to form simply because Tariq and several agents happen to believe in the same cause and, likewise, happen to converge on a shared physical location to make their voices heard (however unlikely this might be). However, the possible courses of practical reasoning available to Tariq concerning how to interact with the police officers would be seriously narrowed if he did not know that his attitude was shared among others in the vicinity; rather, he would feel overwhelmed by the presence of opposing law-enforcement and would consider himself incapable of single-handedly standing up to his potential oppressor. Here, his recognition of solidarity is key to rationalizing his action, and this recognition is itself a function of taking a perspective on his and others' mental states. Similarly, Grace would have no reason to think that the organizational challenge could be conquered if she did not know that different members of the organization intended to perform different actions that jointly address the challenge. Notably, both Tariq and Grace act on the basis of their

commitments; their attitudes are responsive to the reasons they have to act, and they would not act as they do if they did not judge that other agents were in solidarity with them.[169]

My claim, in all of this, is not to suggest that all group action requires self-knowledge. I will be content if some actions are not rationally performable by agents, and will not strike agents as rational to perform, in the absence of knowledge of one's social-epistemic position. Cognizance of group membership is, in these scenarios (*inter alia*), cognizance of the fact that one shares or does not share one's commitments with others. Insofar as many intentional actions are the upshots of practically reasoning from one's commitments as premises, or even from reasoning about facts *about* one's commitments, a lack of self-knowledge can actively prevent agents from conceiving of action-possibilities, and can thus prevent agents from intentionally acting at all. But these are, I think, ubiquitous cases of group action. So ubiquitous, I suspect, that an absence of self-knowledge would drastically diminish our social and collaborative agency, perhaps to a point where it would no longer be possible to make sense of the possibility of large-scale (e.g.,) international institutions (and their actions) at all.

Now, just as with the case of interpersonal argumentation, Tariq and Grace are described as taking both a self- and other-perspective. So the question is sure to strike one yet again: why must it be that only the *self*-knowledge Tariq and Grace possess is privileged and peculiar? After all, if both self-knowledge and other-knowledge are indispensable to their actions, then it is not clear that we have revealed any crucial differences between the kinds of knowledge at issue.

One way to meet this challenge is to think along Colivan lines. Whereas Tariq and Grace must acquire other-knowledge by testimony, observation, or inference, their own commitments

[169] There is a connection, in all this, to my discursive account from §4.6.2. For, if Pettit (2007) is right, then group agency is sometimes necessarily brought about by interpersonal reasoning (and group rationality maintained by it), and so self-knowledge will be on the scene in the very formation of the group.

are simply known to them in virtue of their rational responsibility for them (hence the peculiarity of self-knowledge), and this knowledge is privileged because such easy (conceptually necessitated) knowledge is bound to be especially reliable.

However, Social Agentialism is not dependent on this Colivan line. Take privileged self-knowledge first. Here, I derive inspiration from Charles Siewert (who in turn draws inspiration from Shoemaker 1996a):

> ...social animals that we are, by and large we need cooperation and assistance from others if we are to get what we want. And, because we are also *rational* animals, we do things we take ourselves to have reasons to do; we act in a manner which we would justify in certain ways. Now if the reasons we would offer did not have us acting in ways revealing our actual beliefs and desires [and intentions, I add] to others, we would be much less effective in securing others' cooperation and assistance than we in fact are, in everyday contexts. So, we could not reason about our desires to get them met nearly as well as we actually do, with others' aid and without their hindrance, if we did not represent our thoughts to ourselves and others with reliable accuracy… (2003, p. 139)[170]

Siewert's account is in many ways congenial to my Social Agentialist account, focusing as it does on the interrelations between social agency and self-knowledge. The basic idea, which I share with Siewert, is that cooperation requires that we be especially good at knowing ourselves, even if we also need to know others. I take this to be true of how we are to understand Tariq and Grace, as well as how we are to understand Maya (see §4.6.2). However, there are some key differences between my view and Siewert's.

First, Siewert does not draw a distinction between commissive and dispositional self-knowledge, though I believe he should (for, as I have suggested, it is our commitments that figure chiefly into our cooperative social-epistemic self-locating activities, and so it is only these

[170] This is Siewert's take on a suggestion from Shoemaker (1996a, pp. 27-28). I take the basic idea in this passage to be commensurate with Frith's (2012) view that metacognition enhances joint action.

attitudes we are entitled to think of as being known with privilege, given the sort of argument Siewert advances).

Second, and more importantly, I believe that Siewert only accounts for part of what makes privileged self-knowledge essential to ensuring the levels of collaboration that are characteristic of group action. For, on his view, such knowledge is chiefly to be exercised in an agent's acting in ways that accord with what she says about herself, so that she can be taken as a consistent and therefore reliable collaborator. But I believe that such knowledge would be just as essential for agents who in certain cases do *not* end up engaging in group actions with others. These are cases where an agent (1) knows her own mind, and (2) undertakes to discover who else shares her state of mind, but (3) discovers that nobody does. She might find that nobody in some group is suitably like-minded and hence decline to interact with the group at all. If this agent did not have privileged self-knowledge, her capacity to *consider who* she may want to offer reasons for group action to would be drastically impaired, whether or not she *then* goes on to act "in ways revealing [her] actual beliefs and desires" to those people. This knowledge must be at least as privileged at this stage of the process as it is at the stage at which she goes on to reveal her attitudes to others, for it is equally part of the process of setting out to develop structures of group agency.

The idea that privileged self-knowledge is a key ingredient in enabling human-level cooperation in group action is something that I take to also extend to the cooperative elements of interpersonal reasoning, even when this is argumentative in nature and hence in one sense combative. Interpersonal reasoning, combative or otherwise, still requires collaboration on the part of its participants, insofar as it is a group effort to (e.g.) arrive at truths or to persuade one another into adopting more fitting attitudes (toward whatever ends).

Differences aside, notice that Siewert's argument, like my arguments in this subsection and the previous one, make no use of a putative conceptual connection between rational responsibility and self-knowledge. Rational responsibility is once again beside the point here, even if it turns out to be true that any agents who are competent cooperators and interlocutors are also rationally responsible for their mental states.[171]

§4.6.4—Social-Epistemic Self-Locating in Linguistic Interpretation

Can Social Agentialists say anything about why commissive self-knowledge is not only privileged but also *peculiar*? I believe they can, though the argument requires us to identify yet another social-epistemic self-locating capacity of ours, namely, our capacity to interpret one another's thought and speech. As an initial source of inspiration, I want to draw on two passages from Donald Davidson. Here is the first passage:

> …the standards of rationality and reality on which I depend in understanding others are my own, and there can be no appeal beyond them…insofar as I seek information directly by experiment and observation, I again can do no better or more than employ my own resources. And if I wonder whether the norms of rationality I employ in trying to comprehend others are correct, I can, of course, ask Sebastian whether I am as objective or reasonable as I should be in my account of Basil's thoughts and actions. But my understanding of Sebastian's reply will be one more exercise of my own standards and methods. There is another obvious indication of the irreducible singularity of my direct acquaintance with the contents of my own mind, and this is that such knowledge is unique in that it is, aside from unusual cases, unsupported by observation, evidence, or reasons. This is due, at least in part, to the fact that here interpretation has no application. (2001a, pp. 90-91; originally Davidson 1998)

And the second passage:

> …we are not in a position to attribute thoughts to others unless we have our own thoughts, and know what they are, for the attribution of thought to others is a matter of matching the verbal and other behaviour of others to our own propositions or meaningful sentences. (1991, p. 160)

[171] Is my view compatible with an error theory about rational responsibility, given that I did not deny Coliva's *normative* conception of commitments? While I do not care to defend an error theory here, my point is that we can potentially temper our sense of how deeply, intrinsically normative commitments are (at least if it can also be shown that the *peculiarity* of commissive self-knowledge does not depend on the normativity of commitments).

In both passages we encounter something like the following idea: self-knowledge of some sort is necessary to the process of interpretation. Moreover, in the first passage we find the additional idea that the self-knowledge interpreters utilize in this process cannot itself be based on interpretation in the ordinary run of cases. The view I want to advance from these passages, then, is that we have peculiar self-knowledge of our commitments insofar as we are interpreters.

But do these passages really suggest such a view? Notice that the first passage refers to self-knowledge of "standards of rationality" and "mental contents", whereas the second passage refers to self-knowledge of our own "thoughts" and "meaningful sentences". With the possible exception of mental contents, none of these terms seem to refer to our commitments, let alone our attitudes of any kind. After all, I can entertain the thought that the sky is blue without my thought's constituting the content of an attitude,[172] and so too can a sentence take a proposition as its object without that proposition's being the content of an commitment. Likewise, norms of rationality can place *constraints* on my commitments without being identical to them, though I may also *believe in* these norms, such that they are the *contents* of some of my commitments. How, then, can the above passages generate an account of self-knowledge of one's commitments specifically, peculiar or otherwise? In response to this worry I will argue that, although Davidson sometimes spoke narrowly of self-knowledge of norms, thoughts, contents, and sentences, a proper understanding of the above passages and of the interpretive process suggests that interpretation presupposes self-knowledge of one's commitments, and that this holds the key to understanding why such self-knowledge must also be peculiar.

[172] At least so long as we view thoughts as commitment-neutral 'entertainings' of contents. Mandelbaum (2014) rejects this view of thoughts, though see Street & Richardson (2015) and Street & Kingstone (2017) for counter-evidence.

Here is a preliminary sketch of my account, which takes a certain reading of the second passage as key to illuminating the first. As I read him, Davidson's talk of "matching" in the second passage suggests that interpretation is an essentially metacognitive process. When agents "match" up their commitments with those of other agents (or otherwise *contrast* them, as the case can surely be), this is a metacognitive achievement; it requires taking a perspective on one's own mind *relative to the interpretee's*. Not only is it metacognitive, it is metarepresentational: it does not reduce to a higher-order cognitive *ability*. Now, because interpretation is a metacognitive process in this sense, it follows that interpreters must have non-interpretive access to at least some of their own commitments. The reason is that, when we interpret ourselves instead of others, the basic metarepresentational structure of interpretation does not change: the interpreter must still match up her own mind with the interpretee's by forming thoughts about correspondence or contrasts between mental states, even when the interpreter and interpretee are the same person. But because of this very fact, interpretation cannot so much as get off the ground unless one side of the comparative relata is already grasped: there can be no start to the matching process if neither side of the relata is within the subject's cognitive ken. This is why the second passage illuminates the first: it explains why interpretation has no application in the self-directed case, at least in "the ordinary run of cases". The qualifier can be allowed so long as *when* self-interpretation is called for in order to acquire self-knowledge (as with our dispositional attitudes, oftentimes), its possibility is parasitic on the possession of some *other* more peculiar (here, commissive) form of self-knowledge.

The foregoing does not get us all the way to the claim that peculiar *attitudinal* and *commissive* self-knowledge plays a key role in interpretation. To cross the residual distance, it will help to have a case in view. Suppose, then, that I am working on a project

with Basil. He utters, “Some help on our project would be great right now”. To interpret him, I need to secure knowledge of a number of things: not only what his sentence literally means, but whether it expresses an intention, desire, or whatever. If I am to get a start on this, I must assume that Basil is basically rational: I must assume that he communicates his thoughts in more or less appropriate circumstances, more or less consistently, and that he is more or less rationally responsive to our shared environment. Otherwise, I won’t be able to take facts about Basil’s environment as reliable clues to understanding what his utterances are about. And if I cannot do that, I start out with too impoverished a set of resources to commence the interpretive process.

Underwriting my attribution of basic rationality to Basil are my own norms of rationality, as Davidson says. These are norms that call, e.g., for coherence in what we say and do. This, I take it, means that I must have some knowledge of what these norms are, so that I can take Basil as sharing them. But this is not enough. For one thing, I still need to get a sense for the particular empirical content of his utterance, and this is not something that norms of rationality can tell me all on their own. After all, there may be many different interpretations of Basil’s mental state or the meaning of his utterance that are compatible with taking him as satisfying basic rational norms. But even this is not enough for me to successfully interpret the literal meaning of his utterance by fixing its empirical content. For it may be that his utterance, with its particular content, expresses a *desire* or *hope* or *belief*, and this means that my understanding of what his utterance means underdetermines my understanding of the mental state it expresses. Something more is needed here.

The missing tool, I contend, is self-knowledge of my own commitments. But why think this? After all, it might be thought that I only need to know what commitments I, as a

rational agent, *would have* if I were in Basil's situation. My reply is that this is not so: I must also know my actual commitments toward features of our shared environment and situation. Thus, in the example at hand, interpreting Basil's utterance of "some help on our project would be great right now", I will proceed not only on the basis of my considering what commitments it would be rational for someone in Basil's position to express by way of that utterance. Rather, in addition, I must interpret him on the basis of my current understanding of what attitudes are rationally expressible about our current and *shared* situation. For instance, it might be that I can infer that a desire to get help on our project is rational to have. In understanding this, I will form this desire (as a commitment) if I do not already have it, and I will do so in a self-knowing way, since I am presently invested in using my awareness of my attitudes as a guide to thinking about Basil's.

If this is the correct picture, we cannot always have self-knowledge by self-interpretation, because in order to acquire self-knowledge by self-interpretation we will have to draw on commissive self-knowledge anyway. This follows, once again, by reconstruing the situation so that the interpreter and the interpretee are the same person: at some point, the interpretation will depend on antecedent self-knowledge of some of one's commitments, even if what we are trying to self-interpret is some *other* commitment of ours. Accordingly, commissive self-knowledge is not typically based on inference or observation, these being the basic methods by which we interpret people.

Of course, charitable interpretation can sometimes mandate that we interpret people as merely having dispositional attitudes toward *p*/φ-ing. For this reason, it might seem that we can interpret others in these cases only if we have some grip on our dispositional attitudes, given the foregoing argument. And so it might be wondered anew why I have

been claiming, throughout this subsection, that Davidson provides us with insight into the peculiarity of *commissive* self-knowledge specifically. The question suggests an objection: it suggests that dispositional and commissive self-knowledge might not differ in point of their peculiarity, so long as both are necessary prerequisites for interpreting others.

My reply is that dispositional self-knowledge is not indispensable to all interpretation, whereas commissive self-knowledge is, and this is why we can get away without possessing peculiar dispositional self-knowledge. The reason is that interpretation begins and *ideally* ends in attributing commitments to the interpretee, given the principle of charity. We honor the principle of charity when we interpret others as rational (by their own lights).[173] In the ideal case, then, we only exploit self-knowledge of commitments in interpreting others. Indeed, if we were to begin the interpretive process by attributing dispositional attitudes to agents, we would not get very far. This is because dispositional attitudes are not necessarily shaped by rational pressures, pressures that we can infer from the fact that we share an environment with the interpretee, and so we would have no constraints on *how* and *what* dispositional attitudes we attributed to them if we did not first attribute commitments to them. We would be free to attribute any such attitude, which is tantamount to being so unconstrained in one's endeavour as to make it impossible to begin. This, I take it, is why Davidson's principle of charity is so important: we assume that agents are rational and so try to attribute commitments to them, commitments that they would sensibly come to have on the basis of their rational engagement with our shared environment or situation. Because we are also rational and know something about the kinds of attitudes we have, we have some sense of what the interpretee's mind could be

[173] At least, given one reading of the principle. Another reading suggests that we must attribute *by and large true beliefs* to the interpretee, while yet another suggests that we must ensure widespread *agreement* with the interpretee.

like. We interpret the interpretee as having dispositional attitudes only once we have exhausted our efforts to follow the principle of charity.

To summarize, my point is that it is not necessary to interpretation *in general* that we reach the stage of attributing dispositional attitudes to other agents, even if it is often true (perhaps even always true, as a contingent matter of fact) that we will reach this stage. Moreover, even if I do attribute dispositional attitudes to you, I may do so only by making recourse to my *commissive* self-knowledge, plus my knowledge of the causal forces at work in our shared environment, since there is no reason in general for me to expect Basil and I to share the same dispositional attitudes.[174] So it is not necessary to interpretation that we take advantage of our dispositional self-knowledge, whereas it is necessary that we utilize commissive self-knowledge.[175] This is why I conclude that only commissive self-knowledge need be seen as peculiar.

Here, then, is a summary of my Social Agentialist account. Commissive self-knowledge underwrites our social identities. We utilize self-knowledge of our commitments in order to participate in basic social-epistemic practices.[176] It is commissive mentality that we must possess and put to use in the course of discursive, collaborative, and interpretive social-epistemic

[174] There is no reason for this precisely because these are not the sorts of attitudes that I could expect us to mutually form on the basis of our shared *judgements* about our shared environment.
[175] One could also argue, as Bilgrami (2006b) has, that Davidson would not really ever have embraced the idea of commissive attitudes. *Pace* Bilgrami, I agree with McDowell that there is "nothing in Davidson's basic outlook that should induce him to resist the idea" of such attitudes (2006, p. 69). It is another question whether he would ever have accepted the existence of dispositional attitudes. I believe a case can be made for this interpretation, albeit controversially. Here is just one telling passage for now. The passage is made in response to a suggestion from Stephen Stich: "[t]he…suggestion was that the mental states needed for a scientific psychology, though roughly propositional in character, bear no direct relation to common beliefs, desires, and intentions. These states are, in effect, stipulated to be those that explain behavior, and they are therefore inner or subjective only in the sense of being characterized exclusively by what is physically beneath the skin. But there is no reason to suppose that people can tell without observation when they are in such states and so no reason to call them subjective." (2001a, p. 51)
[176] Perhaps there is interesting work to be done in the future on the question of whether *social groups themselves* require privileged and peculiar self-knowledge. Schwengerer (2020) has made a start in thinking about whether groups themselves *can* have such knowledge, but the former question has not been addressed as far as I can tell.

activities. Commissive self-knowledge, being necessary for all of this, is also peculiar and privileged. It is privileged because the possibility of so much group action that we can participate in depends on our capacity to know our own attitudes with a high degree of reliability, and it is peculiar because it cannot be acquired on the basis of self-interpretation.

§4.6.5—Social Agentialism and Chrisman's Concern

In §2.7.8 I suggested that Neo-Expressivists about authority were left with two questions. One of them was: if the privileged and peculiar self-knowledge agents have of their mental states does not explain the authority of their avowals, what *does* privileged and peculiar self-knowledge explain? I hope to have offered a partial answer to this question by now: privileged and peculiar self-knowledge of *commitments* is crucial to explaining crucial social-epistemic competencies. However, this question was not the question that I referred to as *Chrisman's concern*. I phrased that concern, due to Matthew Chrisman (2009), as follows: how do privileged and peculiar self-knowledge and authority relate, if not by one explaining the other? The concern, in other words, is that it would be strange if, having argued that we are in fact privileged and peculiar self-knowers, there is no deep relationship between this knowledge and the authority of our avowals.

Some Neo-Expressivists could simply deny that we are privileged and peculiar self-knowers, such that there is no need to capture any relationship between the authority of our avowals and such self-knowledge. But for someone in my position, this cannot be done. For, as a Social Agentialist, I contend that we have privileged and peculiar self-knowledge of at least some of our mental states, namely, our commitments. What I want to do here, then, is argue that Chrisman's concern can be addressed by showing why *both* authority (explained along Neo-Expresssivist lines) *and* privileged and peculiar self-knowledge (explained along Social Agentialist lines) come together in crucial cognitive contexts. On this strategy, neither

phenomenon stands to the other as explanans, but neither are we left wondering about the nature of their relationship.

I begin by reiterating a claim that I advanced in §4.6.2: agents engaged in discursive social-epistemic activities like interpersonal argumentation must have a sense for their relative positions in social-epistemic space, and this is a matter of appreciating how one's mental states diverge from or accord with one's interlocutors' mental states. As we eventually saw, such knowledge is also privileged and peculiar. I now want to suggest that, just as privileged and peculiar self-knowledge is indispensable to this process, so too is the capacity, shared by each interlocutor, to directly express their commitments by avowing them, just as the Neo-Expressivist contends.

My argument brings us back to a point that I made about commitments in §4.4.3, namely, that commitments necessarily involve at least one first-order disposition: a disposition to draw inferences from them in at least some cognitive contexts. My argument was that an agent who is not disposed to infer anything whatsoever from a proposition is not really committed to it, for any such attitude will not exhibit the responsiveness to reason that is necessary for having a commitment. In §4.6.2 I added that if one believes that one believes *p* or desires/intends to φ, and if one's self-belief *has the formal character of endorsing* one's commitment, then one will be disposed to draw first-order inferences that are constitutive of having the commitment about which one has this self-belief.

Now consider an interpersonal argument in which A believes that *p* and B believes that ~*p*. Both agents offer reasons for their respective beliefs as the argument proceeds. Moreover, both agents recognize that their interlocutor disagrees over *p* (and over any number of reasons offered for or against *p*). This, as I argued in §4.6.2, means that both agents have knowledge of their

discursive positions in the exchange. Now the first claim I want to make is this: if A opts to avow her commitment to *p* to B, A will be expressing both her self-belief *and* her first-order commitment itself in the avowing act. This is because the very same disposition that (at least partly) constitutes her self-belief is also a part of her first-order commitment.

But does this also vindicate the claim that A and B (should) *treat each other as authoritative* in virtue of the *first-order* expressive function of their avowals? Or is it enough that they take one another as having privileged and peculiar *self-knowledge* of their commitments? I believe it does matter that their avowals (are taken to) express their commitments themselves. Here is one reason: merely appreciating the security of a speaker's self-knowledge is not sufficient to take them as *standing behind* what it is that they have self-knowledge of, and one must take one's interlocutor as really standing behind what she says if one is to have reason to engage—discursively, at least—with one's interlocutor's commitment itself. The reason is that I can have self-knowledge of many mental states that I do not endorse, and so if I self-attribute them it is not necessarily appropriate for you to discursively engage with the mental states I have self-attributed. Sometimes, I merely self-ascribe an attitude in the spirit of reporting its presence, *not* in the spirit of putting it forward in a debate. However, if one's avowal expresses one's commitment itself, the interlocutors in a dispute can recognize one another as *speaking from* that state and, hence, as standing behind what they say.

But perhaps the above reply will not suffice, since A and B might instead recognize one another's self-beliefs as having the endorsing form which I spoke about above. If so, the question arises: might A and B take one another's avowals as authoritative insofar as they take one another's avowals to express self-beliefs *that have the form of endorsing their commitments*, even if they don't take one another's avowals as expressing those commitments themselves? The

answer, I believe, is still no, because taking an agent's avowal to express a self-belief that endorses its first-order object is just another way of taking it to express its first-order object. The reason is as before: when we take an avowal to be authoritative, even granting that we take it as expressing a self-belief that endorses the avower's commitment, this is a way of taking the avower to *self-consciously adopt the perspective of her first-order commitment itself*. This is what it is, on my view, for the self-belief to have the relevant form. As a consequence, when an agent avows her commitment, she expresses her commitment itself *in this self-knowing, endorsing way*, rather than merely expressing her self-belief.

One might now ask: what if the agent is only able to avow her commitment because she self-consciously adopts its perspective? If this is the picture, isn't her being a privileged and peculiar self-knower essential to her being able to authoritatively avow? One thing to say is that this may be so, but that this is no objection to Neo-Expressivism. This is because I have no objection to treating privileged and peculiar self-knowledge as part of the aetiology of avowing (§2.7.6). What matters most is not that privileged self-knowledge is or is not (recognized as) part of the aetiology of avowing (at least so long as this knowledge is not the upshot of *detecting* one's mental states), but that what ultimately grounds the presumptive truth and relative indubitability of avowals is their (being taken as) expressing the commitments self-known. When we take agents to avow their commitments, we defer to their avowals because we understand that they are expressing their commitments. If we also admit that the avower could not do this without possessing privileged and peculiar self-knowledge, this does not mean that our acknowledgement of this knowledge on the avower's part is what grounds her authority. Again, in §2.7.6 I offered a few tentative reasons to avoid thinking in this direction. The basic idea is that, if A takes B to have this knowledge, this is not because B manages to *express* the privilege

and peculiarity with which she knows her own mind, because this is not plausibly something that B expresses and that A is attuned to perceive. Rather, if A takes B to have this knowledge, it is because A makes some sort of assumption to this effect. But this assumption, even if A can and should sometimes make it, will not explain why it is that A makes it only when faced with genuine avowals rather than third-personal self-ascriptions. And because our deferential practices are indeed nuanced in this way (possible mistakes notwithstanding), we have reason to look elsewhere for an explanation of A's deference to B. Neo-Expressivism provides this explanation.

I conclude, therefore, that A should take B's avowals of commitments as authoritative inasmuch as B is in the best position to express her commitments through her avowals. This is true even if we also take her to have privileged and peculiar self-knowledge of her commitments, and even though she must have this knowledge in order to reason discursively with us. Having privileged and peculiar self-knowledge of our commitments as well as being able to express our commitments themselves through our avowals, are *joint* elements of our discursive competencies, and so it is no mystery that they come together even if one does not explain the other. For this reason, my Social Agentialist account of privileged and peculiar commissive self-knowledge is perfectly compatible with a Neo-Expressivist account of authority, at least as regards avowals of commitments. This is only a partial response to Chrisman's concern, seeing as his concern applies to the relationship between Neo-Expressivism and realism about privileged and peculiar self-knowledge for *all* kinds of mental states (toward which we can have such knowledge or authority). But it is a response all the same.

§4.7.1—McGeer and Pettit: Self-Knowledge and Self-Regulation

I have spent the bulk of this chapter developing and defending my Social Agentialist account. In the next few subsections I want to consider one more agentialist account for which I have a good deal of sympathy. In §4.7.4, I will show how this account harmonizes with Social Agentialism.

I will begin by explicating Victoria McGeer and Philip Pettit's account of what they call "self-regulation" (2002). Like Burge, as well as many proponents of the Taking Condition on inference, McGeer and Pettit are at pains to offer an account of higher-order cognition that does not construe our minds as purely "routinised" (2002, 282), i.e., minds where attitude formation and maintenance take place automatically and without cognitive/doxastic agency of any kind. However, whereas Burge takes it that we transcend the automaticity of first-order cognition by way of critical reasoning, McGeer and Pettit's concept of self-regulation is broader than—encompasses more cognitive processes than—critical reasoning. Thus, whereas Burge argues that critical reasoners can influence their attitudes by evaluating them after they have already been generated through first-order reasoning, McGeer and Pettit draw our attention to our capacity to influence the very process of first-order reasoning. One might self-regulate, for example, by rehearsing "the inference that comes naturally to me, studying each step in the transition and guarding against malfunction." Alternatively, one might examine "the habit of inference that takes me spontaneously to the conclusion by seeing where it would lead in other, parallel cases" (2002, p. 289). One self-regulates, in both cases, by influencing where or how well one's first-order reasoning goes.

McGeer and Pettit offer more examples of self-regulation. They consider, first, a gambler who is vulnerable to the gambler's fallacy. Such an agent can self-regulate, the authors argue, by intentionally keeping the gambler's fallacy in mind while at the gambling table, so as to avoid

"backsliding" into its clutches (2002, p. 289). It is because the gambler is *committed* to the falsity of the gambler's fallacy, and because he knows that he is nevertheless *disposed* to act on it, that he does this. A second example involves a pilot who is trained, as pilots are, to ignore proprioceptive information about the orientation of their aircrafts, and to focus instead on the readings provided by the instruments in the cockpit. When pilots do this, they self-regulate against influences from misleading proprioceptive data. It is because they know that the beliefs formed by attending to their instruments are epistemically superior to those formed by proprioception that they do this. Both the gambler and the pilot "succeed in being governed by reason, and in forming their beliefs after an appropriate pattern, only by grace of an intentional effort to school themselves into going by the book" (2002, p. 290).

The above examples involve agents who attend to their internal cognitive processes. Other examples show how agents can proactively manipulate their external environments in order to ensure that their cognitive processes or states function in certain ways. One can self-regulate, for example, by going to the library to acquire evidence for the truth or falsity of some proposition, whether or not one also attends to one's internal cognitive processes once one has acquired the evidence. McGeer (2007, pp. 893-96) also provides an interesting example of using other agents as a means to self-regulation. She takes a case from George Eliot's (Mary Ann Evans's) novel *Middlemarch*. In the novel, Reverend Farebrother finds himself struggling with his love for a woman named Mary. His love produces a struggle because it conflicts with his sworn celibacy as a religious authority. Knowing this about himself, he confesses his love for Mary to another man, Fred. He does this because he knows that Fred also loves Mary, and that Fred's hearing Farebrother's confession will encourage Fred to redouble his efforts in securing Mary's affection. The end result, as intended by Farebrother, is that Fred's securing Mary's love will

make it impossible for Farebrother to secure it. In this way, Farebrother uses Fred to protect himself from giving into his baser desire.

§4.7.2—Self-Regulation as Future-Directed Acts of Self-Control

Tillman Vierkant wonders whether "we really need to be able to ascribe folk psychological states to ourselves in order to intentionally manipulate [i.e., self-regulate] our mental lives" (2013, 282). One skepticism-inducing example comes from research on chimps where it was discovered that, when faced with a choice between minor instant gratification or delayed superior gratification, they could distract themselves by playing with toys while holding out for the delayed superior gratification (Evans & Beran 2007). As Vierkant notes, this looks very much like the chimps used intentional behaviour in order to prevent themselves from caving in to the benefit of the more quickly available, smaller reward. He notices that this case "looks remarkably similar to the pilot case described by Pettit/McGeer" (2012c, p. 282), since both the chimp and the pilot prevent their attitudes from succumbing to certain pressures by paying attention to something else. But Vierkant does not conclude that the chimps have self-knowledge: rather, he concludes that they must self-regulate despite lacking self-knowledge. By parity of reasoning, Vierkant suggests that we can be skeptical about the need for self-knowledge in the piloting case.[177] Put generally: "it is quite possible to manipulate mental states by acting intentionally, even if you don't know what mental states are, as long as the manipulation of mental states is not what you intend to achieve but is connected to (or is, for example, the cause of) you successfully reaching a first order goal" (2012c, p. 282).

[177] Similarly, consider "…the turning of its head by a nervous mouse in order to see whether the buzzard it was watching fearfully earlier on is still on the tree at a safe distance where it had been before. Turning its head clearly allows the mouse to…regulate its beliefs, but it would seem ludicrous to suggest that it does so because it is aware that its belief about the location of the predator is by now possibly false" (2012c, p. 282).

Nevertheless, Vierkant thinks that there is a form of self-regulation to which self-knowledge is indispensable. He argues that:

> …realizing that our first-order evaluations may change enables us to have a completely new level of self-control. As long as the agent does not have psychological knowledge, it will be nigh impossible for her to conceive that something which she very strongly believes to be true now could be judged by her to be false by tomorrow. As she can sincerely see no evidence that would render the proposition in question false, it becomes very difficult, if not impossible, for her to comprehend that she might nevertheless judge very differently tomorrow. (2012c, p. 286)

We can summarize Vierkant's position as follows. To be first-order rational is to conform your attitudes to your current reasons, not those that you might have in the future. For this reason, being rational does not, all on its own, confer a perspective from which one can recognize how one's future attitudes, determined as they currently are by one's current reasons, can later be influenced by other psychological forces. However, once we understand ourselves as psychological creatures and understand our attitudes *qua* attitudes, we can use this information to influence our environment in the interest of insulating our attitudes—indeed, our commitments—from future pressures. Thus, Vierkant's position is that self-regulation necessarily requires self-knowledge whenever self-regulation is geared toward "future-directed acts of self-control" (2013, p. 289).[178]

Now, while Vierkant denies that some forms of self-regulation require self-knowledge, he acknowledges that McGeer's example of Reverend Farebrother does require it. The Reverend's confession to Fred of his love for Mary is undertaken in knowledge of the fact that his love for Mary, liable as it is to sway him from his religious duties, needs to be chained down against the threat of a possible future in which it overwhelms his present piety. In effect, the Reverend's awareness of the possibility of this troubling future leads him to manipulate his current

[178] See Spitzley (2009, p. 86) for a similar account.

environment—here, a *social* environment including Fred, an agent in his own right—to make it impossible to follow through on his baser urge. Unlike the other self-regulatory actions discussed (slowing down one's reasoning, checking one's proprioceptive impulses against one's theoretical knowledge) that seem designed to improve one's occurrent rational activities, the point of future-directed acts of self-control is not to ensure the rationality of one's occurrent first-order cognitive processes or attitudes, but to maintain one's attitudes in the face of possible futures where outside influences may distort them against one's current sense of how they should be. As long as these sorts of cases plausibly involve self-knowledge, then the self-regulation view has legs as an agentialist account of self-knowledge.[179]

§4.7.3—Self-Regulation and Self-Knowledge

McGeer and Pettit also take their discussion of self-regulation to bear on the "recurrent puzzle" (2002, p. 293) of *how* people can know their own beliefs and other commitments in a peculiar way. They therefore take themselves to have a *substantive* agentialist story as well (see §3.1.1).

Take the case of belief for demonstrative purposes. On their view, our capacity for self-regulation is directly relevant to this puzzle because, in virtue of being self-regulators, "self-knowledge of one's beliefs comes about, not as a result of having a special insight into whether one has the dispositions that make one a believer that *p* or that *q*, but as a result of having a special ability to develop the dispositions that make one a believer that *p* or that *q*" (2002, p. 293). They clarify that this is not "the ability to make oneself, depending on what one's will dictates, a believer that *p* or that *q*." Rather, "it is the ability to make oneself, depending on what one's judgment dictates, a believer that *p* or that *q*. It is the ability to believe along the pattern

[179] This will mean narrowing the scope of the importance of self-knowing self-regulation in one sense, though future-directed acts of self-control may still be ubiquitous and highly important.

that judgment sets out." In other words, when an agent avows a belief, her capacity to self-regulate provides an assurance that she can shape her attitudes in accordance with her avowal. To my mind, we should conceive of the sorts of actions that would make one a believer/desirer/intender "along the pattern that judgement sets out" as future-directed acts of self-control, since they involve efforts to ensure that one's avowals now will be true to one's standing attitudes over time.

Now, the idea is not that self-regulation of this sort is intended to ensure that we can form our *dispositional attitudes* by way of self-regulating, in a *Bilgramian/Colivan* sense of 'dispositional attitudes', though perhaps it may sometimes serve this purpose. This is because our dispositions are not rooted in our own judgements, and are not the sorts of attitudes that we are necessarily concerned to preserve (from the first-person point of view). Indeed, McGeer and Pettit also take it that our status as self-regulators conduces to a picture of the human agent as having commissive attitudes, quite apart from whatever dispositional attitudes she also has. At least, they seem to share this picture to some extent: for while they treat commitments as different from purely dispositional states, this seems to be because they (or perhaps just McGeer—see especially her 1996, p. 508) take them to be *prescriptions* like promises rather than *descriptions* about what one's attitudes are. Perhaps contrariwise, Bilgrami, Coliva, and myself are happy to treat commitments as *both* prescriptive and descriptive: as both prescriptions about how one's attitudes ought to be and as standing attitudes that are at least partly constituted by certain sorts of normatively evaluable dispositions.

The above is an account of how we have commissive self-knowledge since it explains how we can ensure that our self-ascriptions are reliable over time.[180] But there is an additional

[180] See also De Bruin et al. (2015).

dimension to their thinking, and to Vierkant's, about self-knowledge: each of these authors embraces what we can think of as an *authorship* theory of self-knowledge. For example, Vierkant writes that "one has special authority about one's own beliefs, because by answering the question [of what to believe], one creates (or at least makes visible) the relevant state" (2013, pp. 284-285).[181] In other words, the authorship view does not commit to a special way of detecting our commitments, and argues instead that our capacity to shape our attitudes—from a reflective point of view—explains the privilege and peculiarity with which we know them. Similarly, McGeer argues that "[a]s no scanning or detecting of first-order states is required or implied, there is no special problem of forming epistemically reliable second-order beliefs about them" (2015, p. 269).[182] This is because, as the very authors of our attitudes, our self-ascriptions bring them about rather than help us discover them.

The idea that our self-knowledge is somehow determined by reflective, deliberative self-authorship might seem distant from the concept of self-regulation. But our authorial and self-regulatory capacities are related. For, by settling the question of whether one believes that *p* or desires to φ (through deliberating about the truth of *p* or the goodness of φ-ing, then self-ascribing the relevant attitude), one will either come to have the relevant attitude automatically in favourable psychological circumstances, or can deploy self-regulative measures to ensure that one comes to have this attitude in less good psychological circumstances. Thus, even when one's self-ascription of an attitude does not automatically serve to create a standing commitment, exercising one's self-regulative powers can greatly increase the chances that one's self-ascription will result in a standing attitude. McGeer's most recent way of putting this idea is that being a

[181] See also Moran (2001, 2003, 2004) and Peacocke (2017).
[182] Thus, second-order reasoning can be a way of acquiring self-knowledge, even if such knowledge is not indispensable to critical reasoning (see §3.2-3.3).

self-regulator consists "in having a capacity…to bring your self-ascribed (or expressed) psychological states into line with your deeds, and your deeds into line with your self-ascribed (or expressed) psychological states" (2015, p. 270).

§4.7.4—Social Agentialism and Self-Regulative Agentialism

Having just exposited 'Self-Regulative Agentialism', understood as both a 'substantive' account of how we have commissive self-knowledge and 'non-substantive' account of its cognitive indispensability,[183] I now want to speak to how Self-Regulative Agentialism relates to Social Agentialism.

The first thing to emphasize is that Self-Regulative Agentialism complements Social Agentialism rather than overtly competing with it. One thing that Self-Regulative Agentialism does nicely, and that is not emphasized by Social Agentialism, is that it better emphasizes the importance of occasionally taking a third-personal perspective on one's commitments. By 'third-personal' I do not mean a self-perspective that one takes in virtue of acquiring third-personal self-knowledge of one's attitudes. Rather, I am referring to the "clinical" perspective we can voluntarily take on our attitudes (no matter how we first came to know them) when we comprehend them as delicate psychological entities that are vulnerable to various pressures from within and without. The self-regulating agent is one who takes up this clinical perspective on her commitments. Farebrother examines the vulnerability of his commitment to celibacy, and he undertakes to stabilize it against his love for Mary. Farebrother teaches us that if one only embodies a first-person commissive perspective, one will be more vulnerable to corrosive psychic interference. Thus, the self-regulator:

[183] As defined in Chapter Three, substantive accounts of self-knowledge tell us *how we acquire it*, whereas non-substantive accounts are transcendental accounts to the effect that we *must* have it.

> …is continually ready to step back from her own character, disempowering the authorizing voice of her own reason to some extent in order to make possible a more objective assessment of her own appetites, needs, weaknesses, and reactive impulses as psychic forces potentially shaping, as much as being shaped by, her own deliberative processes. (McGeer 2008, p. 102)

Nevertheless, there *are* times where a more first-person, non-clinical view on or through our commitments themselves is indispensable, as Social Agentialism argues and as Self-Regulative Agentialism sometimes downplays when considering them in all their fragility and contingency. In debating, in collaborating, and in interpreting, what is at the forefront of our minds is the reasonability of our attitudes and whether these can be transmitted to or joined up with the attitudes of others for many purposes that we are first-personally invested in. No concerns about the fragility of our commitments are salient when we take up this perspective, though it is easy for this point to get lost when we think only about our self-regulatory tendencies. For, if we do not keep in mind the role played by a non-self-regulatory perspective on our commitments, we will begin to lose sight of why they are so worth preserving via self-regulation in the first place.

Now for a point of disagreement. I understand commitments as self-known with privilege and peculiarity not (only) because they can be self-reflectively authored, but because *standing* knowledge of them is indispensable to our diachronically extended, social-epistemic agency. My point is that we require self-knowledge of our commitments long after we have authored them, since after the point of self-authorship we often find ourselves engaging in cognitive activities where self-knowledge of them is required, and where such knowledge is—for reasons argued across §4.6—privileged and peculiar. Thus, while McGeer argues that "no scanning or detecting of first-order states is required or implied" insofar as one brings about one's commitments by self-ascribing them, this does not mean that we do not need a standing form of self-knowledge of

our commitments as well, insofar as such knowledge figures into diachronically extended projects. How we can have *this* self-knowledge is a question for Chapter Five.

§4.8.1—Preamble to Chapter Five

It is possible to argue against me here, i.e., to argue that we are not especially good at accessing our standing attitudes over time. Or one might insist that, in cases where it seems like we are accessing a long-since-settled standing attitude, we are actually just re-making up our minds. However, I am more optimistic about the reality of standing privileged and peculiar commissive self-knowledge. For, first of all, it often seems overly time-consuming and perhaps even indicative of a kind of irrationality or some other cognitive shortcoming if one has to (re)-make up one's mind about something one has believed, desired, or intended for a long time (cf. Van Woudenberg & Kloosterboer 2018, p. 118). For example, if I am asked whether I desire for my friends to succeed in life, having to evaluate the reasons for this desire and make up my mind anew with respect to it would indicate that this seemingly basic and easily-cited desire of mine is actually rather superficial. Second of all, there seem to be cases where we have privileged and peculiar self-knowledge of commitments attitudes that are not open to further deliberation, such as my belief that I am wearing pants (cf. Shoemaker 2003). For these reasons, in Chapter Five I develop an account of self-knowledge that can supplement the self-authorship account.

Chapter Five—A Constitutivist Account of Commissive Self-Knowledge

§5.1.1—Introduction

To offer a 'substantive' account of privileged and peculiar self-knowledge is to account for its provenance (Sorgiovanni 2019). In §4.7.3 I considered the view that self-knowledge of our commitments is privileged and peculiar insofar as the agent herself is able to *determine* her attitudes by avowing them (and, where required, self-regulating one's behaviour in accordance with one's avowals). Therefore, there is no need for the agent to *detect* her mental states in the way that she must detect the mental states of others. But I also suggested that this may be a partial substantive account of privileged and peculiar commissive self-knowledge, because there seem to be cases where one can have privileged and peculiar self-knowledge well after the moment of its self-conscious determination, and in cases where the relevant attitudes are not readily seen as self-authored via deliberation. How are we to account for these cases?

There are a great many substantive accounts of privileged and peculiar self-knowledge that could capture these cases. These vary along several dimensions: how they account for the *warrant* our self-beliefs enjoy, the psychological mechanisms (if any) that provide us with self-knowledge, and the kinds of privileged and peculiar self-knowledge that they aim to explain. Some of these accept the 'detection' metaphor that McGeer rejects (§4.7.3[184]), whereas others do not. Indeed, the sheer diversity of substantive accounts currently on offer has convinced many that we must surely be pluralists about the sources of privileged and peculiar self-knowledge.[185] Fortunately I need not aspire to an exhaustive survey of such accounts. This is precisely because

[184] McGeer is in wider company here—see, e.g., Wright (2001), Coliva (2009, 2012, 2016), and Vierkant (2013). See also Neo-Expressivists like Bar-On (2000, 2004) and Finkelstein (2003) who reject the helpfulness of 'detectivist' accounts of authority.
[185] Moran (2001), Boyle (2009), Coliva (2016), Samoilova (2016), Cholbi (2016), Komorowska-Mach (2019).

I intend to focus only on accounts that could explain our privileged and peculiar *commissive* self-knowledge (without recourse to the self-authorship model, which I accept as a partial account).

Here, then, is the plan for what follows. Over §5.2-5.5 I describe and criticize some historically important and recently developed substantive accounts of privileged and peculiar self-knowledge. This puts us in a position to develop the core of my own account in §5.6. In §5.7 I refine my account by addressing a battery of objections. In §5.8 I conclude this dissertation.

§5.2.1—The Inner Sense Account

The first substantive account that we will consider is usually referred to as the Inner Sense Account.[186] David Armstrong develops the core idea by drawing an analogy between self-knowledge and sense perception:

> By sense-perception we become aware of current physical happenings in our environment and our body. By inner sense we become aware of current happenings in our own mind. (Armstrong 1968, p. 95)

In fact, I briefly discussed this account in connection with Burgean Agentialism (§3.2.2). There, I concluded that *if* Burgean Agentialism is true, then the Inner Sense Account is poorly positioned to explain how we acquire self-knowledge of the sort necessary for critical reasoning. This was because, on Burge's view, critical reasoning is only reasonable if our self-beliefs cannot be warranted yet false. But the Inner Sense Account seems to allow for reasonable yet false self-beliefs. The reason is that, insofar as inner sense is analogous to ordinary sense perception, it should be able to generate brutely erroneous beliefs in the same way that ordinary sense perception can. A brute error, to reiterate, is "one that is simply due to the world failing to cooperate, rather than being due to some kind of failure of the subject's concepts or faculties" (Bar-On 2004, p. 9). In the sense-perception case, this might happen when one is tricked into

[186] Though Coliva (2016, chapter 3) opts for the label "materialist introspectionism".

thinking that there is a barn nearby when, in fact, the barn is a fake. Examples for the inner sense case are harder to imagine, but the thought is that the tight comparison with ordinary sense perception prevents us from arguing that such cases are impossible. And yet it seems that we never take people's self-beliefs to be brutely erroneous when they are first-personally acquired.

We also concluded, however, that Burgean Agentialism suffers from some serious shortcomings. As a result, it may be worth considering the Inner Sense Account once more. Very little was said in §3.2.2 about exactly what inner sense might amount to, and in fact we have every reason to be careful. For, even though some authors are happy to describe inner sense as a kind of inner sense-perception (Armstrong Ibid; Churchland 1984), others have found the analogy with ordinary sense-perception implausible. For, the tighter the analogy between inner and ordinary sense-perception, the harder it is to explain why inner sense is more epistemically secure than ordinary sense-perception (Davidson 1987). And so it is hard to understand why self-knowledge should be privileged, on such a view. I also contend that it is especially hard to make sense of the idea of sensing one's *commitments* via inner perception. This is because it is implausible that commitments have perceptible properties: there seems to be nothing perceptible about one's wanting the world to be a better place, so long as this is understood not as an occurrent urge but as a standing attitude.[187] Similarly, there is nothing perceptible about the commissive belief that the Riemann Hypothesis is true, or an intention to buy Christmas gifts.

Perhaps the idea of inner sense can be divorced from this perceptual analogy. For we might begin instead with the claim that we have some sort of dedicated, *sui generis* causal mechanism for detecting our mental states. As Coliva puts it, "This model gets rid of the Cartesian idea that

[187] Actually, the Neo-Expressivist account of avowals might have it that commitments are perceptible *through avowals*. But even if this is right, it does not help us for present purposes. This is because an agent who has to perceive her own avowal to perceive her own commitment is only in a position to acquire third-personal self-knowledge. Such knowledge is not privileged and peculiar.

self-knowledge is a matter of inner observation: there is simply no observation going on here. There is just a hard-wired mechanism which, given a certain brain state, causally produces another one…" (2016, p. 79).[188]

This provides us with what Sydney Shoemaker refers to as the "core stereotype" of the Inner Sense Account. The core stereotype consists of two features: a causal mechanism dedicated to detecting one's mental states, as well as an immediate implication of this feature, namely, that one's mental states are ontologically independent from the self-beliefs that track them (Shoemaker 1994). Surely, if we have such a mechanism for knowing our own minds, but not for knowing the minds of others, then inner sense yields peculiar self-knowledge. For it to also yield privileged self-knowledge, however, proponents of the account must also argue that this mechanism is especially reliable, such that we are generally in the best position to know our own minds. Now, because what matters is the role played by a distinct causal mechanism dedicated to tracking our mental states, and because the analogy between this mechanism and sense-perception has been abandoned, I will switch from talking about the Inner Sense Account to talking about the Inner Scanner Account.

§5.3.1—The Fallibility of Inner Scanners

Crispin Wright (1998) complains that the Inner Scanner Account is objectionable because it squares poorly with the infallibility of at least some of our avowals.[189] Assuming that we can transpose this concern into talk about self-belief, the concern is something like this: inner scanners, being causal mechanisms designed to detect mental states, are surely not infallible,

[188] The idea is not (necessarily) that mental states reduce to brain states. Rather, the idea is that whenever one is in some mental state, one is in a corresponding brain state, and the relevant causal mechanism tracks these brain states, thereby generating self-beliefs about mental states.
[189] See also Bilgrami (2006a, pp. 91-92), who is more concerned about the *perfect accessibility* or *self-intimating* character of our first-order intentional states.

since there is conceptual space for the possibility of false positives, in just the way that a smoke detector might occasionally go off without actually detecting smoke.[190] But some of our self-beliefs are infallible. So the Inner Scanner Account is objectionable.

However, the avowals or self-beliefs at issue for Wright are phenomenal ones: avowals and self-beliefs about one's pains, itches, and other sensations, as well as some basic emotions like brute fears. But these are not our concern in this chapter, seeing as we are only concerned with commissive self-knowledge, and seeing as these are intentional attitudes rather than essentially phenomenological states. But a second point is even more important, namely, that we need not follow Wright in thinking of *any* self-beliefs as categorically infallible in order to think that there are problems for the Inner Scanner Account. One problem is that this scanning mechanism, being a theoretical postulate, may turn out to be *highly* fallible even if it exists. Proponents of the Inner Scanner Account might stipulate that it is highly reliable, and take as their justification the need to explain prevalent intuitions about privileged and peculiar self-knowledge. However, it is not obvious that Inner Scanner theorists are entitled to *this* stipulation: rather, they may be entitled only to stipulate that there is *likely* to be *some sort* of mechanism or process that yields privileged and peculiar self-knowledge.

Moreover, even if we do not believe that the impossibility of brute error must be closed off if we are to count as Burgean critical reasoners, worries about inner scanners yielding brutely erroneous self-beliefs remain difficult to dislodge. Again, insofar as inner scanners function to scan and detect one's mental states, thereby producing self-beliefs as outputs, it seems perfectly possible that they can yield false outputs in good cognitive conditions, simply in virtue of the brutely causal nature of the mechanism. But this seems true even if we distance the Inner

[190] The example is Byrne's (2018, p. 33). He also notes that Armstrong (1963) himself accepted such possibilities.

Scanner Account from the analogy with inner *perception*. This is because the possibility of brute error seems to arise from the *detectivist* feature of the Inner Scanner Account, not from any subsequently *perceptual* gloss we might place on the idea of an inner scanner. The problem, then, is that the Inner Scanner Account leaves open the possibility that my self-beliefs can be brutely mistaken without any psychological failing on my part—even if my inner scanner is highly reliably, the *kinds* of errors it can yield simply seem to be errors that are not possible for our self-beliefs arrived at in a privileged and peculiar way.

§5.3.2—Inner Scanners and Self-Blindness

I have considered two worries: (1) that inner scanners could fail *often* by generating false positives, and (2) that they could generate false positives that are *brute errors*. But both of these worries may dance around yet another, more fundamental worry, namely, that inner scanners could fail *globally* due to total breakdown, such that they don't produce *any* self-beliefs, let alone false ones. Put differently, the worry is that the Inner Scanner Account is compatible with the possibility of self-blind agents—agents who only ever know their minds by way of third-personal methods like inference, testimony, and observation.

The impossibility of self-blindness was something that we considered first in Chapter Three, though Shoemaker's arguments for this impossibility left much to be desired (§3.6.1). However, self-blindness may still be reasonably viewed as an impossibility. After all, as I argued in Chapter Four, self-knowledge of a privileged and peculiar kind seems to be an essential component of our social-epistemic and perhaps self-regulative agency. And it might be basic to our nature that we are social-epistemic and self-regulating creatures, even if there can be local breakdowns in these agential profiles of ours. The important point is that such breakdowns

cannot conceivably be global, lifelong possibilities. So, if the Inner Scanner Account is compatible with the possibility of self-blindness, then perhaps so much the worse for it.

Alex Byrne, himself no proponent of the Inner Scanner Account, nevertheless defends it against this objection in two ways: first, by arguing that Shoemaker's arguments against the possibility of self-blindness are problematic and,[191] second, by arguing that even if rational agents are not self-blind, this may only entail that, while inner scanners are our primary means of acquiring self-knowledge, rational agents have a "backup" system that protects them against self-blindness in the event that the inner scanning mechanism fails.

I will not rehearse Byrne's objections to Shoemaker, since I have just reiterated that I too find his arguments problematic. What, then, of Byrne's second claim about our so-called "backup" protection against self-blindness? Here I wonder why a backup system for acquiring privileged and peculiar self-knowledge should be relegated to the position of a backup system at all if, unlike the inner scanning system, it is what protects us against self-blindness. Shouldn't such a system be, *ceteris paribus*, superior to the inner scanning system precisely in virtue of providing this additional layer of security? And shouldn't this give us reason to conceive of it as the primary system, whatever exactly it is?

Well, perhaps all things aren't equal: perhaps the inner scanning system is more cognitively streamlined—less demanding of cognitive resources—than the backup system, such that we could not constantly rely on the backup system, on pain of cognitive overload. But this seems implausible. For, if we are admitting that the inner scanning system could break down, thus forcing us to rely on the backup system, then the backup system would *not* protect us from self-blindness if *it too* was at risk of breaking down. Moreover, simply stipulating that this risk is

[191] The most Byrne is prepared to say is that self-blindness "does not actually appear to occur" (2018, p. 49).

non-actual seems questionable: what we would need is a fuller account of the backup system, and at this point it would seem that we are just looking for a new substantive account of privileged and peculiar self-knowledge. For, once again, if our picture of the backup system is of a system that is not prone to breakdown in situations of cognitive overload, it will be tempting to think of this as the primary system, such that any subsequent appeal to the Inner Scanner Account is explanatorily superfluous.

Byrne believes that "the leading objections leave the inner-sense theory pretty much unscathed" (2018, p. 26). Obviously, I disagree. That said, I mentioned above that Byrne himself rejects the Inner Scanner Account. If he does not think the theory goes unscathed, why does he reject it? His reason is that he takes himself to have discovered a more parsimonious substantive account of privileged and peculiar self-knowledge, one that does not require us to posit a *sui generis* causal mechanism dedicated solely to the production of self-knowledge. I consider his account now.

§5.4.1—Transparency to the World Revisited

Byrne's substantive account of our privileged and peculiar access to our mental states is an extension of an insight that was briefly noted in §2.6.2. There, I pointed out that the Neo-Expressivist account of avowals was able to capture an interesting phenomenon, one that is usually referred to as the *transparency to the world* of avowals. Here, again, is Gareth Evans:

> [I]n making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outward—upon the world. If someone asks me "Do you think there is going to be a third world war?" I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question "Will there be a third world war?" (1982, p. 225).

The observation is that a question of what one believes can be settled by directly considering the world (this being what one's belief is about),[192] rather than by turning one's attention inward unto one's own mental state as such. This is the sense in which a question like "do I believe *p*?" is transparent to the question "*p*?" The same is also often thought to hold for other intentional attitudes—e.g., thinking about whether *φ-ing is desirable* as a way of answering questions about what one desires, or thinking about whether *one will φ* as a way of answering questions about what one intends.[193]

Now, while the Neo-Expressivist's use of this transparency idea is that we can look outward as a way of enabling us to directly express our *first-order* states through avowals, Evans's suggestion has also been treated by Byrne (2005, 2011a, 2011b, 2018) and others as suggesting a method by which we can acquire self-knowledge. In developing his account, Byrne attempts to answer two questions left unresolved by Evans: (1) what psychological mechanism explains this movement of mind from world- to mind-directed thought, and (2) how we can be warranted in having the mind-directed thoughts that arise from world-directed ones?

The first question is interesting because an answer to it will mark a stark departure from the Inner Scanner Account. After all, the Inner Scanner Account is *introspective*: it posits a mechanism that looks to detect one's mental states themselves. But the idea of a transparency-based account is *extrospective*: it posits an outward-looking process that yields self-knowledge. The second question is important in light of the facts, e.g., that *believing p* or *desiring to φ* are

[192] As noted in §2.6.2, Evans's "must" is probably too strong.
[193] See also Barz (2014) who extends the transparency suggestion to our self-knowledge of our wishes, and Paul (2015) for an earlier account of how to extend a transparency method to intentions. See Andreotta (2020) for additional takes on transparent self-knowledge of beliefs, desires, intentions, and wishes.

quite different from *p* and *φ-ing's being good to do*, such that there is a prima facie puzzle about how appreciating the latter set of facts could justify the former self-ascriptions.[194]

§5.4.2—The Inferential Transparency Method

Byrne's answer to the first question above—the question of what psychological mechanism generates a movement from world- to mind-directed thought—is that we draw *inferences* from world-directed premises to mind-directed conclusions. I will refer to this as Byrne's *Inferential Transparency Method*, or ITM.

As Byrne sometimes puts it, ITM is the view that we deploy "epistemic rules" (2018, p. 102) to generate self-beliefs from first-order, world-directed premises. Here are some examples of the epistemic rules Byrne envisions, corresponding to how we can acquire self-knowledge of our beliefs, desires, and intentions, respectively:

(BEL): if *p*, believe that you believe that *p.* (2018, p. 102)

(DES): If ϕ-ing is a desirable option, believe that you want to ϕ (2018, p. 161)

(INT) If you will ϕ, believe you intend to ϕ. (2018, p. 169)

Drawing inferences in accordance with these epistemic rules—these transparent inference rules, you might say—respects Evans's original insight because one moves from a first-order, world-directed premise to a self-ascriptive conclusion; the method is extrospective rather than introspective. Moreover, this account is more ontologically parsimonious than the Inner Scanner Account, since it only requires us to appeal to our general capacity for inference; no *sui generis* self-scanning mechanism is required.

Now that I have explained how ITM conceives of the psychological process involved in acquiring self-knowledge, I must say more about the warrant we have to draw the relevant

[194] Hence O'Brien's calling it the "two topics problem" (2007, p. 103).

inferences. Before proceeding, however, I want to note and address a few objections. The first objection is that, because one cannot follow any of these transparent inference rules without judging that their antecedents are true, one can only use ITM to acquire self-knowledge of one's attitudes that are based on one's own judgements. This has been a source of discontent for some commentators,[195] since it leaves unanswered the question of how one can know those of one's attitudes that do not correspond to one's own judgements. However, I will not be adding myself to the pile-on here. This is because my aim is to account for the self-knowledge that we have of our commitments specifically, which *are* attitudes that are brought about by our judgements (§4.4.2). If we are interested in how we know our commitments specifically, the restricted applicability of ITM is no objection.

A second objection will lead us directly into Byrne's defense of ITM. This is the concern that, on the face of it, ITM won't be able to speak to the *peculiar* character of our commissive self-knowledge. This is because it treats self-knowledge as the result of an inference and, since inference is a ubiquitous route to acquiring all kinds of knowledge, one might doubt that it can reasonably be viewed as a *peculiar* route to self-knowledge. Not only this, but one might even wonder how privileged such knowledge could be, since in general we can easily draw inferences to false conclusions. Why, then, should transparent inferences fare any better?

As we proceed, I will occasionally draw our attention to how the different transparent inference rules fare with respect to these concerns. However, for ease of exposition, I will largely focus on what I take to be Byrne's most promising transparent epistemic rule, namely (BEL). To reiterate:

(BEL): if *p*, believe that you believe that *p*. (2018, p. 102)

[195] See, e.g., Gertler (2011b).

Focusing on (BEL), we can offer a few reasons for thinking that it can yield peculiar self-knowledge of our commissive beliefs. One reason concerns a point that Byrne repeatedly stresses about (BEL), which is that one can acquire self-knowledge via (BEL) *even when p is false*. In such a case, one's *attempt* to follow (BEL) will involve falsely believing that *p* and moving to self-ascribe a false first-order belief. In that event, the falsity of one's first-order belief does not falsify one's second-order belief, since one is just self-ascribing *one's own* false first-order belief. This is why Byrne describes (BEL) as "self-verifying" (2018, p. 104): the mere attempt to follow it yields true self-beliefs. It would seem, then, that such inferences are more or less *infallible*. This, surely, makes (BEL)-inferences highly distinct as a kind, and so makes a case for the peculiarity of transparent inferences as a route to self-knowledge.

A second reason for optimism about (BEL)'s capacity to yield peculiar self-knowledge is that, as Byrne also stresses, transparent inference rules like this can *only* deliver self-knowledge. We can see this by imagining a third-personal variant of (BEL), e.g., *(Pete-BEL)*: *if p, believe that Pete believes p*. (Pete-BEL) is nowhere close to self-verifying because my first-order beliefs don't tell me anything about Pete's (self-)beliefs. Thus, because only the agent herself can acquire transparent self-knowledge, we have another reason to think of ITM as a peculiar method for acquiring self-knowledge. Equally importantly, all of this explains why self-knowledge acquired via (BEL) is privileged: (BEL)-inferences are self-verifying and, hence, bound to be more reliable than others' inferences about one's mind.

Actually, while I have just described (BEL) as self-verifying according to Byrne, this is not *exactly* right. After all:

> …since inference is not instantaneous, there is no cast-iron guarantee that one's belief in the premiss will remain by the time one reaches the conclusion, in which case one's belief that one (now) believes that *p* will be false…

Byrne's own response to this sort of scenario—a scenario involving what we might call *inferential belief extinction*—is that it is no real threat to (BEL)'s self-verifying status, because: "…[s]ince the chain of reasoning [via BEL or other transparent epistemic rules] is as short as it gets, this possibility can be ignored" (2011, p. 206).[196]

Let us grant that inferential belief extinction is at most a rare possibility. Whether these errors are to be construed as *brute* errors is a further question, one that I am not sure admits of an easy answer. The reason is that it is hard to imagine the conditions under which inferential belief extinction might occur. Perhaps it is true that, in most cases of inferential belief extinction, something would have to be psychologically amiss with the agent herself if the belief-base from which she draws her (BEL) inference ceases to exist in the course of drawing that inference. That being said, there may also be cases where one moves from a short-lived belief to a false self-ascription simply because of the nature of the belief—perhaps a belief about one's distance relative to an object that dissipates just as soon as one's distance relative to that object changes (perceptual beliefs in general strike me as good candidates for being occurrent beliefs with potentially very short shelf lives). If inferential belief extinction were to occur in such a case, I doubt that the result would be easily diagnosed as resulting from a psychological failing on the agent's part.[197] So there may be a problem here for Byrne's account if we continue to suppose that avoidance of brute error is a criterion for a workable account of privileged and peculiar self-knowledge.

It might be possible to avoid this concern by accepting the Taking Condition (§3.4) and, in turn, embracing a particular conception of the "taking-beliefs" or taking-attitudes involved in inference. Here is how. On one account of what it is to take one's premise(s) to support one's

[196] See also Byrne (2018, p. 104).
[197] See Winokur (2021) for more on this issue.

conclusion, inference occurs when an agent attaches "inferential force" (Hlobil 2019) to an ordered set of propositions that are all kept in mind at the same time. So, on this view, they must all be there before my mind in order for me to attach inferential force to them, since inferential force is conferred upon each proposition in the inference all at once.[198] On another account, my conclusion-belief is a conclusion-belief because it possesses the *form* of a conclusion belief, but it possesses this form *if and only if* it stands in a formal relation to my premise-belief(s). Crucially, if I cease to have my premise-belief(s) before I form my would-be conclusion-belief, then it does not take on the form of a conclusion at all, and so I have not actually inferred anything.[199] If something roughly like either of these conceptions of taking-attitudes is right, then Byrne can potentially help himself to an even stronger conception of (BEL)'s epistemic pedigree. For he might be able to say that inferential belief extinction is not really possible after all. This being said, we will see in §5.5.1 that embracing TC actually undermines ITM.

Before considering the above issue, I want to wrap up my general presentation of ITM. One further point to note here is that, while Byrne sees inferential belief extinction as the only possible fail-case for (BEL), he acknowledges additional room for error for other transparent inference rules. Take (DES), by way of example. I may look to the world and discern that φ-ing is a desirable option and thereby self-ascribe a desire to φ, as (DES) calls for me to do.[200] Alas, while "[o]ne's desires *tend to* line up with one's knowledge of the desirability of the options" they *merely* tend to so align (2018, p. 161, emphasis mine). So errors may be possible here. Put differently, because the inference base for a (DES)-inference is a *belief* about the desirability of φ-ing, there is no guarantee that moving from this *belief* to a self-ascription of *desire* will be true.

[198] In fact, if inference begins and ends with a single act of attaching inferential force, this may mean that inference is not a temporally extended mental movement from one mental state to another.
[199] Kietzmann (2017, p. 300).
[200] Byrne's transparent epistemic rule for desire: "DES: If ϕ-ing is a desirable option, believe that you want to ϕ".

But so long as our desires do indeed tend to line up with our beliefs about what is desirable, Byrne thinks that we can still view (DES) as a highly reliable transparent inference rule. It is, in other words "*practically* self-verifying" (2018, p. 162). Accordingly, he thinks that (DES) can still provide us with privileged self-knowledge of our commissive desires (and a similar lesson might be apply to (INT)).

§5.5.1—Standard Objections About ITM's Reasonability

I have just given a general defense of ITM as an account of privileged and peculiar self-knowledge. Nevertheless, it has been argued that ITM cannot yield warranted self-beliefs. Focusing again on (BEL), the objection is that (BEL)-inferences cannot be warranted because, as Matthew Boyle argues, any such inferences would be "mad" (2011, pp. 230-231). This is because (BEL)-inferences are neither deductively valid (*p* does not entail that I believe that *p*) nor inductively strong (there are indefinitely many first-order issues that I have never considered and hence have no first-order beliefs about).[201]

In reply, Byrne has simply reiterated that the self-verifying nature of (BEL)-inferences makes them reasonable despite their lacking other canonical good-making properties of inference. In particular, such inferences yield *safe* self-beliefs. A belief is *safe* if it could not have easily been false, i.e., is not merely luckily or accidentally true. Self-beliefs acquired by Byrne's transparent inference rules are safe because self-verifying or practically so. Byrne concludes that the only reason for skepticism about (BEL)'s reasonability is skepticism about the sufficiency of safety for knowledge. But safety *is* sufficient for knowledge, he claims.[202]

[201] See also Sorgiovanni (2018, en. 6).
[202] Byrne says that "[s]afety is a plausible necessary condition for knowledge." Moreover, "provided it is emphasized that the relevant sense of 'could not have easily been false' cannot be elucidated in knowledge free terms, there is no obvious reason to suppose that it is not also sufficient" (2018, 110).

A related charge, also raised by Boyle, is that an agent who follows (BEL) cannot explain why, from her own perspective, her self-belief is reasonable. The charge is related because, as Boyle reminds us, *p* is not good evidence for the fact that one believes *p* (nor, again, is φ's being desirable good evidence that one desires to φ). In reply, however, Byrne notes that this is exactly what we should expect. For, like many other philosophers, he takes it as a datum that privileged and peculiar self-knowledge is not based on first-personally available evidence.[203] This is a datum that is only disputed by some Rylean types.[204] Rather than serving as a crucial objection,

[203] See, e.g., (Davidson 1984; Wright 2001; Bar-On 2004; Roessler 2013)

[204] While we are on the subject, let me address a different sort of inferentialist account of self-knowledge that might be thought of as an improvement on the Rylean view, since it can at least respect the appearance of the evidential groundlessness of self-knowledge (you can feel free to read this footnote after reading the rest of §5.5, so as to not break the flow of my discussion of INT). I have in mind Gopnik's (1983) "theory-theory" of self-knowledge, which argues that self-knowledge seems evidentially groundless because we have become such skilled self-mindreaders that the inferences we draw about our mental states on the basis of our behaviour leave no phenomenological trace. In this way, we are like the seasoned scientist who can *simply see* highly complex phenomena play out before her eyes in the lab without actually being in direct contact with them. Now, preserving the contrast between the seeming privilege of self-knowledge and non-privilege of knowledge of other minds depends, for Gopnik, on the claim that we are expert self-mindreaders but novice other-mindreaders. But why should this be? The intuition that our self-knowledge would still seem privileged even if we were comparing it to the knowledge we have of a very well-known friend or family member's mind gives the lie to Gopnik's strategy here (cf. Bilgrami 2006, p. 17).

Moreover, it is hard to see how an inferential account of knowledge of any mind is to avoid circularity. On the one hand, if these accounts understand inference as a person-level phenomenon, they must reject TC, and this shoulders them with various further burdens (see §3.4). A second circularity worry is that it is hard to see how we could draw inferences about the mental states of a person solely from premises about mere bodily movements, i.e., behaviour non-intentionally described (McDowell 1983; Bilgrami 2006, pp. 19-20). In other words, the inferential base, if it is to be understood this way, is too impoverished to license reasonable inferences about mental states. Indeed, this gives reason to think that knowledge of other minds is at least sometimes perceptual, precisely as Neo-Expressivists believe. The trouble for Gopnik, then, is this: her account requires that we begin from premises about *intentionally described* behaviour, and so presupposes a degree of knowledge of the mind being read (see also Coliva 2016, pp. 86-87). Similarly, it is hard to think about self-mindreading in terms that do not presuppose self-interpretation. But as we saw in §4.6, that self-interpretation presupposes self-knowledge (cf. Davidson 1990, 1991).

Now, what if the behavioural basis for our inferences incorporates elements other than behaviour? Here we might follow more recent inferentialists like Lawlor (2009), Carruthers (2011) and Cassam (2014). Some of these authors argue that our knowledge of our intentional attitudes comes from interpretations of 'inner promptings' like feelings or other phenomenological elements of occurrent thought (Lawlor restricts her attention to desire). Cassam admits that these promptings themselves aren't inferentially known, since otherwise we would not yet have an inference base. Now, because the inference base is not mere (externally observable) behaviour, the objection to Gopnik does not arise. Finally, the suggestion is that these inferences take place sub-personally, i.e., at the system 1 level (so TC is not violated). However, this last point is problematic. For, if the inference takes place sub-personally, then its results may also be confined to the level of system 1 cognition. If so, we must ask the question anew: how can the agent make use of her system 1 self-ascription at the system 2 level? At this point it would seem that we must introduce some further mechanism to make this system 1 deliverance accessible to higher-order cognition. But in my view, that is simply to ask the question of how we acquire self-knowledge all over again. It is to ask how *the agent* acquires self-knowledge, not how her subpersonal modules do (cf. Coliva 2015, p. 251; Coliva 2016, pp. 87-88).

then, we are simply reminded that the first-person perspective is silent on the nature of the warrant one's self-beliefs enjoy (cf. Roessler 2013).

For these reasons, I take it that Boyle's arguments against the reasonability of (BEL)-inferences (and transparent inferences more generally) are ineffective against ITM. However, all is not rosy, or so I shall argue. For, even though Boyle's criticisms are inconclusive against ITM, there are a few crucial problems for the account.

§5.5.2—The Inferential Transparency Method Violates the Taking Condition

In §5.4.2 I suggested that Byrne may not need to allow for the possibility of *inferential belief extinction*, and so for one potential case of (potentially) brute error, so long as he embraces TC. For, as we saw, at least some accounts of what it is to take one's premise(s) to support one's conclusion metaphysically necessitate that we hold to our premises at the time we draw our conclusion. This closes off the possibility of a temporal gap between one's holding a first-order belief and one's forming a self-belief about it in a transparent inference, and so renders inferential belief extinction impossible. However, focusing again on (BEL) as an example, I will now show that ITM cannot be paired with TC (though the objection generalizes to all of his transparent inference rules).

To begin, note that inference turns out to be a self-conscious activity on many TC-based accounts,[205] wherein an agent must already be aware of her premise-belief(s) in the course of taking them to support her conclusion-belief. If this is the case, then ITM cannot explain how we acquire self-knowledge, for one must already know that one believes *p* before one can "take" one's belief that *p* to support *q* (where *q*, in a (BEL)-inference for example, is a self-ascription of a belief that *p*). Now, one might try to push back by arguing that one's taking-attitudes can be

[205] See Rödl (2007), Boyle (2011), Koziolek (2017), Kietzmann (2018).

attitudes toward propositions and relations of epistemic support, rather than attitudes towards one's own mental states *as such* (this being an argument that I myself pursued in §3.4).[206] On this strategy, taking *p* to support *q* does not require explicitly viewing *p* and *q* as the propositional contents of one's own beliefs, and so no self-knowledge would be presupposed in a (BEL)-inference that moved from a belief that *p* to a self-belief that one believes that *p*.

The problem is that, even granting the general possibility of such 'second-tier' inferences (§3.3.1), the idea of a second-tier (BEL)-inference is unintelligible. For consider a taking-mediated (BEL)-inference, e.g., my taking *p* (*there are fifty sports fans in this bar right now*) to epistemically support a self-belief that *p* (*I believe that there are fifty sports fans in this bar right now*). Our question now is: what epistemic support relation do I appreciate here? As observed in §5.5.1, it cannot be that I appreciate the former proposition as providing deductive or inductive support for the latter. To my mind, then, it can only be that I take the former to support the latter because self-ascribing a belief on the basis of the former is self-verifying. But how can I appreciate *this* except by taking it that the former's *being what I believe* is what makes its self-ascription self-verifying? Because this sort of appreciation presupposes self-knowledge, it cannot be appealed to as part of an account of how I acquire self-knowledge. For most inferences, where I can take my premises to support my conclusions by recognizing inductive or deductive support relations between them, my taking-attitudes could perhaps have mere propositions and logical/evidential support relations as their objects. But for (BEL)-inferences, there is nothing for a second-tier taking-attitude to be: I can only take *p* to support its self-ascription by recognizing *p qua content of my belief*.[207]

[206] See also Winokur (forthcoming).
[207] I believe that this is a way of decisively developing a similar point made by Boyle (2011, p. 231).

At this point, I believe Byrne must grant that ITM cannot be reconciled with TC. And indeed, this seems to be his view:

> An assumption of this book is that the pertinent kind of reasoning is relatively undemanding: it can occur without self-knowledge, and without an appreciation of one's evidence or reasons as evidence or reasons. (2018, p. 100, fn. 1)

Unfortunately, it is not so easy to reject TC. For Byrne to reject it, he will have to offer some other account of inference's agential nature (something that he should like to preserve, given his conception of inference as a matter of rule following). He will also have to explain how inference differs from mere association, how we can count as doxastically responsible reasoners, how we can explain Moore-paradoxical phenomena associated with inference, and so on (seem again, §3.4).

§5.5.3—The Inferential Transparency Method and Laypeople

I have argued that ITM is incompatible with TC, and that Byrne faces a significant burden in rejecting TC. However, I also think that there is a more general problem for ITM in the vicinity. In brief: there is a problem with ITM because it is hard to understand ordinary agents as using the relevant epistemic rules.

To see the problem, note first that while not all rule following is conscious or self-conscious, it is usually relatively easy for the cognitively mature among us to recognize the reasonability of the rules we follow in different theoretical or practical domains, at least when those rules are explicitly articulated to us by someone else. Even semantic rules (e.g., *apply 'green' only to green things*), these being plausible candidates for rules we usually learn to follow without ever first doing so self-consciously (Boghossian 2016, p. 8), can be reflected on by ordinary agents without one's suddenly finding them unreasonable.

The problem now is that grasping the reasonability of rules like (BEL), (DES), and (INT) seems to require knowledge of sophisticated philosophical arguments like those provided by Byrne. These arguments are sophisticated to the extent that they require us to recognize the rationality of transparent inferences as a product of their self-verifying and epistemically safe status despite the fact that they actively flout basic inductive and deductive rules of inference with which most of us are simultaneously (again, even if only tacitly) familiar. I worry that most ordinary, cognitively mature agents have not thought, nor would easily recognize, that such inferences are reasonable despite their easily appreciated logical shortcomings. But according to ITM, ordinary agents acquire self-knowledge by way of ITM, and so it has to be that agents are already following (BEL), (DES), and (INT) despite this fact.

What this means is that, for most ordinary agents, it is implausible that they would find it implausible that they follow (BEL), (DES), or (INT) if it were suggested that they do so. And this strikes me as (admittedly defeasible) evidence that ordinary agents do not follow them: if they followed them, suggesting these rules to them would not strike them as potentially bizarre or overly difficult to endorse. So ITM seems to be a poor psychological account of how ordinary agents acquire self-knowledge.[208]

Note that the problem remains even if we reject the Taking Condition. For, while we may not think that drawing an inference necessarily requires one to be cognizant of epistemic support relations between its premise(s) and conclusion, this is beside the point here. Rather, my claim is that, whether or not agents necessarily take their premise(s) to epistemically support their conclusions in the course of drawing an inference, an ordinary agent who uses Byrne's

[208] Not only this, but I suspect that if they *subsequently* took these rules to be *bad* ones for whatever reason (say because Boyle had convinced them), it would be strange if we had to keep understanding them as tacitly following these rules.

transparent inference rules is likely to be unsure whether they really use them *once they consider whether they do so*. Note also, and finally, that I am not denying Byrne the opportunity to say—as he does to Boyle—that we follow these rules sub-consciously (which is what explains the apparent groundlessness, from the first-person perspective, of so much of our self-knowledge). For I am not suggesting that *as one self-ascribes* a mental state in a privileged and peculiar way, the method that one uses to do so (and its source of warrant) must be transparent to one. Again, the point is only about what might happen when agents become aware of all this. Even if we follow rules sub-consciously, if we are really following them it seems odd to suggest that they should strike us as bizarre rules once they are made consciously available to us.

I am going to move on from transparency accounts at this point, despite not having discussed every possible version of them in detail.[209] Still, as we will see, this does not require abandoning Evans's basic insight. This is because the insight that many of our self-beliefs and self-ascriptions are transparent to the world can be accommodated within an account of self-knowledge even without supposing that there must be a transparency *method* at work.

[209] One might especially wonder why I have not offered a detailed discussion of Moran (2001, 2003, 2004), who likely deserves the bulk of the credit for the recent explosion of interest in transparency methods among philosophers. The reason is that I concur with O'Brien (2003) and Coliva (2016) that Moran does not offer a transparency account of self-knowledge so much as he offers an account of how we can exercise authority as authorship over our attitudes. So his account of self-knowledge, to the extent that it is one, does not explain the sort of self-knowledge I am interested in here and is, instead, better viewed as an account along the lines endorsed by McGeer and Vierkant (see §4.8). The same is true of Peacocke's (2017) recent account, which I take to be another play on the Moranian self-authorship theme.

There are other transparency accounts that aim to explain the sort of self-knowledge I aim to explain, such as Shah and Velleman's (2005) account according to which I put to myself the question whether "p?" as a brute stimulus and see what "response" echoes back (see also Barz 2019). However, I agree with Moran that "[s]imply hearing oneself coming out with something in response to a brute stimulus will provide no more reason for thinking this represents one's belief [desire, intention, etc.] about something than if one were to sneeze in response to the stimulus" (2012, p. 221). Next, there is the 'bypass' transparency account due to Fernández (2013), though see Coliva (2014, 2016) for what I think are powerful criticisms. Finally, there is a transparency method recently proposed by Andreotta (2020). I confess that I have not had time to consider his account in sufficient detail, as this chapter was already substantially written by the time I came across it.

On another, final note, one might care to look at Parent's (2017) transparency-inspired account of how we can *know when* we have self-knowledge, though this account does not explain how we can *acquire* self-knowledge.

§5.6.1—Constitutivism About Peculiar Self-Knowledge: A Crude First Pass

The Inner Scanner Account and transparency accounts could not be more different in their details—one being quintessentially introspective and causal, the other being quintessentially extrospective and rule-based. Still, they share a key assumption: that there must in fact be a *means* by which we acquire privileged and peculiar self-knowledge. In the remainder of this chapter, I will develop an account that rejects this assumption. That is, I will argue that we can account for privileged and peculiar commissive self-knowledge in *metaphysical* rather than *epistemic* terms. The sort of account I will be developing in the remainder of this chapter, then, is a form of *Constitutivism*: it is an account according to which our commitments are partly constituted by our self-knowledge of them.

Just to get us started, consider the following:

> **Constitutivism-Crude**: if an agent is in some mental state, M, she believes that she is in M, and if she believes that she is in some mental state M, she is in M.

Focusing on the first conditional, commissive self-knowledge is privileged because our mental states are "self-intimating" (Shoemaker 2009): they are self-known simply in virtue of our being in them. Likewise, such self-knowledge would also be peculiar, since nobody else relates to one's mental states in such a way that they know them whenever one is in them. Focusing on the second conditional, the logic is similar. If an agent is in M whenever she believes that she is, then her self-beliefs are maximally reliable—more so than any other empirical beliefs—and so amount to self-knowledge that is privileged relative to other kinds of empirical knowledge.

Unfortunately, CC is not at all plausible. First, it introduces no qualifications to capture the possibilities of self-ignorance or self-error. Second, it likely generates a vicious regress. For, if any mental state M is self-known whenever one is in it, then CC dictates that the possessor's self-knowledge of it, being itself a kind of warranted (self-)belief, will also be a mental state that,

in turn, must be self-known to its possessor, *ad infinitum*. Third, it applies indiscriminately to all types of mental states, whereas a refined version likely ought to restrict the scope of Constitutivism to a subset of mental states. A refined version of Constitutivism will address these and other problems. My point in discussing CC, then, is not to present it as a plausible version of Constitutivism, but to offer a sense of what it is to propose a metaphysical account of self-knowledge—an account according to which self-knowledge simply comes from being in a mental state or believing that one is in it, rather than from some *method* that puts one in touch with one's mental states.

Before I turn to the task of developing a superior constitutive account in detail, let me give a few initial reasons to think that the Constitutivist strategy—understood simply as a metaphysical account of privileged and peculiar self-knowledge—is promising. First, Constitutivism affords maximal respect to the intuition that we can have groundless self-knowledge, for there is indeed no epistemic method on which such knowledge is based, and this is precisely how things appear from the first-person perspective. Granted, other accounts we have considered can accommodate the first-personal appearance of groundlessness. Indeed, we have considered both introspective and extrospective accounts that appeal to sub-conscious psychological goings-on. But these accounts, being at pains to offer a hidden method or process by which we acquire self-knowledge, betray their uneasiness about the groundlessness in a way that Constitutivism does not.

Second, a version of Constitutivism is already plausible as regards those self-beliefs that *bring about* the very mental states they self-ascribe (see §4.7.3). Thus, McGeer (2015) is a kind of Constitutivist: she argues that no scanning, tracking, or detecting of our commitments is required because we ourselves constitute our commitments by self-ascribing them (see also

Wright 2001; Heal 2002; Coliva 2009, 2012, 2016). Of course, this does not bring us all the way to CC. This is because McGeer does not argue that commitments that are not determined by their very self-ascription are self-intimating. Nor does she argue that self-ascriptions that do not determine their objects are nevertheless true. Still, the point for now is just that the very idea of constitutive self-knowledge is not, as a wholly general matter, implausible.

Third, it is easy to read many agentialist/rationalist non-substantive accounts of self-knowledge as at least implicitly friendly to Constitutivism. Recall, for example, Shoemaker's rationalist view that self-knowledge of at least some of our attitudes supervenes on a certain degree of conceptual sophistication, intelligence, and rationality (see §3.5.1). On such a view, no epistemic method must be utilized in order to acquire self-knowledge so long as one meets these conditions. This, on his thinking, is a consequence of the impossibility of self-blindness. Now, whether we agree with Shoemaker that Constitutivism follows directly from the impossibility of self-blindness,[210] there is something intuitive to the more basic idea that self-knowledge of one's commitments, being necessary to many of our basic rational/agential capacities, is constitutive of having them. Constitutivism, being a metaphysical account of self-knowledge, can be viewed as the best way to preserve deep links between rationality, agency, and self-knowledge.[211]

Fourth, there is good reason to think that a constitutive account of self-knowledge is a natural fit with Neo-Expressivism.[212] The reason is that Neo-Expressivism denies that avowals are backed by episodes of introspective self-detection, and Constitutivism also denies this. Of

[210] See, for skepticism, De Brasi (2015, p. 243) and Parrott (2017). De Brasi argues for the Inner Scanner Account, which I have rejected, whereas Parrott argues for a view that is in some respects similar to Constitutivism. I discuss Parrott's account in §5.7.5.

[211] Horgan and Kriegel (2007) advance a constitutive account of self-knowledge of phenomenal experience generally, while Shoemaker (1994, 1996a, 1996b) and Coliva (2016, chapter 8) advance constitutive accounts of self-knowledge of sensory states. Finally, Giananti (2020) advances a constitutive account of perceptual experience. I won't evaluate these accounts in this dissertation.

[212] Though Bar-On (2004, 2009, 2015, ms.) is not sympathetic to this marriage.

course, the Neo-Expressivist will deny that the agent's constitutive self-knowledge explains the authority of her avowals, but this is no objection to Constitutivism. For, by itself, Constitutivism is just a view about how one's mental states relate to one's self-beliefs. Whether this constitutive relation should be taken to explain authority is a further question that the Constitutivist can admit, if she likes, is best explained along *non-epistemic* Neo-Expressivist lines.

Fifth, the Constitutivism/Neo-Expressivism pairing gives us a way to see Constitutivism as less objectionable than it has sometimes appeared. The contemptible Cartesian view that so many philosophers now vehemently deny is the view that we have self-intimating *private* states of mind. It is the privacy component of this view that is so contemptible—the other component is less obviously so, at least so long as we offer a compelling story of the proper scope of self-intimation. Now, if we accept that expressibility is the mark of the mental (as Bar-On 2004, 2009 does), rather than privacy or incorrigibility, then Constitutivism is not of a piece with false Cartesian dogma. For the mere fact that a mental state may intimate itself to its possessor does not mean that it is intrinsically private, nor that one's mental states cannot fail to self-intimate in certain conditions. After all, if one is also a Neo-Expressivist, it can be publicly manifested through avowing.

Sixth, and finally, because Constitutivism dispenses with recognitional/detectivist understandings of self-knowledge, the impossibility of brutely erroneous self-beliefs is readily explicable. Again, a brute error about one's mental state would be one that occurs through no psychological failing on one's part. Constitutivism can be construed in such a way that there is no possibility of brute error. For there is no mechanism by which a psychologically well-functioning agent could fail to detect her mental states, whether due to some fault of her own or due to the world's failure to cooperate. This does not mean that self-ignorance and self-error are

impossible. It only means that they cannot be explicated in terms of mechanistic failures or failures in the use of a method.

§5.6.2—Refining Constitutivism about Commissive Self-Knowledge

In the previous subsection I presented a maximally crude version of Constitutivism, CC. In this section I refine the core constitutive thesis and argue that it can meet some of the standard replies facing Constitutivist accounts.

CC was a conjunction of two conditionals, one claiming that being in M entails self-believing that you are in M, and another claiming that self-believing being in M entails being in M. As aforementioned, these conditionals are implausibly strong. Wisely, therefore, most Constitutivists go in for weaker conditionals. Consider, for example, what Bilgrami refers to as *authority* (A) and *transparency* (T):

> (A). It is a presumption that: if S believes that she desires (believes) that p, then she desires (believes) that p. (2006, p. 30)[213]

> (T). It is a presumption that: if S desires (believes) that p, then S believes that she desires (believes) that p. (2006, p. 31)[214]

There are two major respects in which (A) and (T) are more modest than those that make up CC. First, (A) and (T) only concern constitutive self-knowledge of beliefs and desires, whereas CC indiscriminately applies to all mental state types. Moreover, (A) and (T) only refer to *presumptions* of constitutive connections between beliefs, desires, and self-beliefs. Much of Bilgrami's effort is expended, in his subsequent discussions of (A) and (T), on accounting for the conditions in which these presumptions actually obtain. Those conditions, he writes, are what I

[213] Authority is not to be confused with *first-person* authority, as discussed in Chapters One and Two.
[214] Transparency is not to be confused with Evans's thesis about the transparency-to-the-world of self-knowledge.

referred to in §4.2.1 as *conditions of agency*. For Bilgrami, conditions of agency are conditions in which one can act freely and responsibly.[215]

Tying all of this together, and adding intentions to the mix as another kind of commitment, we can now consider another (still provisional) version of Constitutivism:

> **Constitutivism-Provisional**: Given C, one believes/desires/intends that P/to φ iff one believes that one believes/desires/intends that P/to φ.

While Constitutivism Provisional (CP) is inspired by Coliva (2016, p. 164), I believe that it captures much of what Bilgrami has in mind with (A) and (T). One difference is that, whereas Bilgrami writes of conditions of agency, Coliva's reference to 'C'—or 'C-conditions'—serves as a placeholder for whatever the relevant conditions are in which the biconditional holds. One C-condition that I accept, obviously, is that the relevant first-order attitudes be understood as commitments. Before turning to other plausible C-conditions, I want to focus in on the idea of mental constitution itself.

Minimally, as I have said, 'constitution' refers to a metaphysical relation between a mental state and one's self-belief about it. Is there more that can be said about this? Here I will mention three proposals, and I'll settle on my preferred variant. We can begin with Bilgrami's proposal:

> No account, not even the constitutive account, should deny the existence of some such underlying causal mechanism linking first-order intentional states and second-order beliefs about them. If first- and second-order states are different states, which they surely are, then some causal link must be connecting them when we have self-knowledge, even if self-knowledge is constitutive of mental states. (2006, p. 37)

On my reading, Bilgrami endorses a relatively weak notion of constitution because he clings to the ontological independence of the attitudes thus related. But even though his aim is to be metaphysically modest about what constitution entails, I doubt that his proposal is all that stable. For, if the constitutive relationship is preserved by an independent causal link, it is unclear

[215] I discussed this in §4.2 without explicitly formulating Bilgrami's account as a version of Constitutivism.

whether we can articulate plausible C-conditions in which such a link is immune to breakdown. For how can a causal link generate a metaphysical relation as strong as constitution?

Granted, Bilgrami adds that, while causality is part of the aetiology of constitution, it is not part of the explanation of what makes such knowledge privileged and peculiar. Rather, what explains this is the conceptual relations between commitments, freedom, responsibility, and self-knowledge (§4.2). And yet, this conceptual story makes a metaphysical claim to the effect that one's commitments are necessarily known in C-conditions. So we are simply returned to the initial problem: how can an intervening causal link between one's commitments and one's self-beliefs about them be *anywhere* part of his account of self-knowledge?

Another option comes from Matthew Boyle, who denies that "in the normal, non-alienated case, being in a given mental state M and believing oneself to be in M are two distinct psychological conditions" (2011, p. 235). On his view, self-beliefs and the states they are about are actually numerically identical, contra Bilgrami. Taking the case of belief as an example, Boyle says that "in the normal and basic case, believing *p* and knowing oneself to believe *p* are not two cognitive states; they are two aspects of one cognitive state—the state, as we might put it, of knowingly believing *p*" (2011, p. 228).[216] No causal link between distinct mental states obtains in "normal, non-alienated" cases of self-knowledge.

Shoemaker can be interpreted the same way when he says that "…we needn't suppose that each belief or desire produces a separate state which is the knowledge or belief in its own existence" and that "[p]erhaps instead it is the case that insofar as a person is rational [and has the concepts required to self-ascribe her mental states], each belief and desire tends to double as

[216] Similarly, Casey Doyle writes that "[r]ather than thinking of self-knowledge as knowledge of a fact about oneself, we should think of it as a way of being in a first-order state" (2015, p. 5). Indeed, for the cases at issue in this chapter, it is a way of *commissively* being in that state.

knowledge or belief in its own existence" (1996a, p. 33). In meeting the required conditions, it is not that this "…pushes the creature into a new state, distinct from any it was in before…It is rather that adding this enables the core realization of the first-order belief to play a more encompassing role" in cognition (1996a, p. 244). Alternatively, he suggests the possibility that one's first-order propositional attitudes and self-beliefs could have distinct core realizations in the brain, albeit overlapping total realizations. On this view, just like the previous one, there is no total ontological separation between first- and second-order states.

As far as I can tell, however, there is a difference between Shoemaker's proposals and Boyle's. The difference is that, whereas Boyle's constitution relation seems to be one of wholesale identity,[217] Shoemaker's seems to be one of mereology.[218] I am inclined to think that a mereological conception allows us to more easily imagine distinctive second- versus first-order components of the part-whole relations that constitute commissive self-knowledge, so that we are not at risk of collapsing the fruitful *conceptual* distinction between the two, while also being able to argue—*pace* Bilgrami—that the relevant states are not *entirely* ontologically distinct. I also think that the mereological conception of constitution can help us to make sense of cases where constitutive self-knowledge breaks down (more on this later on).

This mereological conception of constitution also squares nicely with my conception of commitments. In Chapter Four I argued that to have a commitment is at least partly to be disposed to draw some number of inferences from it, and I also argued that this disposition partly constitutes one's self-belief that one has this commitment, given that one's self-belief has the

[217] I share this reading with Parrott (2017).
[218] Parent (2017, p. 159) also proposes that our self-beliefs might have their objects as a proper part, though he relies on a language of thought hypothesis to get there (Fodor 1975).

requisite *ratifying form* (§4.6.2, §4.6.5).[219] On a mereological conception of constitution, we can say that this second-order self-belief has a first-order commitment as a proper part, and that this enables one's commitment to figure into quintessentially second-order cognitive processes such as those involved in interpersonal argumentation (§4.6.2), complex group actions (§4.6.3), linguistic interpretation (§4.6.4), and future-directed acts of self-control (§4.7.2).[220]

Having suggested that a mereological gloss on the operative notion of constitution is fruitful, consider CP again:

> **Constitutivism-Provisional (CP)**: Given C, one believes/desires/intends that P/to φ iff one believes that one believes/desires/intends that P/to φ.

Paraphrased, we can say that, given some set of background conditions, one's commissive beliefs/desires/intentions stand in part-whole relations to self-beliefs about one's commitments.

We can now return, in a fuller way, to the question of what the 'C-conditions' of Constitutivism ought to be. In fact, I have already mentioned a number of such conditions in the preceding discussion. Thus, following Shoemaker, it seems reasonable to restrict the thesis to agents who are sufficiently rational, intelligent, and conceptually equipped (1996a, p. 243).

---

[219] See Kietzmann (2018) for a similar idea. In this way, my argument for the ontological indistinctness of commitments and self-beliefs is different from Bilgrami's (2012), since mine draws on the dispositional elements of each component state whereas Bilgrami's relies on the idea that both states come with a disposition to criticize oneself for failing to live up to one's commitments. Schwitzgebel (2009) seems to hold a similarly mereological view of belief and self-belief, except that he does not distinguish commitments from dispositions and only thinks that a constitutive, mereological relationship holds *when we form self-beliefs* in ordinary circumstances about our first-order beliefs, rather than (additionally) that first-order beliefs formed through first-order cognition can bring self-beliefs in their train.

[220] This may be a good time to forestall a potential concern. Williamson (see also Mandelbaum 2014) argues that "the difference between believing *p* and merely fancying *p* depends in part on one's dispositions to practical reasoning and action manifested only in counterfactual circumstances, and one is not always in a position to know what those dispositions are" (2000, p. 24). So one might think that Constitutivism is off to a bad start if knowing one's own commitments requires knowledge of their dispositional structure. However, my claim is not that having constitutive self-knowledge involves being aware of one's disposition to infer various things from one's commitment *under this sort of description*. Rather, one knows one's commitment insofar as one can appreciate at least some of what follows from it (by one's own lights), where this is a *manifestation* of the relevant disposition rather than an *introspection* of it.

Likewise, as aforementioned, we can add that the relevant first-order attitudes must be commitments.

These conditions make CP more plausible because they ward off otherwise obvious and fatal objections, e.g., that CP requires us to implausibly attribute self-knowledge to lower animals and human infants. We can see that the first, crude version of Constitutivism CC forces this implausible attribution, at least assuming that lower animals and human infants have beliefs, desires, and the like, since that view says that being in a mental state entails self-believing that you are in it.[221] This objection fails on CP, at least when we define C as above. For, first of all, the aforementioned C-conditions ensure that only creatures with the requisite conceptual repertoire have constitutive self-knowledge. Moreover, because it also applies to self-knowledge of commitments specifically, it is plausible that infants and animals do not qualify precisely because (or to the extent that) they lack commitments, even if they have other mental states.

Coliva adds that only agents who are "cognitively lucid and alert" (2016, p. 178) as well as "attentive" (Ibid., p. 182) enjoy Constitutivist self-knowledge. While lucidity and alertness strike me as friendly additions, the idea of attentiveness might raise some eyebrows. The worry here is that attentiveness might mean something like *attending to one's mental states*. This would be a problematic C-condition to add, since it would introduce an epistemological (introspective) dimension that Constitutivists reject.[222] Because Coliva is herself a Constitutivist, we should wonder whether she had something else in mind here. My thought is that she may have had in

[221] See Burge (2013, p. 192).

[222] On this note, I worry about Sorgiovanni's gloss on Boyle's "reflectivist" version of Constitutivism, according to which we transition from "tacit" self-knowledge—built already into one's world-oriented attitudes—to explicit self-knowledge by *reflecting* on one's implicit awareness of one's orientation toward the world, thus "making explicit what is implicit in one's subjective awareness" (2018, p. 16). Such an account does not really strike me as Constitutivist: rather, a certain measure of introspection is required. At the very least, worries about the "accuracy conditions" of reflection (introspective or otherwise) need to be addressed (Golob 2015, p. 244). All of this holds even if we are happy to say that there is such a thing as tacit self-knowledge of one's attitudes (as we might be in the case of phenomenal introspection—cf. Giustina 2019).

mind something like attentiveness to the *contents* of one's commitments. That is, in being sufficiently rational, conceptually equipped, lucid, alert, and *attentive to the goodness of φ-ing* (for example), an agent will have self-knowledge of her commissive desire without further epistemic effort.

This attention-to-the-world condition allows us to salvage Evans's insight about transparency, despite not involving the application of any transparency *method*. To see this, consider the following passage from Aaron Zimmerman:

> It is of course true that a rational agent will not believe in the inevitability of a third world war until she has assessed the evidence for and against its occurrence. Thus, insofar as I have not made up my mind as to whether we will go to war, I must first assess the political climate, form a judgment on the matter, and only then report on my attitude toward the issue. But why not think that the evidence for war grounds my first-order belief that war will occur, and that it is the fact that I have this belief that both leads me to believe that I have it and grounds the belief that I have it? It seems that the transparency and direct access [i.e., constitutive] accounts jibe equally well with 'what it is like' to deploy Evans's procedure. (2008, p. 338)

In a nutshell, whereas transparency *method* theorists like Byrne argue that we must do something to transition from a first-order world-direct attitude to a second-order self-belief, Zimmerman argues (on behalf of Constitutivists) that we can acquire self-knowledge simply by attending to the content that determines our first-order attitude. In acquiring the first-order state, that state itself automatically becomes the reason (both epistemically and psychologically) for our self-belief.[223] And since this begins from looking outward, as it were, Evans's original insight is respected.[224]

[223] We can also follow Borgoni in thinking of Constitutivism as friendly to "[t]ransparency-as-coordination" (2018a, p. 57). 'Transparency-as-coordination' refers to the fact that being in a position to express the first-order, world-directed content of an intentional state typically also means that one is in a position to avow one's state, such that one's first- and second-order attitudes are in this way 'coordinated'. Borgoni adds that our non-commissive beliefs do not exhibit transparency-as-coordination: having, e.g., Freudian desire (a dispositional attitude, on Bilgrami's, Coliva's, and my view) does not mean one is in a position to avow it.

[224] This condition also comports well with Tooming's (2020) distinction between "easy" and "hard" self-knowledge of desire (i.e., privileged and peculiar versus its opposite) as a difference between the extent of one's familiarity with its content.

So far, I have clarified a number of features of CP. First, I have clarified that it is a thesis about commissive self-knowledge specifically, and that self-knowledge is simply assured by the mereological relation between our commitments and self-beliefs. Finally, I have clarified that these mereological relations only obtain when we are in the right psychological conditions. These are conditions such as: having the requisite rationality, intelligence, and conceptual capacity, while being sufficiently lucid, alert, and attentive.

Let us now turn to another question, namely, the question of what *warrants* one's self-beliefs in C-conditions, such that in C-conditions we have not only true self-beliefs about but also self-*knowledge* of our commitments. Many Constitutivists are entirely silent on this issue (Bilgrami 2006, 2012; Boyle 2011; Coliva 2009, 2012, 2016). The reason for this silence might be that questions of warrant can seem ridiculous if our self-beliefs literally constitute their objects; it might simply seem obvious that such self-beliefs are warranted owing to their infallibility in C-conditions. However, the above passage from Zimmerman also gestures at an account of Constitutivist epistemic warrant: since we self-believe that we are in a given state *because* we are in that state, one's first-order state is the epistemic reason why one self-believes it. So, it is one's commitment itself that warrants one's self-belief. Perhaps this can seem a little bit mysterious given that the self-belief is not genuinely ontologically distinct from its object, since it can sound tantamount to declaring that commitments are self-warranting. But this is another place where the mereological conception of constitution helps more than the identity-conception. On the mereological conception, a first-order commitment can warrant a self-belief by warranting its higher-order *part*. On the identity conception, we cannot make any such move.

As it happens, this conception of warrant has made contact with certain Neo-Expressivist proposals. As Dorit Bar-On and Drew Johnson argue:

> Like the Constitutivist, the Neo-Expressivist can insist that a true basic mental self-belief will have epistemic grounding – in the very state the belief is about (in addition to the believer being entitled to it by default). Not so when it comes to the corresponding false self-belief. When one is not in M, one can still believe that one is in M, and even be entitled by default to that belief; but one is not (fully) warranted, since one's belief fails to be epistemically grounded. (2019, p. 335)

Notably, however, Bar-On and Johnson identify an additional possibility for warranted self-beliefs, namely, a default *entitlement* to them. Bar-On and Johnson never tell a full story about this. However, I believe it is possible to connect up this entitlement with one's preferred agentialist account of self-knowledge. For a Social Agentialist like myself, this entitlement could be said to derive from the fact that we are the sorts of agents who require self-knowledge of our commitments in order to discursively interact with and interpret one another's speech, and who are therefore entitled to rely on our self-beliefs about what commitments we have (even if, outside of C-conditions, we can go wrong about ourselves).[225]

Because we have added a lot of content to our working constitutive thesis—content that is not explicitly represented in CP—I now propose the following:

> **Constitutivism-Final**: Given C-conditions (rationality, intelligence, conceptual competence, lucidity, alertness, and attentiveness)—one's beliefs/desires/intentions that P/to φ as a commitment stand in a part-whole relation to one's beliefs that one believes/desires/intends that P/to φ as a commitment.[226]

Notice that Constitutivism-Final (CF) does not actually include a biconditional operator like 'if and only if'. Nevertheless, it is intended as a biconditional: if having a commitment partly

[225] Coliva (2016, pp. 157-158) rejects the idea that being in a commissive mental state can itself justify one's self-belief. Her concern is that we cannot ground our warrant for self-beliefs in the very mental states they are about because this will produce the absurd result that even self-beliefs that seem reasonable from one's self-perspective won't be so much as warranted at all when the mental states are not present. However, we can now see that this is not quite right, since *some* degree of warrant can remain even if the warrant provided by M for one's self-belief that one is in M is no longer available.

[226] The reader will note that ccurrent mental actions closely associated with our commitments, such as judgement, doubting, and questioning, are not covered by this thesis, nor passing thoughts that might arise in streams of conscious thinking (see Doyle 2018a, 2020 for more on self-knowledge of conscious thought; see also Alshanetsky 2020 for some tricky cases). Finally, since I am concerned here with commissive self-knowledge, I do not address possible cases of 'first-personal' self-knowledge of attitudes that one does *not* endorse (cf. Hunter 2011; Leite 2018; Doyle 2018b).

constitutes one's self-belief about it in C-conditions, then one has a commitment in C-conditions if and only if one has a self-belief about it.

I began our discussion of Constitutivist views by offering some reasons for optimism, and I eventually regimented my particular Constitutivist account in such a way that it stands in the face of some common objections. It is now time to address a range of further objections. Obviously, my formulation of Constitutivism draws from the work of several other Constitutivists, and so is not endorsed in full by any of the authors whose versions have been objected to in the literature. Still, I will present these objections as though they target Constitutivism-Final, seeing as they also ought to be addressed by a proponent of CF. My preferred responses will sometimes draw on and sometimes diverge from those of other Constitutivists (who also, in some cases, offer no answers at all to the objections). In the end, my claim will be that CF can substantively account for our privileged and peculiar self-knowledge of our commitments.[227]

[227] Here I want to flag one objection that is not addressed to Constitutivists directly, but that is likely worth addressing by Constitutivists. This is the objection that our having privileged and peculiar self-knowledge of commitments is obviously undermined by empirical psychology, given the bevy of studies concerning our "confabulatory" tendencies (e.g., Nisbett & Wilson 1977; Haidt 2001). In these studies, subjects are set up to give answers to various questions or perform certain association tasks, and it is clear that they tend to engage in post-hoc rationalizations as to their true motivating reasons for the attitudes they express. Thus, in the Haidt study, subjects encounter a fictional story about a consensual incestuous relationship. They are then asked to weigh in on its moral status. Inevitably, many agents answer that it is immoral. But when asked to give their reasons, they cite factors such as health risks and psychological traumas that might arise from incest, only to be told again that these factors are ruled out by the description of the case. I agree that the subjects might lack self-knowledge in these cases, but what they lack is knowledge of their *motivating* reasons for their commitments, not knowledge of their commitments (cf. Andreotta 2019).

Now, one might worry that this still leaves us with a disorienting picture of how well we know ourselves, since it leaves us unable to know when it is our commitments, rather than other facts (perhaps, e.g., our merely dispositional propositional attitudes) that serve as the motivating reasons for our responses to various questions. But I do not think that things are really as dire as this, since we may have relatively good knowledge of when our commitments are also motivating reasons (cf. Keeling 2019a, 2019b). I also have a thought that I would like to develop down the road: it may be that our commitments are still properly regarded as motivating reasons for one another even in cases where confabulation occurs. One way this might be argued comes from Smith & Miller (1978), who argue that the reasons we cite in confabulation scenarios might still be *part* of the causal chain leading from initial confabulation to eventual response to the test prompts. But another way to go is to argue that our commitments are our motivating reasons in a specific sense: they are the *self-conscious sustaining sources* of our answers to the test prompts. In other words, they may be the reasons for why we *continue* to hold to the answers we give, and so count as our motivating reasons in a sense, even if they do not motivate our initial answers.

§5.7.1—Constitutivism and the Episodic Requirement

One objection to CF is that it does not account for *episodic* self-knowledge. This is because CF claims that all of an agent's commitments are self-known in C-conditions, even though it is surely true that not all of one's commitments—and self-beliefs about them—are occurrently tokened at any given time in episodes of thought (such as avowals). If this is right, then CF is best read as an account of *standing* commissive self-knowledge—self-knowledge that one is, at best, disposed to occurrently exercise in a variety of contexts, whether or not one does so at any particular time. And so we remain in need of an explanation of how we ever have episodic commissive self-knowledge. Things get worse if one goes so far as to argue that *all* self-knowledge must be episodic, such that all self-knowledge always requires *actively* avowing one's mental state in an occurrent "doxastic episode" (Bar-On, ms.).

One partial reply involves challenging what is required for episodic self-knowledge. Thus, it may be argued that commissive self-knowledge need not require actively avowing one's commitments. For instance, Shoemaker argues that one has constitutive self-knowledge when one's attitudes are *access conscious*. Access conscious attitudes are those that one draws on in occurrent mental processes (e.g., reasoning) whether this involves actively self-ascribing them or not. Access-conscious attitudes are, in turn, *available* "in the sense that their subjects are poised to assent to their contents, to use them as premises in reasoning, and to be guided by them in their behaviour" (2009, p. 35). So perhaps CF can contend that all of our commitments are self-known in C-Conditions insofar as they are access conscious. One qualification I would like to add is that one's commitments be access conscious in the sense that they are available to distinctively second-order cognitive processes, such as those explored in §4.6. The reason for

this is that access-consciousness can otherwise be understood as sufficient for a kind of *first-order* awareness, one that might not suffice for making strong claims about self-knowledge.

However, there seem to be problem cases for this response. Following Tyler Burge (2007a, p. 386; 2007b), Daniel Stoljar offers the following case:

> …after thinking about philosophy for some time, your thoughts are completely crowded out by images of "a rainy night in Salisbury". At the beginning of the process, your thoughts about philosophy are access conscious; after all, you are drawing conclusions from them, weighing up their relative plausibility and so on, and these are aspects of the sort of functional role that is at issue in cases of access consciousness. At the end of the process, however, they fail to be conscious in any sense, and so fail to be access conscious, since you have become completely absorbed by something else. And yet it may remain possible that the thoughts are poised [to play a functional role]; for example, perhaps hearing the whispered name of a prominent philosopher is all it would take to have them come flooding back. (2019, p. 2071)

If Stoljar is right about the case, then it can happen that one's commitments are available (in Shoemaker's sense) but not access conscious. This means that not all of our commitments are always access conscious, even in cases like the above where it does not seem like any plausible C-condition has been violated, and so we might conclude that our commitments are not always episodically self-known in C-conditions. In that case, we need a further account of what enables us to transition from merely standing to episodic-qua-access-conscious commissive self-knowledge, in C-conditions. Not only this, but depending on one's conception of episodicity, we may still need an account of what enables us to actively avow our commitments in C-conditions, over and above an account of what enables us to have episodic qua access-conscious attitudes. This will be an account of how we transition from either standing self-knowledge *or* episodic-qua-access-conscious commissive self-knowledge to episodic avowals of our commitments.

Stoljar thinks he can solve one of these problems by bringing *attention* into the mix. Thus, as he argues, our standing attitudes will be access conscious and hence episodically self-known (in one sense of episodically self-known) if they are: available (such that, on my terminology,

they are self-known in a standing way), *and* "if one attends to a sufficiently great extent to their content" (Ibid., p. 2080). However, one may still have doubts about what explains movements from non-avowed access-conscious commitments to avowed access-conscious commitments, in C-conditions. Here, I contend that we might reasonably make the following amendment to Stoljar's account: a commitment will be both access-conscious *and* actively avowed in episodic thought when (1) one attends to the content of a commitment while (2) engaging in a second-order thought process that triggers one's disposition, constitutive of having a commitment in C-conditions, to actively avow one's commitment. These might be trains of thought that my Social Agentialist account carves out (§4.6), or that a Self-Regulative Agentialist account carves out (§4.7), or something else entirely.

I contend, then, that Constitutivists can account for the episodicity of self-knowledge, when it *is* episodic, by appealing to (and, to be fair, elaborating on) a C-condition that was already in place for CF, namely, the C-condition of being sufficiently first-order attentive. It is still true, though, that our commitments are not all episodically self-known all at the same time. Still, we can account for episodic commissive self-knowledge (in the access-conscious sense *and* the active avowing sense) without countenancing an epistemic method that yields such self-knowledge, thereby respecting the central Constitutivist idea.

§5.7.2—Constitutivism and the Regress Objection

I will now consider whether CF is vulnerable to a vicious regress—a regress that threatened our first formulation of Constitutivism, CC (§5.6.1). To reiterate: if self-knowledge is constitutive of our mental states, and self-knowledge involves self-beliefs that are themselves mental states, then our self-beliefs must also be self-known, and so on *ad infinitum*. Note that, for this regress objection to matter for CF, it must threaten commissive self-knowledge specifically.

Existing responses to the regress objection are sometimes concessionary. For example, in response to Richard Greene (2003)—the first author, as far as I know, to raise the regress objection for Constitutivism—Tom Stoneham (2003) argues that we can block the regress by simply forfeiting the "self-intimation" direction of the Constitutivist accounts like CF—the direction, roughly, such that "if you believe [desire, intend] that *p*, then you believe that you believe [desire, intend] that *p*"—while holding onto its other direction, namely, that "if you are in the state of believing that you believe [desire, intend] that *p* then you do believe [desire, intend] that *p*" (2003, p. 151). This concession straightforwardly blocks the regress, because it is never claimed that our mental states are, in an infinite ascension, constituted by higher-order self-beliefs about them.[228]

The sacrifice made by Stoneham here would be disastrous for my purposes, however, since my aim in this chapter has been to make sense of how we have access to our commitments, rather than to make sense of what assures the privileged status of our self-knowledge when we happen to have it. Stoneham acknowledges that his reply is highly concessionary, though only to say that we should separate the task of explaining the epistemology of self-knowledge (which, he thinks, is accounted for by preserving the other direction of the constitutive biconditional) from its aetiology (Stoneham 2003, p. 154). But I think that we can do better than this; I do not believe that we need to reject the self-intimation conditional of CF in the first place.

One strategy brings us back around to the C-conditions of CF. For recall that, on this thesis, one C-condition is that one has the psychological concepts required for having self-beliefs about one's intentional attitudes. But it is also worth noting that, while concept possession requires *competence* with a concept, such competence is often if not always finite. Thus, one

[228] A similar strategy is taken by Shah & Vavova (2014, pp. 635-636) in response to a different sort of regress discussed by Kornblith (2012).

way to block the regress might be to say that the regress ends wherever our competence with respect to our attitude concepts does. How far does this regress go, then? In Bilgrami's view, the answer seems to be "not very far" (2006, p. 115). This strikes me as plausible as well, seeing as I hardly understand what it is to self-ascribe a fourth-order self-belief, let alone a three-thousandth-order one. My competence with the concepts of belief, desire, and intention does not extend this far.

To better see why this isn't a kind of cheat response, consider a common objection to doxastic infinitism.[229] Doxastic infinitism contends that doxastic justification depends on my having an infinitely long chain of further beliefs-as-justifiers. The problem with saying that this chain need only extend as far as I am psychologically able to entertain it is that the objector will simply reply by saying that my belief goes unjustified as soon as we reach the last belief in the chain (unless, of course, the infinitist now pivots to foundationalism, coherentism, or whatever). The objector is dialectically entitled to this reply because, as the infinitist herself argues, the chain of justification collapses if it is not constituted by infinite doxastic states. On my reply to Greene's regress for Constitutivism, however, there is no obvious problem with simply allowing that the regress ends where my psychological capacities do. This is because nothing about CF suggests that cutting three-thousandth-order self-beliefs from my psychology impugns the warrant I have for my lower-order self-beliefs.[230]

For those unpersuaded, however, a final reply strikes me as the strongest of all (and shows, I think, that Bilgrami has missed out on the simplest answer to the regress). As we have seen, Greene's regress objection purports to get a grip on self-beliefs as well as ordinary, first-order

[229] See Klein (2009) for discussion.
[230] That is, if my self-belief is warranted by its *lower-order* object, or by a transcendental entitlement, then it is not warranted by my three-thousandth order self-belief, and so the excision of such a self-belief from my psychology has no worrisome epistemic consequences for my lower-order self-beliefs.

states, or at least those that agents possess in C-conditions. But it is important to remember that Constitutivists can—as I do—restrict their view to propositional attitudes as commitments. With this in mind, the regress only obtains if self-beliefs about our commitments are themselves commitments. But are they? The answer is that they are not. Rather, on my view, such self-beliefs are just *related in a special way* to commitments: a constitutive way. Commitments—as both Coliva (2016) and I understand them—have certain essential properties, such as being brought about by judgements based on assessments of evidence (see §4.4.1). Our self-beliefs, however, are not brought about by judgements based on assessments of evidence. Rather, our self-beliefs are evidentially groundless. Indeed, this is a crucial part of what gives rise to common philosophical puzzlement about their provenance and warrant. We do not need to conceive of our self-beliefs as themselves commitments, and so the regress never launches in the first place.[231]

§5.7.3—Constitutivism and Self-Deception

In the previous subsection we considered an objection to the self-intimation direction of the CF biconditional. In this subsection we consider a—really, *the*—objection for the "authority" direction of the CF, i.e., the conditional according to which self-believing that one has an attitude constitutes and hence entails having that attitude as a commitment (given C-conditions). The objection in question concerns the possibility of self-deception.

Self-deception is a seemingly common phenomenon. A man's brow tenses as he drops his plate of food, only to then proclaim to his friend: "I'm not angry!" He is self-deceived: he seems

[231] Bilgrami takes any beliefs that rationalize actions to be commitments, since he does not constrain his conception of commitments to judgement-based attitudes. Thus, he concedes that "[p]erhaps clever people will devise intelligible examples of how beliefs of higher and higher order may rationalize actions…So there is no need to be embarrassed by the climb to higher orders *if* the rationalizing and the agential element is intelligibly present in each case" (2006, p. 115), whereas I can block the regress immediately at the second-order level.

to have a false self-belief. A wife constantly finds herself checking her husband's phone for signs of infidelity, and yet avows: "I know he would never do anything to hurt me!". She is self-deceived: she cannot acknowledge her insecurity. If this is right, then one might think that the Constitutivist must show that self-deception is impossible in C-conditions. But how plausible is this move? Can't we be self-deceived even if we are attentive, alert, rational, and conceptually competent? It seems that we can. Self-deception can get the better of highly functioning individuals. Moreover, simply introducing a *no-self-deception* C-condition seems ad hoc (Coliva 2016, *pace* Heal 2002).

To get a sense for the Constitutivist's option space, we should look at a case of self-deception. I take the following case, briefly alluded to above, from Coliva (2019a):

> Jane is married to Jim. They have been married for several years and have a daughter. Jane is often at home, on her own, attending to domestic chores. From time to time, she feels lonely and wishes that she had pursued her own career. More often than not, however, she feels much rewarded by the fact that her family is so serene. Indeed, when she meets with her friends, who sometimes complain about their husbands, she cannot help remarking that her life makes her happy and that her husband is adorable and completely trustworthy. Still, it often happens that, while preparing the laundry, Jane carefully searches Jim's pockets. While tidying up his study, she opens and examines all the drawers. While dusting the furniture, she lingers on the screen of his laptop, left open on the incoming messages. One day, Jane hears about Freud's theories concerning the unconscious. Little by little, the deep significance of a whole series of previously meaningless actions is disclosed to her. Ashamedly, she realizes that all that attention spent over the content of her husband's pockets was a sign of her being insecure about him. All that dusting the screen of his laptop, a symptom of her thinking that he might have some intimate correspondence with another woman. Still, Jane knew all too well that Jim had always been the most truthful of men. The thought popped into her head: "I believe that Jim is unfaithful to me, although he is not."

It is worth working with so lengthy a case because it includes two crucial psychological stages in Jane's life. We can describe Jane at the first stage—where she remarks that her husband makes her happy—as the Jane who is outright self-deceived. We can describe Jane at the second stage as the Jane that has become aware of her self-deception and issues an utterance that we might

ordinarily describe as Moore-paradoxical. While the core phenomenon is obviously located at the first stage (and so I will focus mostly on this first), paying attention to the second stage will also bear some fruit in just a moment.

It is true, on the standard view of self-deception, that Jane is simply wrong about herself. She does not really express her first-order belief when she remarks that her husband makes her happy. She deceives herself into believing that she believes this, but the deception ensures the falsity of her self-belief. "Given her overall behavior," the standard view tells us that "she should be taken to believe only the opposite [of what she self-believes], albeit unconsciously" (Ibid., p. 3). However, Coliva denies that the standard view gives us the best account of Jane's situation. Rather, following Bilgrami (2006a, 2012), Coliva argues that Jane's self-belief is *true*.

The key to understanding how this could be depends, once again, on distinguishing between commitments and dispositions. Once we appreciate this distinction, the following account is available: Jane really does express her first-order *commissive* belief that her husband makes her happy when she avows this to her friends. This is her belief that is responsive to the evidence about the serenity of her homelife. What renders her self-deceived is that she *also* has a countervailing *dispositional* belief, one that tacitly guides her behaviour, often without her awareness.[232,233] If this is right, then self-deception is no counter-example to the authority direction of the Constitutivist bi-conditional after all.[234] Moreover, when she becomes aware of her self-deception, we can understand why her eventual admission—"I believe that Jim is

[232] I concur with Shoemaker (2009) that there is a related issue in many cases of self-deception: if someone discovers that she has a dispositional belief, she may remain ignorant of the fact that her dispositional belief has a stronger pull on her behaviour than her commissive belief does.
[233] This picture differs from Schwitzgebel's (2010), who does not distinguish commitments from dispositions and, instead, understands the beliefs expressed by avowals as 'in-between' or 'indeterminate' beliefs that conflict with other, more determinate beliefs.
[234] Though Davidson (1986) does not invoke a commitment/disposition distinction, it is possible to understand his account of self-deception as at least compatible with Bilgrami's and Coliva's account, since for him self-deception involves a "partitioning" of the mind, wherein one does not realize that one harbors conflicting attitudes.

unfaithful to me, although he is not"—does not seem to be Moore-paradoxical (see §4.4.2). The commitment/disposition distinction is what bears fruit here: it explains why we do not take her utterance to be Moore-paradoxical. For, even though she expresses formally contradictory attitudes as regards their contents, only one is her commitment and so the clash is not one of straightforward self-contradiction. The intuition is that her situation is more subtle than this, and the commitment/disposition distinction explains this while preserving the truth of her avowal.

Coliva provides additional incentive to accept this account of self-deception:

> To deny that Jane really believes that her husband is faithful to her, despite her saying so, is tantamount to denying either that sincere assertion expresses belief, or that Jane's assertion is sincere. The former horn of this dilemma is difficult to maintain. For if an assertion is sincere, it may not be true, or it may not be justified by one's available evidence, but it would seem to go against the grain to deny that it expresses one's belief. The latter horn is thus *prima facie* more plausible. Yet there is no obvious reason to think that cases of self-deception involve some kind of insincerity on the subject's part. Insincerity requires saying something one believes to be false and doing so intentionally, in order to get someone else to believe it. Yet, how could a subject deploy this complex plan towards herself, to have herself believe something she thinks to be false and do so intentionally? (Ibid., p. 3)

I agree with Coliva that the second horn of this dilemma for the standard view is more plausible prima facie. But Coliva's subsequent dismissal of it may be premature. This is because it might be false that insincerity "requires saying something one believes to be false and doing so intentionally." This may be the standard account of insincerity, but perhaps there are alternatives worth exploring.

To see where I am going with this, consider Kevin Falvey's claim that self-deception does involve some degree of insincerity (2000, pp. 89-91), albeit insincerity that is subtler than outright lying. He offers the following case by way of example:

> Sam is a young professor of philosophy attending a reception for the new dean of his college, who is from the comparative literature department. A small group of faculty from various departments are extolling the virtues of Derrida's work, and bemoaning the fact that analytic philosophers do not appreciate him. When the dean asks for Sam's

> opinion, he says, "I think analytic philosophers have been a bit too hard on Derrida," whereupon Sam's friend and colleague Jane rolls her eyes and mutters under her breath, "You don't really believe that!" When Jane taunts Sam about his remark afterward, he initially tries to defend himself, reiterating that he does think that many philosophers have been a *bit* too hard on Derrida's work, but he eventually admits that Jane is right. Sam was merely going along with the crowd, trying to ingratiate himself to the new dean, and he feels duly ashamed. (2000, p. 90)

What is going on here? On Falvey's view, Jane (it's always Jane) takes Sam's avowal to be false, because she has never seen Sam defend Derrida before. Sam's subsequent "defense" of his avowal is really just a more qualified version of it. After that too is challenged, he immediately capitulates: his avowal is, by his own lights, a false one. *Ex hypothesi*, however, Sam did not outright lie. It is just that he was blinded to the pretenses behind his avowal by his desire to please those around him.

Now, one might reasonably doubt Falvey's own take here, since the ease with which Sam gives up in his exchange with Jane might just as easily indicate that he is, to some degree, aware of his own insincerity from the jump. After all, Jane does not need to engage in any sophisticated psychologizing to convince him; she merely challenges his sincerity and, after a very brief back and forth, one in which he does not so much as offer a single justification for his claim, he concedes. This provides some evidence that Sam's avowal does not express a sincere second-order belief in the first place. But if this is right, then it cannot be a *false* self-belief that, therefore, undermines CF.

It might help if Falvey had an account of the pretensive character of Sam's avowal. The case is supposed to speak for itself, but it doesn't seem to. Fortunately, Tamar Gendler has offered an account of pretense and its role in self-deception (2007). She understands pretenses as a kind of *imagination* or *make-believe* that *p*. On this account, pretenses involve a kind of representation of *p*, and can motivate action in the way that beliefs do, but they fall short of

beliefs or other doxastic attitudes toward *p*. They are, as Xintong Wei puts it, matters of "pretending the world in a *p*-like way" (2020, p. 2).

Unfortunately, while such pretensive attitudes may exist, taking them to be the states that an agent expresses in her self-deceived avowals is highly problematic. First, if pretenses are not doxastic attitudes, then they are not truth-aiming (as Gendler herself contends). But this seems wrong in cases of self-deception: self-deceived agents seem to avert their gazes from countervailing evidence for their attitudes, and they tend to appeal to evidence in favour of them (Wei 2020, p. 4). Moreover, if pretenses are not doxastic states, then there can be no reasonable epistemic demand for the self-deceived agent to revise her self-deceived state (Ibid., p. 5). But it seems like we do (reasonably) demand this of the self-deceived. Finally, pretenses ordinarily have restricted motivational effects; somebody pretending to be a doctor won't go so far as to actually prescribe a treatment to a friend (Ibid., p. 5). This is so even if we think about *highly immersed* pretenses: even an actor heavily engrossed in his role as a doctor will drop the pretense before actually prescribing medicine (Ibid., p. 6). But self-deceived agents can often remain self-deceived well beyond the motivational perimeter of a pretense.

Wei concludes from this that states of self-deception are not constituted by pretenses, though she goes on to suggest that pretenses can play an important role in the *process* of becoming self-deceived. The picture is, roughly, as follows. In adopting the imaginative pretense that *p*, one will go on to act in ways commensurate with believing *p*, at least to some extent. After a certain point, one may begin to experience various psychological benefits. By pretending one's book is excellent, for example, one might experience pride that one would not otherwise experience in clear-mindedly acknowledging the book's poor sales. This can create a feedback loop that motivates one to deepen the pretense, perhaps by immersing oneself in a social

environment comprised of a bunch of yes-men who call one's book a masterpiece. At this point, one's evidence regarding one's belief about the book's quality will have changed, since one will now be receiving positive feedback from one's peers. At this point, one can genuinely come to believe that the book is good.[235]

This may be a fruitful account of self-deception. But notice that, on this account, one's self-deception consists in having come to change one's *first-order* belief about the goodness of the book after engaging in pretenses to the effect that it was good. This means that, if one self-ascribes this first-order belief, one is saying something *true*. The self-deception does not consist in one's coming to have a false self-belief, but rather in the fact that one has altered one's first-order beliefs by engaging in motivated reasoning, all without being aware that this is what she has done.[236] The fact that she is not aware of the role that pretense has played in shaping her first-order attitude is not a problem for CF, because her pretenses are not her commitments and so her lacking self-knowledge of them is not a counter-example to CF.

Is Falvey's case of Sam amenable to this model? It might be. Perhaps, to get himself to go along with the crowd, he engages in an immersive, motivated process of pretending that Derrida's work has been unduly dismissed by analytic philosophers. He might call to mind imagery of Derrida being slandered in similar meetings, and he might imagine Derrida despairing over his mistreatment. If this is right, then Sam's pretense can serve as a precursor to the formation of a genuine first-order belief that he, in turn, *truly* self-ascribes. Of course, in

[235] The pretense account differs from other, seemingly similar accounts. For example, McGeer & Pettit argue that: "the self-regulating mind can intentionally seek to discover what truth-related constraints require, so a desire to maintain certain beliefs can lead it intentionally not to pursue such questions" (2002, p. 296), and Davidson (1982, 1986) argues that self-deception involves an intention to self-deceive. These accounts tend to emphasize the role of motivated reasoning in the process of self-deception. But because pretenses are not doxastic states, I doubt Wei would endorse this exact treatment of the process of self-deception.

[236] Perhaps such reasoning is sub-personal, or personal yet second-tier (see §3.4). Or perhaps the agent is aware of her reasoning but lacks knowledge of its being motivated by her subterranean attitudes.

Wei's example, the agent who comes to believe that her book is good acquires genuine (albeit misleading) evidence from her yes-men peers, and so she forms a first-order attitude that we might rightly think of as a *commitment* on this basis. To find a strict parallel with Sam's case, we will have to imagine that his imaginative pretenses can lead him to acquire *evidence* for his belief. I see no reason why we should not read Sam's case this way, for we can imagine his imaginative pretenses as leading him to take seriously the testimony of those in the meeting who claim that Derrida has been mistreated. In this way, he draws on genuine testimonial evidence in forming his belief, though perhaps he may not have taken such evidence as sufficient if he had not engaged in any pretensive preliminaries.

So, we have the following structure for the case of Sam. Sam has a desire (we can say, a merely dispositional desire of which he is unaware) to go along with the crowd. He encounters testimony about Derrida's mistreatment. He then engages in a series of pretensive imaginings that prime him to become more receptive to this testimony. He then forms a first-order commitment to the effect that Derrida's work has been treated unfairly and, for this reason, *truly* avows it. Jane rolls her eyes, taking him to be insincere. But she does this only because she lacks access to the role that Sam's pretensive imaginings played in leading up to his response. These pretenses lead him to take the testimony of the others more seriously than he otherwise would have (and they may also cause him to screen out considerations of countervailing reasons). This also explains why Sam's replies to Jane are so weak. His only real evidence for the truth of his claim is that his colleagues have made the same claim. He cannot provide any independent reasons for thinking that Derrida's work has been treated unfairly, and it would be embarrassing to simply admit that the testimony of these other professors is his reason for believing what he does. His commitment stands on flimsy ground, and it quickly crumbles under Jane's scrutiny.

Nevertheless, it *is* his commitment that crumbles, and so his initial avowal of it was true, not false.

Notice that this take on Sam's case is not identical to Coliva's take on Jane (the wife, not Sam's colleague). In that case, Jane truly avows her belief in her husband's faithfulness, although she is self-deceived in virtue simultaneously harbouring an incompatible, subconscious dispositional belief. In Sam's case, he does not (necessarily) harbor an incompatible dispositional belief that Derrida's work is bad when he avows that it is good. Rather, he simply forms a commitment on flimsy grounds and lacks awareness of some of the pretensive mental activity that got him there. So it seems that there are two ways, not one, to accommodate self-deception along Constitutivist lines. Either one is self-deceived when one's self-belief about one's commitment is true despite one's harboring an incompatible dispositional attitude unawares, or one's self-belief about one's commitment is true and one is self-deceived insofar as one is unaware of the pretenses that influenced one's commitment.[237]

§5.7.4—Stage Setting: Constitutivism and the Indistinct Existence Thesis

In the next several subsections I turn to a final set of objections for my Constitutivist account. These objections are each offered in the spirit of rendering Constitutivism comparatively less plausible than an alternative account that has recently been developed by Matthew Parrott (2017). As we will see, Parrott's preferred account is supposed to do better than Constitutivism at accounting for the possibilities of self-ignorance and self-error, and at accommodating further plausible metaphysical and epistemological phenomena. My aim, in addressing all of this, will be to show that Constitutivism remains more plausible than this alternative.

[237] Note that the Constitutivist account of self-deception advanced here is to be distinguished from Bagnoli's (2012) self-proclaimed Constitutivist account of self-deception, according to which we are *responsible* for our self-deception because we ourselves bring about (constitute) our self-deception.

In denying Constitutivism, Matthew Parrott accepts:

> **Distinct Existence Thesis**: For any subject *a* and psychological state *M*: (i) it is not the case that part of what it is for *a* to be in *M* is for *a* to believe (first-personally) that *a* is in *M*, and (ii) it is not the case that part of what it is for *a* to believe (first-personally) that *a* is in *M* is for *a* to be in *M*. (2017, pp. 2-3)

Now, because Constitutivists *deny* Distinct Existence Thesis, and because Parrott focuses his efforts *against* Constitutivism, it will be useful to explicitly formulate Parrott's critical target, which is just the inverse of Distinct Existence Thesis, namely:

> **Indistinct Existence Thesis**: For any subject *a* and psychological state *M*: (i) part of what it is for *a* to be in *M* is for *a* to believe (first-personally) that *a* is in *M*, and (ii) part of what it is for *a* to believe (first-personally) that *a* is in *M* is for *a* to be in *M*.

Something like IET indeed characterizes Constitutivist views. Note, however, that while IET talks of 'first-personal' self-knowledge, I will continue to talk of privileged and peculiar self-knowledge. Nothing will hinge, for my purposes, on this alteration.

§5.7.5—Constitutivism Versus Rational Fundamentalism

Parrott offers three reasons to be skeptical of IET and, hence, of Constitutivism. First, he refers us to what is sometimes called Hume's Dictum (or Hume's Doctrine). Hume's Dictum states that, for any entities that are distinguishable in thought, "they may exist separately, and have no need of anything else to support their existence".[238] Constitutivists seem to be committed to denying this. And yet, to many philosophers, Hume's Dictum is "tautological" (Ayer 1956) or "plausibly analytic" (Stoljar 2007, p. 266).

A second objection concerns what Parrott sees as a crucial motivation for Constitutivism, namely, its ability to accommodate the impossibility of self-blindness. Parrott agrees with

[238] Hume (1741, 1.4.5.5.). Of course, Hume's foremost application of this doctrine was causal relations, while the mental states that Constitutivists take to be necessarily related are not typically viewed as causal relata (though, as we saw, Bilgrami seems strangely inclined to say that they are, even while denying that these causal relations are *explanatory* of self-knowledge).

Shoemaker that rational, psychologically well-functioning, conceptually equipped agents cannot be self-blind. Contra Shoemaker, however, Parrott does not think that Constitutivism follows as a result. For, while Shoemaker takes the impossibility of self-blindness to suggest that there must be something about the nature of our *mental states* that renders them self-intimating, Parrott takes the impossibility of self-blindness to establish something about *rational agents*, not about the mental states of rational agents.

Parrott puts this in the form of the following thesis:

> **First-Personal Dispositions**: Necessarily, for any rational subject *a* and psychological attitude *M*, if *M(a)*, then *a* is disposed to believe [first-personally] that *M(a)*. (2017, p. 15).

Parrott takes First-Personal Dispositions (FPD) to follow from the impossibility of self-blindness. It is also the core of his non-Constitutivist account of self-knowledge, an account that I will refer to as *Rational Fundamentalism*. I call this account Rational Fundamentalism because its core thesis, FPD, posits a disposition, fundamental to rational agency, the triggering of which suffices for possessing self-knowledge of one's mental states.[239] Self-knowledge is privileged and peculiar, according to Rational Fundamentalism, because nobody else is disposed to form beliefs about one's mind simply in virtue of being rational agents, and it is privileged because this disposition will deliver self-knowledge whenever it is not masked, whereas no knowledge of other minds is so reliable. Unlike IET, however, FPD only entails that we be *disposed* to form higher-order self-beliefs about our first-order attitudes. There is nothing in this thesis about mental states themselves (commissive or otherwise) that *necessitates* self-knowledge of them. In other words, these attitudes are ontologically distinct.

[239] This is Parrott's conception: he is explicit that we are not to understand FPD as positing a unique disposition to believe (first-personally) that *M(a)* for every psychological attitude M.

FPD is said to entail the impossibility of self-blindness because it ensures that rational agents have self-knowledge without the possibility of an intervening causal mechanism that could break down. Moreover, if an agent is never in good enough cognitive circumstances for the relevant disposition to trigger (non-deviantly, we might add), we can reasonably wonder whether she is a rational agent.[240] If she is not, then she might be self-blind. But even Shoemaker argues that self-blindness is impossible only for rational agents (with the necessary conceptual repertoire). If all of this is right, Rational Fundamentalism undercuts a key dialectical advantage of Constitutivism, for it now follows that Constitutivists are not the only ones capable of ruling out self-blindness. This result gets worse when paired with Parrott's first criticism of Constitutivism, since Rational Fundamentalists can respect Hume's Dictum and rule out self-blindness, whereas it seems that Constitutivists can only do the latter.

We will have more to say about Rational Fundamentalism and its central thesis, FPD. For now, we should get Parrott's third objection to IET on the table. The criticism is that, if our self-beliefs metaphysically constitute their objects, it is hard to see how self-ignorance and self-error are possible. Contrariwise, FPD leaves open the possibility that, due to fatigue or distraction, one's fundamental self-knowledge-delivering disposition can be masked, leading to self-ignorance (such that one lacks a self-belief). Alternatively, though Parrott does not suggest this himself, it may be that one's disposition can also be triggered in deviant ways, resulting in self-*error* (such that one has a mistaken self-belief). In the next several subsections I will address these and other objections to IET and, hence, Constitutivism.

[240] See Coliva's (2016) notion of "thick" rationality for a similar suggestion.

§5.7.6—Rational Fundamentalist Objections to Constitutivism: Hume's Dictum and Self-Blindness

We have already seen how Constitutivists try to capture self-ignorance and self-error; they do so by admitting that these are possible outside of C-conditions. The idea, then, is only that there are *conditionally* necessary connections between the mental states covered by a given Constitutivist account. I will consider Parrott's problem with this move in a moment. For now, I want to point out that it is not enough to honor Hume's Dictum. This is because Hume's Dictum has the consequence that, if one's mental states can exist independently at any possible worlds, then they never necessarily exist together. As a proponent of Hume's Dictum, Parrott is therefore skeptical about the very possibility of conditional necessities. He writes:

> If there are some conditions in which an X can fail to stand in a relation to Y, then it seems like the two are not really necessarily connected. They may well stand in that relation in every world in which some further condition C holds, but we might question whether that is sufficient for necessity…it might still be true that in ordinary psychological conditions, if one *a* is in *M* then *a* will believe that *M*(*a*) (or vice versa), but since this isn't necessary it is compatible with the *Distinct Existence Thesis* [i.e., ~IET]. (2017, p. 19, fn. 35)

Notice, though, that anyone who takes self-blindness to be impossible should have a problem here. This is because saying that there are no conditions that necessitate self-knowledge is tantamount to saying that, for any agent *A* at time *t* in conditions *C*, it is possible for her not to know her mind first-personally (i.e., with privilege and peculiarity). But this is just a way of saying that self-blindness *is* possible. In other words, no matter how good one's cognitive conditions are, no matter how rational one is, no matter how attentive, lucid, and conceptually sophisticated, these conditions can never *entail* that one has privileged and peculiar self-knowledge. So, Parrott seems forced to accept:

> **Unavoidable Failures**: For any agent *A* at time *t* in conditions *C*, it is possible for her fundamental disposition to self-knowledge to be masked or to be deviantly triggered.

My point now is that we cannot both deny the possibility of self-blindness and accept Hume's Dictum, for we can specify no principled limitations on the ubiquity of self-ignorance or self-error if we are Rational Fundamentalists: an unavoidable instance of self-error or self-ignorance is, after all, an instance that can occur in optimal psychological/rational conditions. For this reason, Rational Fundamentalism begins to appear like a version of the Inner Scanner Account, albeit while supplanting a dedicated causal mechanism in the brain for a functionally identical disposition that supervenes on our rational agency.

I take it that this is a bad result for Rational Fundamentalism. It means that Constitutivism is, after all, dialectically advantaged in one respect, since it does not accept Unavoidable Failures. This alone does not disprove Hume's Dictum. But because it has also struck many philosophers that the impossibility of self-blindness is analytic (cf. Davidson 1995, p. 234; Shoemaker 1990, p. 50; Coliva 2016, Chapter 3), the mere fact that other philosophers feel similarly about Hume's Dictum does not put Constitutivists at an automatic disadvantage.[241]

I take these points to provide responses to two of Parrott's objections: the objection from Hume's Dictum and the objection that Constitutivism is not better positioned to accommodate the impossibility of self-blindness than Rational Fundamentalism. In the next subsection I say more about Constitutivism's prospects for explaining self-ignorance and self-error.

§5.7.7—Rational Fundamentalist Objections to Constitutivism: Self-Ignorance and Error

As we have seen, the Constitutivist's view is that there can be worlds at which agents fail to have constitutive self-knowledge because they are outside of C-conditions. Perhaps we can think

[241] See, e.g., Shoemaker (1996), Setiya (2011) and Coliva (2016). Moreover, Wilson (2010) shows that many arguments for Hume's dictum fail to establish it, despite the fact that Parrott cites her article approvingly. She does point to a few defenses of it near the end of the article, but these are only developed in a highly cursory way. My take-away from Wilson's article is that, even if defenses of Hume's dictum can be developed, it is not tautological.

about worlds at which an agent is heavily fatigued, distracted, or conceptually unsophisticated. In these circumstances, we might take the dual-order mereological complexes (that normally constitute our commitments) to be destabilized or destroyed.

Parrott worries, however, that "the Constitutivist is forced to turn to comparatively more complicated explanations for instances of self-error or ignorance" (2017, p. 19) than the Rational Fundamentalist:

> For example, [if one is a Rational Fundamentalist] one might pursue the promising idea that we sometimes make mistakes about our beliefs because distraction or fatigue masks our standing disposition (*First-Personal Dispositions*) to form higher-order beliefs about them. This is a straightforward causal explanation, but it would some lose credibility if we were to accept a constitutive theorist's picture about how a subject's psychological attitudes are necessarily connected to her higher-order beliefs about them. (2017, pp. 19-20)

I confess that I am simply unsure why this causal explanation is so much better than the Constitutivist alternative. There are worlds at which certain conditions are not ripe for mereological relations to obtain, and there are worlds at which one's fundamental disposition to self-knowledge might be masked. If the Constitutivist has to deny that her story is causal, it is not obvious that it is comparatively more complicated. At any rate, even if one thinks that introducing a causal element produces a comparatively straightforward explanation of self-ignorance and self-error, it comes at the expense of ontological parsimony. After all, it is only Rational Fundamentalists who have to countenance a fundamental self-knowledge-producing disposition, since Constitutivists simply embrace a metaphysical account of the relationship between our commitments and self-beliefs about them. Given that this is so, it is not obvious which account scores better in the game of theoretical virtue counting.

§5.7.8—Rational Fundamentalist Objections to Constitutivism: Anti-Luminosity

There is one further concern about self-error, noted by Parrott, that may seem harder to accommodate along Constitutivist lines than Rational Fundamentalist lines. This is Timothy Williamson's well-known *anti-luminosity* argument. The argument is intended to establish that we human beings have no "cognitive home" (2000, p. 93), which is to say that no cognizable domain of facts is "inherently accessible" to us despite our *de facto* success as cognizers in many domains. It is because our own minds are so often taken to be especially epistemically accessible, as per common theses about privileged access, that Williamson focuses his anti-luminosity argument on self-knowledge in particular.

Williamson's anti-luminosity argument has been variously interpreted. On what is perhaps its most common interpretation, the argument has a "sorites-like" structure (Berker 2008, p. 1).[242] The key premise of the anti-luminosity argument is:

> **Luminosity**: For every case α, if in α condition c obtains, then in α one is in a position to know that c obtains.

Luminosity should readily remind us of the self-intimation conditional of Constitutivist theses. There is a difference, seeing as Luminosity emphases 'being in a position to know' rather than outright self-believing or self-knowing. Still, this is close enough to the Constitutivist's view to be worth considering, since its falsity would surely entail the falsity of the stronger Constitutivist thesis that one *knows* c when in α.

Williamson's test case for a luminous (i.e., self-intimating) condition is *being cold*. Being cold paradigmatically has a highly salient phenomenology. So Williamson thinks that, if being cold is not luminous, it is highly unlikely that anything else is. This is why it is instructive to consider this case, despite the fact that my Constitutivist thesis CF does not make any

[242] Though see McGlynn (2014, pp. 148-149) for reservations.

luminosity/self-intimation claims about sensory states like feeling cold. Another reason why it will be helpful to consider this case is that, as we will see, standard strategies for defending the luminosity of cold cannot be repurposed to defend the luminosity of propositional attitudes like commitments.

Here is Williamson's case, from which he eventually derives his anti-luminosity conclusion:

> Consider a morning on which one feels freezing cold at dawn, very slowly warms up, and feels hot by noon. One changes from feeling cold to not feeling cold, and from being in a position to know that one feels cold to not being in a position to know that one feels cold. If the condition that one feels cold is luminous, these changes are exactly simultaneous. Suppose that one's feelings of heat and cold change so slowly during this process that one is not aware of any change in them over one millisecond. Suppose also that throughout the process one thoroughly considers how cold or hot one feels. One's confidence that one feels cold gradually decreases. One's initial answers to the question 'Do you feel cold?' are firmly positive; then hesitations and qualifications creep in, until one gives neutral answers such as 'It's hard to say'; then one begins to dissent, with gradually decreasing hesitations and qualifications; one's final answers are firmly negative. (2000, pp. 96-97)

Say that T_1 is the crack of dawn, at which the agent—who we can call Mr. Davis (following Weatherson 2004)—is coldest, with various further time slices T_2, T_3 leading up to T_n at which Mr. Davis is warmest. All the while, Mr. Davis attends carefully to his condition, and we can stipulate that his cognitive conditions are excellent. Thus, he should be in a good position to know whether he is cold or not at a given time. As Wolfgang Barz puts it, if phenomenal states like being cold are luminous conditions, then they are "waterproofed against epistemic luck, at least if they are based on careful introspection" (2017, p. 482).

Now, at some interval between $T_1 \ldots T_n$ there will be a 'borderline case', T_i, at which Mr. Davis is cold and then finally transitions to not being cold at T_{i+1}. But here is where Williamson thinks we have reason for concern. For, at this borderline case T_i, even granting that Mr. Davis

has a true self-belief about his condition, it seems that he must lack self-knowledge. The reason is that Mr. Davis' self-belief is *unsafe*. To see this, consider:

> **Safety**: Only safe beliefs count as knowledge, so whenever you know that C obtains, C obtains in all similar cases. (Weatherson 2004, p. 373)

Safety entails that knowing that you're in some condition means that you must know it in sufficiently similar situations—otherwise, your knowledge will *not* be "waterproofed against epistemic luck" (Barz 2017, p. 482).

The trouble comes when we consider the conjunction of Luminosity and Safety. For, while Luminosity says that Mr. Davis is in a position to know that he is cold when he is cold, Safety says that he therefore knows he is cold at the next time slice, since the next time slice is indexed to an extremely close possible world, such that it counts as a similar case to the previous time slice. But this yields an absurd result: it entails that, if you know you are cold at T_1, you know that you're cold at every subsequent time slice, because otherwise your belief at any particular time slice would not be safe. And so, because knowledge requires safety, and because you do know at T_1 that you are cold, knowing you are cold at T_1 entails that you know you are cold at all the way through to T_n where, *ex hypothesi*, you are *not* cold. To avoid this absurd result, we can either drop Safety or drop Luminosity. Williamson takes it that we should drop Luminosity.

Even if one thinks that we should drop Safety instead, it is also possible to formulate an anti-luminosity argument using a weaker-than-Safety premise. Consider, for example:

> **Safety-Weaker**: A true belief b, actually held by a subject S, is unsafe if there is an extremely similar possible situation in which S continues to form b in the same way as in the actual situation, even though b is false. (Barz 2017, p. 483)

Safety-Weaker does not entail that you *know p* at *all* sufficiently nearby worlds to T_i; it only requires that there is *no* sufficiently nearby world to T_i at which your belief is *false*. But even with this weaker safety principle on hand, the problem is that, "though, in actuality, Mr. Davis, at

t_{i+1}, no longer believes that he feels cold, he might well have continued to believe that he felt cold at t_{i+1}" (Barz 2017, p. 483). So Safety-Weaker and Luminosity are also problematic in conjunction. For every case α, if in α condition c obtains, then in α one is in a position to know that c obtains. But at T_{i+1}, one's belief that one is cold would be false (since one would not be cold). So there is a sufficiently nearby world at which one's belief would be false. So, one's belief is unsafe and fails to count as knowledge, despite what Luminosity claims.

Many opponents of Williamson's anti-luminosity argument focus on criticizing Safety, though these arguments don't address versions based on *Safety*-Weaker.[243] It is also possible to keep toying with safety principles until one is compatible with Luminosity, though these attempts have been variously criticized.[244] Yet another approach is based on David Chalmers's (2003) notion of a *direct phenomenal concept*, i.e., a concept that directly incorporates the phenomenal state it is about. Taking advantage of this notion, one can argue that Mr. Davis will not be able to utilize the direct phenomenal concept of feeling cold at T_{i+1} that he uses at T_i, because he is not cold at T_{i+1} and so cannot incorporate the state of feeling cold into the very self-belief he forms at T_{i+1}. For this reason, he will not be able to form the *false* self-belief that he is cold at T_{i+1}. This ensures that his belief at T_i is safe by the lights of both Safety or Safety-Weaker. Unfortunately, appealing to direct phenomenal concepts won't help when we turn to thinking about commitments, because no direct phenomenal concepts figure into our self-beliefs about them.

We can see, then, that the case for the luminosity of cold is difficult to make, and that one promising strategy for overcoming this difficulty has no analogue when we turn to thinking about non-phenomenal states like commitments. Now, Williamson himself never developed an

[243] See, e.g., Brueckner & Fiocco (2002).
[244] See Srinivasan (2015).

anti-luminosity argument directly addressing commitments or, really, any propositional attitudes. Again, the reason is that he took phenomenal states to be the best prima facie candidates for being luminous, such that refuting their luminosity could cast doubt on the luminosity of any other state of affairs, mental or otherwise.

Still, I do not think matters are fully settled. After all, according to Rational Fundamentalism and Constitutivism, propositional attitudes like commitments are not known via introspective attention in the way that states like being cold often are. This suggests that we might need to develop a different case through which to construct an anti-luminosity argument for propositional attitudes like commitments.

I take a case from the only author who (to my knowledge) has developed one, namely, Nico Silins (2012). Here is Silins with the initial setup:

> Consider a series of times during which Belle's confidence that *p* very gradually decreases, so that she believes that *p* at the beginning but does not believe that *p* at the end. Throughout the series Belle carefully considers whether she believes that *p*, so that whenever she is in a position to know she believes that *p*, she does know she believes that *p*. Finally, her confidence that she believes that *p* likewise gradually decreases throughout the series. (2012, p. 313)

Perhaps similar cases can be modelled for thinking about the firmness of one's intentions or the intensity of one's desires (neither property of which need be specified in phenomenal terms). Keeping with the case of belief for the sake of example, the "crucial premise" is that "if Belle knows at a given moment in the series that she believes that *p*, then she believes that *p* at the next moment in the series" (Ibid., p. 313). This is structurally isomorphic to *Safety* considered above. Silins's anti-luminosity follows from here:

> At the first moment in the series, Belle knows that she believes that *p*. By the crucial premise, it follows that, at the second moment in the series, she does believe that *p*… at the second moment in the series, she is in a position to know (introspectively) that she believes that *p*. By the setup concerning her attentiveness, at the second moment in the series she does actually know that she believes that *p*. This reasoning can be

> repeated until we reach the conclusion that Belle believes that *p* at the last time in the series. Something has gone wrong! (Ibid., p. 314)

Once again, we have an absurd result. Parrott seizes on this as providing further support for Rational Fundamentalism. The reason is that the Rational Fundamentalist's key thesis, FPD, can accommodate this anti-luminosity result, whereas Constitutivists allegedly cannot. To see this, recall FPD and Luminosity:

> **First-Personal Dispositions***:* Necessarily, for any rational subject *a* and psychological attitude *M*, (if *M*(*a*), then *a* is disposed to believe [first-personally] that *M*(*a*)). (2017, p. 15)

> **Luminosity**: For every case α, if in α condition c obtains, then in α one is in a position to know that c obtains.

The reason why FPD comports with the denial of luminosity is that *being disposed to believe* that one has a propositional attitude is not equivalent to *being in a position to know* that one has it. So, even if we reject Luminosity, FPD stands untouched. Perhaps, in the borderline cases where one's credences (i.e., epistemic "confidence", in Silins's terminology) verge on either side of the .5 threshold, one's disposition to form higher-order beliefs about her beliefs will be masked or be vulnerable to deviant triggering. On the other hand, because Constitutivism ties our first- and second-order states together, we cannot cheerfully accommodate the possibility of false self-beliefs (or a lack of self-belief) at borderline worlds. Or so the objection goes.

One reply to this objection is inspired by a suggestion from Casey Doyle. Doyle suggests that, as a general matter, Constitutivists can explain self-error and self-ignorance "as resulting from problems with *the first-order states themselves*" (2015, p. 13). On a mereological model of constitution, this suggestion might be reframed as follows: a self-belief will have a first-order state as a proper part, but if that part is itself unstable or indeterminate, this compromises the stability and determinacy of the whole dual-order complex.

Applied to the case at hand, a Constitutivist like myself could say that commitments backed by borderline credences are bound to exhibit some instability or indeterminacy. For example, if commitments are partly constituted, as I argue, by a disposition to reason in accordance with them, then one could expect the stability of this disposition to be compromised as one's credences draw nearer to the .5 threshold, for one should be less disposed to draw clear inferences from states about which one has tenuous epistemic confidence. If so, then if our self-beliefs about our commitments fail to count as knowledge in such cases, this will be because our commitments themselves are unstable. On this picture, it is true that we lack self-knowledge in borderline cases, but the explanation actually makes use of a Constitutivist account of commissive mentality.

In light of this, we might even introduce a new C-condition on CF: that one's commitments be sufficiently determinate. On this view, we can agree that one's commitments may become unstable as one's associated credences draw toward the .5 borderline, and that in such cases we lack self-knowledge. However, if we continue to think of Constitutivism as a mereological account of self-knowledge, it should be no surprise that one's self-beliefs also exhibit instabilities in these situations. Because we can reasonably expect this result, the Constitutivist might have a principled reason to amend her C-conditions. This is a principled reason, rather than an ad hoc maneuver, insofar as it is plausible that commitments, being *commitments*, must exhibit a certain degree of determinacy.

A final possibility is this. Because our commitments are partly constituted by self-beliefs, these self-beliefs might always be true even if they do not always count as knowledge. For example, at T_i, one's self-belief that one commissively desires to φ will be true, and at T_{i+1} when one no longer commissively desires to φ one's self-belief about one's new state will be true.

Luminosity is the premise that one is always in a position to *know* that some state of affairs obtains when that state of affairs obtains, but a Constitutivist might admit that self-*knowledge* is compromised in borderline cases, all while preserving the metaphysical claim that our commitments have a dual-order structure such that, in the right conditions, they automatically count as knowledge. This might be enough for Constitutivists to argue that the real core of their account has not been defeated (notice that CF does not invoke the term *knowledge* anywhere—whatever we say about the warrant our self-beliefs enjoy is not folded directly into it).

One wrinkle might obtain if, following Zimmerman (see §5.6.2), a Constitutivist continues to argue that self-beliefs are warranted simply by the first-order states they are about. This is a wrinkle because, if our self-beliefs are always true, then (on this account of warrant) they should always be warranted, and this might mean that they should always count as knowledge. This contradicts the present argument, but the contradiction can be defused if it is argued that this source of warrant—derived from the presence of the first-order state—can be defeated in borderline cases precisely because the first-order state begins to exhibit indeterminacy. Alternatively, the warrant conferred by the first-order state on the self-belief component might not arise to the threshold required for knowledge.

§5.7.9—Rational Fundamentalist Objections to Constitutivism: Knowledge as Achievement

Parrott argues that Constitutivism is objectionable because it means denying that self-knowledge is a cognitive achievement.[245] As regards CF, the objection is that it paints self-knowledge of our commitments as metaphysically assured by our possession of these attitudes in C-conditions, rather than as resulting from any cognitive/epistemic activity. Therefore, constitutive self-

[245] This objection, Parrott notes, has been levied against Constitutivists by several authors, such as Boghossian (1989) and Fernandez (2013).

knowledge is made to stand out as radically unlike other varieties of knowledge. Rational Fundamentalism, on the other hand, is supposed to have no problem explaining the sense in which self-knowledge is a cognitive achievement (more on this in a moment).

I am not, in general, inclined to think that knowledge must be a cognitive achievement. Cognitive achievements are the sorts of things that *produce* knowledge, but for this very reason cognitive achievements can be pulled apart both conceptually and metaphysically from knowledge itself. So my first, flat-footed response is simply to deny that there is a terribly compelling problem for CF here.

However, it may not even be the case that constitutive self-knowledge is alone among kinds of knowledge—even propositional knowledge—in not involving cognitive achievements, and so Constitutivists might find strength in numbers here. For example, J. Adam Carter and Duncan Pritchard (2015) argue that many cases of testimonial knowledge involve no cognitive achievement on the hearer's part, the idea being that the onus is entirely on the speaker to transfer knowledge to the hearer. Granted, the hearer must *cognize the testimony* in order to acquire knowledge via testimony. But this does not imply that testimonial knowledge is different from constitutive self-knowledge in virtue of involving a cognitive achievement. After all, to have constitutive self-knowledge, one must cognize the (putative) first-order evidence that leads one to form one's commitments in the first place.

Finally, it is not clear that Rational Fundamentalism really fares any better than CF in allowing us to treat self-knowledge as a cognitive achievement. Rational Fundamentalism claims that self-knowledge is the result of a disposition, possessed by rational agents, to form higher-order beliefs about their minds. But how does it follow that the agent has cognitively achieved something when it triggers? Its triggering is a product of her rational nature, rather than being the

product of any cognitive or epistemic activity. Granted, whether or not a disposition triggers can sometimes be a matter of things one can achieve. One can, for instance, take steps to avoid fatigue or other conditions of cognitive poor-functioning that would impede the triggering of a disposition. But proponents of CF can make similar claims: an agent can take steps to avoid fatigue such that, *ceteris paribus*, she will be in a dual-order state of mind. There are things agents can do, in other words, to assure that they are in C-conditions. And if we want to say that cognitive achievements go into this, then so be it. I only deny that, once one is in C-conditions, something additional must be done or must happen in order for one to have self-knowledge of one's commitments.

§5.7.10—Rational Fundamentalism and Constitutivism: Summary

Across the last several subsections I have argued that various objections to Constitutivism—and indeed, my particular constitutive thesis CF—can be overcome. On my view, CF does no worse than Rational Fundamentalism, and likely fares better than Rational Fundamentalism in one crucial respect (think back to Unavoidable Failures, from §5.7.5). I conclude that:

> **Constitutivism-Final**: Given C-conditions (rationality, intelligence, conceptual competence, lucidity, alertness, and attentiveness)—one's beliefs/desires/intentions that P/to φ as a commitment stand in a part-whole relation to one's beliefs that one believes/desires/intends that P/to φ as a commitment.

…is a plausible substantive account of our self-knowledge of commitments. It nicely captures the anti-introspectionist/non-epistemic bent pursued by Neo-Expressivist analysis of avowals without denying that our avowals do, in fact, express privileged and peculiar self-knowledge as well as the first-order commitments they are about. Moreover, it nicely accounts for the apparent impossibilities of self-blindness and brutely erroneous self-beliefs. Finally, it nicely accounts for

the thought that commissive self-knowledge is necessitated by our social-epistemic and self-regulative agency.

§5.8.1—Conclusion

In this dissertation I reflected on the common claim that agents have authority with respect to their mental states. Understood as a quasi-social phenomenon, I argued that there are many senses in which agents have authority: their avowals ought to be treated as relatively indubitable and they (ought to) receive a distinctive presumption of truth.

In Chapter Two I argued that we can explain the various dimensions of our authority by embracing Neo-Expressivism, which is the view according to which avowing is a way of putting our first-order mental states directly on display for others to perceive. However, it is commonly thought that avowals are authoritative for a different reason, namely, because they express privileged and peculiar self-knowledge of our first-order mental states. Like some other Neo-Expressivists, I did not deny that we have such self-knowledge. However, this raised anew the question of what such self-knowledge explains—what functional roles it has in our psychological economies, if not that of explaining our authority.

The search for a plausible answer to this question took us through Chapters Three and Four, albeit with a narrowed focus on privileged and peculiar self-knowledge of propositional attitudes. In Chapter Four I developed a Social Agentialist account which claimed that privileged and peculiar self-knowledge of our propositional attitudes *as commitments* is indispensable to multiple social-epistemic capacities, and to our self-regulative capacity for future-directed acts of self-control. I also argued that coupling a Neo-Expressivist account of authority with a Social Agentialist account of privileged and peculiar self-knowledge of our commitments could jointly

explain why such self-knowledge and our authority (explained along Neo-Expressivist lines) often come as a package, despite neither explaining the other.

In this fifth and final chapter I have argued that we can account for the source of privileged and peculiar self-knowledge of our commitments by embracing a version of Constitutivism. On this account, self-knowledge is privileged and peculiar because it is constitutive of having commitments, given a plausible set of background conditions. Constitutivism is, moreover, congenial to Neo-Expressivism. In avowing, we express our commitments as well as our self-beliefs (that count as knowledge), because these constitute one another. The resultant picture is one on which there is no tension between our self-conception as privileged and peculiar self-knowers on the one hand, and as authoritative, Neo-Expressive avowers on the other.

Bibliography

Alshanetsky, E. 2020. The Meno Paradox of Reflection. *The Journal of Philosophy*, https://doi.org/10.5840/jphil2020117414

Alston, W. P. 1967. Expressing. In Black, E. (Ed.), *Philosophy in America*, 15-34. Cornell University Press.

Andreotta, A. 2019. Confabulation Does Not Undermine Introspection for Propositional Attitudes. *Synthese*, https://doi.org/10.1007/s11229-019-02373-9

Andreotta, A. 2020. Extending the Transparency Method Beyond Belief: A Solution to the Generality Problem. *Acta Analytica*, https://doi.org/10.1007/s12136-020-00447-9

Anscombe, G. E. M. 1963. *Intention*. 2nd Edition. Oxford: Blackwell.

Arango–Muñoz, S. 2019. Cognitive Phenomenology and Metacognitive Feelings. *Mind & Language*, 34(2): 247-262.

Armstrong, D. 1963. Is Introspective Knowledge Incorrigible? *Philosophical Review* 72: 417 - 432.

Armstrong, D. 1968. *A Materialist Theory of Mind*. Routledge.

Ayer, Alfred J. 1956. What is a Law of Nature?. *Revue Internationale de Philosophie*, 10:144–65.

Ayer, A. J. 1963. Privacy. In his *The Concept of a Person and Other Essays*. London, UK: MacMillan.

Bagnoli, C. 2012. Self-Deception and Agential Authority: A Constitutivist Account. *Humana.Mente Journal of Philosophical Studies*, 20: 99-116.

Bar-On, D. 2000. Speaking My Mind. *Philosophical Topics*, 28:1-34.

Bar-On, D. & Long, D. 2001. Avowals and First-Person Privilege. *Philosophy and Phenomenological Research*, 62: 311-335.

Bar-On, D. & Long, D. 2003. "Knowing Selves: Expression, Truth, and Knowledge" In Gertler, B. (Ed.), *Privileged Access: Philosophical Accounts of Self-Knowledge*, 179-212. Routledge.

Bar-On, D. 2004. *Speaking My Mind: Expression and Self-Knowledge*. Oxford University Press.

Bar-On, D. 2009. First-Person Authority: Dualism, Constitutivism, and Neo-Expressivism. *Erkenntnis*, 71(1): 53-71.

Bar-On, D. 2010. Avowals: Expression, Security, and Knowledge: Reply to Matthew Boyle, David Rosenthal, and Maura Tumulty. *Acta Analytica*, 25: 47-63.

Bar-On, D. 2011. "Neo-Expressivism: Avowals' Security and Privileged Self-Knowledge" in Hatzimoysis (Ed.), *Self-Knowledge*, 189-201.

Bar-On, D. 2012. "Externalism and Skepticism: Recognition, Expression, and Self-Knowledge" In Coliva, A. (Ed.), *The Self and Self-Knowledge*, 189-211. Oxford University Press.

Bar-On, D. 2015. Transparency, Expression, and Self-Knowledge. *Philosophical Explorations*, 18(2): 134-152.

Bar-On, D. & Johnson, D. 2019. Epistemological Disjunctivism: Perception, Expression, and Self-Knowledge. In Doyle, C., Milburn, J. & Pritchard, D. (Eds.), *New Issues in Epistemological Disjunctivism*, 317-344. Routledge.

Barz, W. 2014. Transparent Introspection of Wishes. *Philosophical Studies*, 172(8): 1993-2023.

Barz, W. 2017. Luminosity Regained. *Pacific Philosophical Quarterly*, 98(S1): 480-496.

Barz, W. 2018. Is There Anything to the Authority Thesis? *Journal of Philosophical Research*, 43. doi:10.5840/jpr2018712122

Barz, W. 2019. The Puzzle of Transparency and How to Solve It. *Canadian Journal of Philosophy*, DOI: 10.1080/00455091.2019.1565620
Bettcher, T. 2009. Trans Identities and First-Person Authority. In Shrage, L. (Ed.) *You've Changed: Sex Reassignment and Personal Identity*, 98-120. Oxford University Press.
Bilgrami, A. 1998. Self Knowledge and Resentment. In Wright C., Smith B., and MacDonald C. (Eds.), *Knowing Our Own Minds*, 207-241. Oxford University Press.
Bilgrami, A. 2006a. *Self-Knowledge and Resentment*. Harvard University Press.
Bilgrami, A. 2006b. Some Philosophical Integrations. In Macdonald C. and Macdonald M. (Eds.), *McDowell and his Critics*, 50-65. Wiley-Blackwell.
Bilgrami, A. 2012. The Unique Status of Self-Knowledge. In Coliva A. (Ed.), *The Self and Self-Knowledge*, 263-278.
Blackwood, Stephen. *Self-Knowledge and Rationality*. Doctoral Dissertation, Wilfrid Laurier University, 2010.
Boghossian, P. 1989. Content and Self-Knowledge. *Philosophical Topics*, 17(1): 5-26.
Boghossian, P. 2014. What is Inference? *Philosophical Studies*, 169(1): 1-18.
Boghossian, P. 2015. Reasoning and Reflection: A Reply to Kornblith. *Analysis*, 76(1): 41-54.
Boghossian, P. 2018. Delimiting the Bounds of Inference. *Philosophical Issues, 28, Philosophy of Logic and Inferential Reasoning*. doi: 10.1111/phis.12115
Boghossian, P. 2019. Inference, Agency, and Responsibility. In Balcerak Jackson, B. & Balcerak Jackson, M. (Eds.), *Reasoning: New Essays on Theoretical and Practical Thinking*, pp. 101-124.
Bonjour, L. 2009. *Epistemology: Classic Problems and Contemporary Responses*. Rowman & Littlefield. Plymouth, UK. Second edition printed in 2010.
Borgoni, C. 2015. On Knowing One's Resistant Beliefs. *Philosophical Explorations*, 18(2): 212-225.
Borgoni, C. & Luthra, Y. 2017. Epistemic Akrasia and the Fallibility of Critical Reasoning. *Philosophical Studies*, 174: 877-886.
Borgoni, C. 2018a. Unendorsed Beliefs. *Dialectica*, 72(1): 49-68.
Borgoni, C. 2018b. Basic Self-Knowledge and Transparency. *Synthese*, 195: 679-696.
Borgoni, C. 2019. Authority and Attribution. The Case of Epistemic Justice in Self-Knowledge. *Philosophia*. https://doi.org/10.1007/s11406-018-0002-x
Boyle, M. 2009. Two Kinds of Self-Knowledge. *Philosophy and Phenomenological Research*, 78(1): 133-164.
Boyle, M. 2011. Transparent Self-Knowledge. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 85(1): 223-241.
Boyle, M. 2015. Critical Study: Cassam on Self-Knowledge for Humans. *European Journal of Philosophy*, DOI: 10.1111/ejop.12117
Boyle, M. 2019. Transparency and Reflection. *Canadian Journal of Philosophy*, DOI: 10.1080/00455091.2019.1565621
Brandom, R. 1994. *Making It Explicit. Reasoning, Representing, and Discursive Commitment*. Harvard University Press.
Broome, J. 2014. Comments on Boghossian. *Philosophical Studies*, 169: 19-25.
Broome, J. 2019. A Linking Belief is not Necessary for Reasoning. In Balcerak Jackson, B. & Balcerak Jackson, M. (Eds.), *Reasoning: New Essays on Theoretical and Practical Thinking*, pp. 32-43. Oxford University Press.

Brueckner, A. & Fiocco, M. 2002. Williamson's Anti-Luminosity Argument. *Philosophical Studies*, 110 (3): 285-293.
Brueckner, A. 2011. 'Neo-Expressivism.' In A. Hatzimoysis (ed.), *Self-Knowledge*, 170–88. Oxford University Press.
Burge, T. 1988. Individualism and Self-Knowledge. *The Journal of Philosophy*, 85(11): 649-663.
Burge, T. 1993. Content Preservation. *The Philosophical Review*, 102(4): 457-488.
Burge, T. 1996. Our Entitlement to Self-Knowledge. *Proceedings of the Aristotelian Society*, 96(1): 1-26.
Burge, T. 2007a. Two Kinds of Consciousness. In Burge, T. (Ed.), *Philosophical essays, vol. II: Foundations of mind*, 383–391. Oxford University Press.
Burge, T. 2007b. Reflections on two kinds of consciousness. In Burge, T. (Ed.), *Philosophical essays, vol. II: Foundations of mind*. Oxford University Press, 392-419.
Burge, T. 2013. *Cognition Through Understanding: Self-Knowledge, Interlocution, Reasoning, Reflection*. *Philosophical Essays, Volume 3*. Oxford University Press.
Butterfill, S. 2016. Tracking and Representing Others' Mental States. In Andrews, K. and Beck, J. (Eds.), *Routledge Companion to the Philosophy of Animal Minds*, 269-279. Routledge.
Byrne, A. 2005. Introspection. *Philosophical Topics*, 33(1): 79-104.
Byrne, A. 2007. Review Essay of Dorit Bar-On's *Speaking My Mind*. *Philosophy and Phenomenological Research*, 83(3): 705-717.
Byrne, A. 2011a. "Knowing What I Want" In J. Liu & J. Perry (Eds.), *Consciousness and the Self: New Essays*. Cambridge University Press.
Byrne, A. 2011b. Transparency, Belief, Intention. *Aristotelian Society Supplementary Volume*, 85, 201-221.
Byrne, A. 2011c. Review Essay of Dorit Bar-On's Speaking My Mind. *Philosophy and Phenomenological Research*, 83(3): 705-717.
Byrne, A. 2018. *Transparency and Self-Knowledge.* Oxford University Press.
Campbell, L. 2018. Self-Knowledge, Belief, Ability (and Agency?). *Philosophical Explorations*, 21(3): 333-349.
Carroll, L. 1895. What the Tortoise Said to Achilles. *Mind*, 4(14): 278-280.
Carruthers, P. 2011. *The Opacity of Mind: An Integrative Theory of Self-Knowledge.* Oxford University Press.
Cassam, Q. 2009. The Basis of Self-Knowledge. *Erkenntnis*, 71: 3-18.
Cassam, Q. 2015. *Self-Knowledge for Humans*. Oxford University Press.
Chrisman, M. 2009. Expressivism, Truth, and (Self-) Knowledge. *Philosopher's Imprint*, 9(3): 1-26.
Cholbi, M. J. 2016. Review of *Self-Knowledge for Humans* by Quassim Cassam. *Philosophy*, 91(3): 441-446.
Chudnoff, E. *Intuition*. Oxford University Press.
Churchland, P. 1984. *Matter and Consciousness*. MIT Press.
Coliva, A. 2009. Self-Knowledge and Commitments. *Synthese*, 171: 365-375.
Coliva, A. 2012. "One Variety of Self-Knowledge: Constitutivism as Constructivism." In Coliva, A. (Ed.) *The Self and Self-Knowledge*, 212-242. Oxford University Press.
Coliva, A. 2014. Review of Jordi Fernández's *Transparent Minds*. *Theoria*, 81: 443-449.
Coliva, A. 2015. Review of Quassim Cassam *Self-Knowledge for Humans*. *Analysis*.
Coliva, A. 2016. *The Varieties of Self-Knowledge*. Palgrave Macmillan.

Coliva, A. 2019a. Disagreeing With Myself: Doxastic Commitments and Intrapersonal Disagreement. *American Philosophical Quarterly*, 56(1): 1-13.
Coliva, A. 2019b. Replies to: Commentators. *Philosophia*, 47(2): 343-352.
Dain, E. 2019. Wittgenstein, Mindreading and Perception. *European Journal of Philosophy*, 27: 675-692.
Davidson, D. 1973. Radical Interpretation. *Dialectica*, 27: 314-28.
Davidson, D. 1984. First Person Authority. *Dialectica*, 38(2-3): 101-112.
Davidson, D. 1986. Deception and Division. In Elster, J. (Ed.), *The Multiple Self*. Cambridge University Press. Reprinted in Davidson 2004, *Problems of Rationality*, OUP.
Davidson, D. 1987. Knowing One's Own Mind. *Proceedings of the American Philosophical Association*, 60(3): 441-458.
Davidson, D. 1989. What is Present to the Mind. *Grazier Philosophische Studien*. Netherlands: Rodopi. 197-213.
Davidson, D. 1991. Three Varieties of Knowledge. *Royal Institute of Philosophy Supplement*, 30: 153-166. Reprinted in Davidson (2001a) *Subjective, Intersubjective, Objective* (pp. 205-220).
Davidson, D. 1995. Davidson, Donald. In Guttenplan, S. (Ed.), *A Companion to the Philosophy of Mind*, 231-236. Oxford University Press.
Davidson, Donald. 1998. The Irreducibility of the Concept of the Self. *Philosophie in Synthetischer Absicht*. Stuttgart: Klett-Cotta. Reprinted in Davidson 2001a, 85-91.
Davidson, D. 1999. The Emergence of Thought. *Erkenntnis*, 51(1): 7-17. Reprinted in Davidson 2001a.
Davidson, D. 2001a. *Subjective, Intersubjective, Objective*. Oxford University Press.
De Brasi, L. 2015. The Peculiarity and Contingency of the Introspection of Belief. *Filosofia Unisinos*, 16(2): 100-118.
De Bruin, L., Jongepier, F. & Strijbos, D. 2014. Mental Agency and Self-Regulation. *Review of Philosophical Psychology*, 6: 815-825. DOI 10.1007/s13164-014-0190-7
Descartes, R. 1984. *Meditations*, in Philosophical Writings of Descartes, Volume II, transl. J. Cottingham, R. Stoothoff and D. Murdoch, Cambridge University Press.
Doyle, C. 2015. *Four Essays on Self-Knowledge*. Dissertation: University of Pittsburgh. http://d-scholarship.pitt.edu/24970/
Doyle, C. 2017. Engagement, Expression, and Initiation. In Peters, M.A. & Stickney, J (Eds.), *A Companion to Wittgenstein on Education*, 467-479. DOI 10.1007/978-981-10-3136-6_31
Doyle, C. 2018a. Agency and Observation in Knowledge of One's Own thinking. *European Journal of Philosophy*: 1-14.
Doyle, C. 2018b. Deferring to Others About One's Own Mind. *Pacific Philosophical Quarterly*, DOI: 10.1111/papq.12268
Doyle, C. 2019. Ringers for Belief. In Doyle, C., Milburn, J. & Pritchard, D. (Eds.), *New Issues in Epistemological Disjunctivism*, 345-365. Routledge.
Doyle, C. 2020. The Sense of Agency in Conscious Thinking. *Erkenntnis*, https://doi.org/10.1007/s10670-020-00317-1
Dretske, F. 1973. Perception and Other Minds. *Noûs*, 7(1): 34-44.
Esken, F. 2012. Early Forms of Metacognition in Human Children. In Beran, M., Brandl, J., Perner, J., and Proust, J. *Foundations of Metacognition*, 134-145. Oxford University Press.
Evans, G. 1982. *The Varieties of Reference*. Oxford University Press.

Evans, T. & Beran, M. Chimpanzees Use Self-Distraction to Cope With Impulsivity. *Biology Letters*, 22(3): 699-602.
Falvey, K & Owens, J. 1994. Externalism, Self-Knowledge, and Skepticism. *The Philosophical Review*, 103(1): 107-137.
Falvey, K. 2000. The Basis of First-Person Authority. *Philosophical Topics*, 28(2): 69-99.
Fernández, J. 2013. *Transparent Minds*. Oxford University Press.
Finkelstein, D. 2003. *Expression and the Inner*. Harvard University Press.
Finkelstein, D. 2011. Replies to my Commentators. *Teorema: Revista Internacional de Filosofía* 30(3): 79-95.
Flanagan, O. 1992. *Consciousness Reconsidered*. MIT Press.
Fodor, J. 1975. *The Language of Thought*. Harvard University Press.
Frege, G. 1979. *Posthumous Writings*. Hermes, H., Kambartel, F. and Kaulbach, F. (Eds.), Basil Blackwell.
Fricker, E. 2006. Second-Hand Knowledge. *Philosophy and Phenomenological Research*, 73(3): 592-618.
Fricker, M. 2007. *Epistemic Injustice: Power & The Ethics of Knowing*. Oxford University Press.
Frith, C. 2012. The Role of Metacognition in Human Social Interactions. *Philosophical Transactions of the Royal Society*, 367: 2213-2223.
Gallois, A. 1997. *The World Without, The Mind Within: An Essay on First-Person Authority*. Cambridge University Press.
García Rodríguez, A. 2019. Expression and the Transparency of Belief. *European Journal of Philosophy*, 27: 136-147.
Geach, P. 1971. 'Assertion' in J. Rosenberg and C. Travis (Eds.), *Readings in the Philosophy of Language*, 250-61. Prentice-Hall.
Gendler, T. S. 2007. Self-Deception as Pretense. *Philosophical Perspectives*, 21(1): 231-258.
Gendler, T. S. 2008. Alief and Belief. *The Journal of Philosophy*, 105(10): 634-663.
Gertler, B. 2000. The Mechanics of Self-Knowledge. *Philosophical Topics*, 28(2): 125-146.
Gertler, B. 2011a. *Self-Knowledge*. Routledge.
Gertler, B. 2011b. "Self-Knowledge and the Transparency of Belief" in A. Hatzimoysis (Ed.), *Self-Knowledge*, 125-145. Oxford University Press.
Gertler, B. 2016a. Critical Notice of *Self-Knowledge for Humans*, by Quassim Cassam. *Mind*, 125(497): 269-280.
Gertler, B. 2016b. Self-Knowledge and Rational Agency: A Defense of Empiricism. *Philosophy and Phenomenological Research*, 96(1): 91-109.
Gertler, B. Forthcoming. Rational Agency and the Struggle to Believe What Your Reasons Dictate. In Borgoni, C., Kindermann, D., and Onofri, A. (Eds.), *The Fragmented Mind*, Oxford University Press.
Giananti, A. 2020. I Know How I Know: Perception, Self-Awareness, Self-Knowledge. *Synthese*, https://doi.org/10.1007/s11229-020-02726-9
Giustina, A. 2019. Introspection Without Judgment. *Erkenntnis*, https://doi.org/10.1007/s10670-019-00111-8
Goldman, A. 2006. *Simulating Minds*. Oxford University Press.
Golob, S. 2015. Self-Knowledge, Transparency, and Self-Authorship. *Proceedings of the Aristotelian Society*, 115(3-3): 235-253.
Gopnik, A. 1993. How We Know Our Own Minds: The Illusion of First-Person Knowledge of Intentionality. *Behavioral and Brain Sciences*, 16: 1-15.

Gray, K. et al. 2011. More Than a Body: Mind Perception and the Nature of Objectification. *Journal of Personality and Social Psychology*, 101(6): 1207-1220.
Green, M. 2007. *Self-Expression*. Oxford University Press.
Greene, R. 2003. Constitutive Theories of Self-Knowledge and the Regress Problem. *Philosophical Papers*, 32(2): 141-148.
Grimm, S. 2011. Understanding. In D. Pritchard and S. Bernecker (Eds.), *The Routledge Companion to Epistemology*.
Hacker, P.M.S. 1986. *Insight and Illusion: Themes in the Philosophy of Wittgenstein*. Clarendon Press.
Hacker, P. M. S. 1990. *Wittgenstein: Meaning and Mind. Part I: Essays*. Basil Blackwell.
Haidt, J. 2001. The Emotional Dog and its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychology Review*, 108: 814-834.
Heal, J. 2002. First Person Authority. *Proceedings of the Aristotelian Society*, 102-1: 1-19.
Helton, G. 2018. If You Can't Change What You Believe, You Don't Believe It. *Noûs*, 17(3): 1-26.
Hlobil, U. 2014. Against Boghossian, Wright, and Broome on Inference. *Philosophical Studies*, 167(2): 419-429.
Hlobil, U. 2018. We Cannot Infer by Accepting Testimony. *Philosophical Studies*, 176: 2589-2598.
Hlobil, U. 2019. Inferring by Attaching Force. *Australasian Journal of Philosophy*, 97(4): 701-714.
Horgan, T & Kriegel, U. 2007. Phenomenal Epistemology: What is Consciousness That We May Know it So Well? *Philosophical Issues*, 17: 123-144.
Hume, D. 1741. *A Treatise of Human Nature*. Norton, D. (Ed.). Oxford University Press.
Hunter, D. 2011. Alienated Belief. *Dialectica*, 65(2): 221-240.
Hunter, D. ms. Inference as a Mental Act.
Hursthouse, R. 1991. Arational Actions. *The Journal of Philosophy*, 88(2): 57-68.
Hyman, J. 2015. *Action, Knowledge and Will*. Oxford University Press.
Jacobsen, R. 1996. Wittgenstein on Self-Knowledge and Self-Expression. *The Philosophical Quarterly*, 46(182): 12-30.
Jacobsen, R. 1997a. Self-Quotation and Self-Knowledge. *Synthese*, 110(3): 419-445.
Jacobsen, R. 1997b. Semantic Character and Expressive Content. *Philosophical Papers*, 26(2): 129-146.
Jacobsen, R. 2008. The Duck Quacks Back: A Reply to A. Minh Nguyen. *Dialogue*, 48: 655-663.
Jacobsen, R. 2009. Davidson and First-Person Authority: Parataxis and Self-Expression. *Pacific Philosophical Quarterly*, 90: 251-266.
Jacobsen, R. ms. Truth and Sincerity: Self-Knowledge and the Nature of the Mental.
Johnson, C. N., & Wellman, H. M. 1980. Children's Developing Understanding of Mental Verbs: Remember, Know, and Guess. *Child Development*, 51(4): 1095–1102.
Jongepier, F. 2020. The Value of Transparent Self-Knowledge. *Ethical Theory and Moral Practice*, https://doi.org/10.1007/s10677-020-10118-8
Kahneman, D. 2011. *Thinking, Fast and Slow*. Macmillan.
Keeling, S. 2019a. The Transparency Method and Knowing Our Reasons. *Analysis*, doi:10.1093/analys/anz031

Keeling, S. 2019b. Knowing Our Reasons: Distinctive Self-Knowledge of Why We Hold Our Attitudes and Perform Actions. *Philosophy and Phenomenological Research*, DOI: 10.1111/phpr.12655
Kiefer, A. 2017. Literal Perceptual Inference. In Metzinger, T. and Wiese, W. (Eds.), *Philosophy and Predictive Processing: 17*.
Kietzmann, C. 2018. Inference and the Taking Condition. *Ratio*, 31(3): 294-302.
Klein, P. 2009. Contemporary Responses to Agrippa's Trilemma. In Greco, J. (Ed.), *The Oxford Handbook of Skepticism*, 485-500. Oxford University Press.
Knappik, F. 2015. Self-Knowledge About Attitudes: Rationalism Meets Interpretation. *Philosophical Explorations*, 18(2): 183-198.
Knappik, F. 2020. Sellars on Self-Knowledge. In Brandt, S. and Breunig, A. (Eds.), *Wilfrid Sellars and Twentieth-Century Philosophy*, 221-239. Routledge.
Knobe, J. & Prinz, J. 2008. Intuitions About Consciousness: Experimental Studies. *Phenomenology and the Cognitive Sciences*, 7: 67-83.
Komorowska-Mach, J. 2019. Introspection—One Or More? Pluralism about Self-Knowledge. *Filozofia Nauki*, 1(105): 5-25.
Koreň, L. 2019. Have Mercier and Sperber Untied the Knot of Human Reasoning? https://doi.org/10.1080/0020174X.2019.1684988
Kornblith, H. 2012. *On Reflection*. Oxford University Press.
Kornblith, H. 2016. "Epistemic Agency" in Vargas M. A. F. (Ed.), *Performance Epistemology: Foundations and Applications*, 167-182. Oxford University Press.
Koziolek, N. 2017. Inferring as a Way of Knowing. *Synthese*, https://doi.org/10.1007/s11229-017-1632-4
Koziolek, N. 2018. Belief as an Act of Reason. *Manuscrito*, 41(4): 287-318.
Lackey, J. 2006. The Nature of Testimony. *Pacific Philosophical Quarterly*, 87(2): 177-197.
Lackey, J. 2008. *Learning from Words: Testimony as a Source of Knowledge*. Oxford University Press.
Lawlor, K. 2008a. Review of *Self-Knowledge and Resentment* by Akeel Bilgrami. *Mind*, 117: 469-372.
Lawlor, K. 2008b. Knowing Beliefs, Seeking Causes. *American Imago*, 65(3): 335-356.
Lawlor, K. 2009. Knowing What One Wants. *Philosophy and Phenomenological Research*, 79(1): 47-75.
Leite, A. 2018. Changing One's Mind: Self-Conscious Belief and Rational Endorsement. *Philosophy and Phenomenological Research*, 97(1): 150-171.
Levine, S. 2009. Expressivism and I-Beliefs in Brandom's *Making it Explicit*. *International Journal of Philosophical Studies*, 17(1): 95–114.
Locke, D. 1967. *Perception and Our Knowledge of the External World*. Allen & Unwin.
Lohmann, H., & Tomasello, M. 2003. The Role of Language in the Development of False Belief Understanding: A Training Study. *Child Development*, 74(4): 1130-1144.
Ludlow, P. 1995. Externalism, Self-Knowledge, and the Prevalence of Slow-Switching. *Analysis*, 55: 45-49.
Ludwig, K. 1994. First-Person Knowledge and Authority. In Preyer, G., Siebelt, F. & Ulfig, A. (Eds.), *Language, Mind, and Epistemology: On Donald Davidson's Philosophy*. Dordrecht: Kluwer Academic Publishers.
Lycan, W. 1996. *Consciousness and Experience*. MIT Press.

Macdonald, C. 2014. In My 'Mind's Eye': Introspectionism, Detectivism, and the Basis of Authoritative Self-Knowledge. *Synthese*, 191(15): 3685-3710.
Malcolm, N. 1954. Wittgenstein's Philosophical Investigations. *The Philosophical Review, 63*, 530–559. https://doi.org/10.2307/2182289
Malcolm, N. 1958. Knowledge of Other Minds. *The Journal of Philosophy, 55*, 969–978. https://doi.org/10.2307/2021905
Malmgren, A. 2019. On Fundamental Responsibility. *Philosophical Issues: A Supplement to Noûs*, 29(1): 198-213.
Mandelbaum, E. 2014. Thinking is Believing. *Inquiry: An Interdisciplinary Journal of Philosophy*, 57(1): 55-96.
Manning, R. 2014. Sellarsian Behaviorism, Davidsonian Interpretivism, and First-Person Authority. *Philosophia*, 42: 433-456.
Marcus, E. 2016. To Believe is to Know that You Believe. *Dialectica*, 70-3: 375-405.
Marcus, E. 2020. Inference as Consciousness of Necessity. *Analytic Philosophy*, 61(4): 1-19.
Marcus, E. & Schwenkler, J. 2018. Assertion and Transparent Self-Knowledge. *Canadian Journal of Philosophy*, 49(7): 873-889.
McDowell, J. 1983. Criteria, Defeasibility, and Knowledge. *Proceedings of the British Academy*, 68: 455-479.
McDowell, J. 2006. Response to Bilgrami. In Macdonald C. and Macdonald M. (Eds.), *McDowell and his Critics*, 66-72. Wiley Blackwell.
McGeer, V. 1996. Is "Self-Knowledge" Really an Empirical Problem? Renegotiating the Space of Philosophical Explanation. *The Journal of Philosophy,* 93(10): 483-515.
McGeer, V. 2007. The Moral Development of First-Person Authority. *European Journal of Philosophy,* 16(1): 81-108.
McGeer. V. 2015. Mind-Making Practices: the Social Infrastructure of Self-Knowing Agency and Responsibility. *Philosophical Explorations*, 18(2): 259-281.
McGeer, V. & Pettit, P. 2002. The Self-regulating Mind. *Language and Communication*, 22: 281–99.
McGinn, C. 1982. *The Character of Mind*. Oxford University Press.
McGlynn, A. 2014. *Knowledge First?* Palgrave Macmillan.
McHugh, C. & Way, J. 2016. Against the Taking Condition. *Philosophical Issues, 26, Knowledge and Mind*: 314-331.
McHugh, C. & Way, J. 2018. What is Reasoning? *Mind*, 127(505): 167-196**.**
Mercier, H. 2011. Reasoning Serves Argumentation in Children. *Cognitive Development*, 26(3), 177-191.
Mercier, H & Sperber, D. 2009. Intuitive and Reflective Inferences. In Evans J. S. B. T. and Frankish K (Eds.), *In Two Minds*. Oxford Unversity Press.
Moll, H. 2013. Ontogenetic Precursors of Assertion and Denial. In Rödl, S. and Tegtmeyer, H. (Eds.), *Sinnkritisches Philosophieren*, 337-346. De Gruyter.
Moran, R. 2001. *Authority and Estrangement.* Princeton University Press.
Moran, R. 2003. Responses to O'Brien and Shoemaker. *European Journal of Philosophy,* 11(3): 402-419.
Moran, R. 2004. Replies Heal, Reginster, Wilson, and Lear. *Philosophy and Phenomenological Research*, 69(2): 455-472.
Moran, R. 2012. Self-Knowledge, 'Transparency', and the Forms of Activity. In Smithies D. & Stoljar D (Eds.), *Introspection and Consciousness*., 211-236. Oxford University Press.

Müller, A. 2019. Reasoning and Normative Beliefs: Not Too Sophisticated. *Philosophical Explorations*, 22(1): 2-15.
Nguyen, A. 2008. The Authority of Expressive Self-Ascriptions. *Dialogue*, 47: 103-136.
Nguyen, A. 2015. What Good is Self-Knowledge? *Journal of Philosophical Research*, 40: 137-154.
Nisbett, R. & Wilson, T. 1977. Telling More than We Can Know: Verbal Reports on Mental Processes. *Psychological Review*, 85(4): 231-259.
O'Brien, L. 2003. Moran on Agency and Self-Knowledge. *European Journal of Philosophy*, 11(3): 402-419.
O'Brien, L. 2005. Self-Knowledge, Agency and Force. *Philosophy and Phenomenological Research*, 71(3): 580-601.
O'Brien, L. 2007. *Self-Knowing Agents*. Oxford University Press.
Onishi, K. H., and R. Baillargeon. 2005. Do 15-Month-Old Infants Understand False Beliefs? *Science*, 308: 255–258.
O'Shaughnessy, B. 2000. *Consciousness and the World*. Oxford University Press.
Owens, D. 2000. *Rationality without Freedom*. Routledge.
Owens, D. 2006. Testimony and Assertion. *Philosophical Studies*, 130: 105-129.
Owens, D. 2011. "Deliberation and the First-Person." In Hatzimoysis A. (Ed.), *Self-Knowledge*. Oxford University Press, 261-278.
Owens, J. 2007. Review of Dorit Bar-On's *Speaking My Mind: Expression and Self-Knowledge*. *Notre Dame Philosophical Reviews*, https://ndpr.nd.edu/news/speaking-my-mind-expression-and-self-knowledge.
Parent, T. 2017. *Self-Reflection for the Opaque Mind: An Essay in Neo-Sellarsian Philosophy*. Routledge.
Parrott, M. 2010. Agency and First-Person Authority. Dissertation.
Parrott, M. 2015. Expressing First-Person Authority. *Philosophical Studies*, 172(8): 2215-2237.
Parrott, M. 2017. Self-Blindness and Self-Knowledge. *Philosophers' Imprint,* 17(16): 1-22.
Paul, R. & Elder, L. 2008. *Critical Thinking: Concepts and Tools*. Foundation for Critical Thinking Press.
Paul, S. 2015. The transparency of Intention. *Philosophical Studies*, 172(6): 1529-1548.
Peacocke, A. 2017. Embedded Mental Action in Self-Attribution of Belief. *Philosophical Studies*, 170: 353-377.
Peacocke, C. 1996. Our Entitlement to Self-Knowledge: Entitlement, Self-Knowledge and Conceptual Redeployment. *Proceedings of the Arisototelian Society*, 96(1): 117-58.
Peacocke, C. 1998. Conscious Attitudes, Attention, and Self-Knowledge. In Wright, C., Smith, B. & Macdonald, C. (Eds.), *Knowing Our Own Minds*, 63-98. Oxford University Press.
Peterson, J. 2020. The Value of Privileged Access. *European Journal of Philosophy*. DOI: 10.1111/ejop.12594
Pettit, P. 2007. Rationality, Reasoning, and Group Agency. *Dialectica*, 61(4): 495-519.
Pettit, P. 2016. Broome on Reasoning and Rule Following. *Philosophical Studies*, 173: 3373-3384.
Pitt, D. 2004. The Phenomenology of Cognition or What is it Like to Think that P? *Philosophy and Phenomenological Research*, 69(1): 1-36.
Piñeros Glasscock, J. 2019. Practical Knowledge and Luminosity. *Mind*, https://doi.org/10.1093/mind/fzz056

Pritchard, D. & Carter, J. 2015. Knowledge-How and Cognitive Achievement. *Philosophy and Phenomenological Research*, 91(1): 181-199.
Proust, J. 2013. *The Philosophy of Metacognition: Mental Agency and Self-Awareness.* Oxford University Press.
Quilty-Dunn, J. & Mandelbaum, E. 2018. Inferential Transitions. *Australasian Journal of Philosophy*, 96(3): 532-547.
Richard, M. 2019. Is Reasoning Something the Reasoner Does? In Balcerak Jackson, B. & Balcerak Jackson, M. (Eds.), *Reasoning: New Essays on Theoretical and Practical Thinking*. Oxford University Press, 101-124.
Rödl, S. 2007. *Self-Consciousness*. Harvard University Press.
Roessler, J. 2013. The Silence of Self-Knowledge. *Philosophical Explorations*, 16(1): 1-17.
Roessler, J. 2015. Self-Knowledge and Communication. *Philosophical Explorations*, 18(2): 153-168.
Rorty, R. 1970. Incorrigibility as the Mark of the Mental. *Journal of Philosophy*, 12: 399-424**.**
Rosa, L. Reasoning Without Regress. *Synthese*, DOI 10.1007/s11229-017-1535-4
Rosenthal, D. 1993. Thinking That One Thinks. In Davies, M. & Humphreys, G. W. (Eds.), *Consciousness*. Basil Blackwell.
Rosenthal, D. 2005. *Consciousness and Mind*. United Kingdom: Oxford University Press.
Russell, B. 1920. The Nature of Inference, The Athenæum 4694: 514–15.
Ryle, G. 1949. *The Concept of Mind*. University of Chicago Press. Reprinted in 2009 by Routledge.
Samoilova, K. 2015. First-Person Privilege, Judgement, and Avowal. *Philosophical Explorations*, 18(2): 169-182.
Samoilova, K. 2016. Transparency and Introspective Justification. *Synthese*, 193: 3363-3381.
Scanlon, T. M. 1998. *What We Owe to Each Other*. Harvard University Press.
Sawyer, S. 2019. Concepts, Conceptions, and Self-Knowledge. *Erkenntnis*, https://doi.org/10.1007/s10670-019-00109-2
Schwengerer, L. 2019. Beliefs Over Avowals: Setting Up the Discourse on Self-Knowledge. *Episteme*, doi:10.1017/epi.2018.56
Schwengerer, L. 2020. Toward Collective Self-Knowledge. *Erkenntnis*, https://doi.org/10.1007/s10670-020-00235-2
Schwitzgebel, E. 2008. The Unreliability of Naïve Introspection. *Philosophical Review*, 117: 245-273.
Schwitzgebel, E. 2009. Knowing Your Own Beliefs. *Canadian Journal of Philosophy Supplementary Volume*, Volume 35 (Belief and Agency): 41-62.
Schwitzgebel, E. 2010. Acting Contrary to Our Professed Beliefs or the Gulf Between Occurrent Judgement and Dispositional Belief. *Pacific Philosophical Quarterly*, 91(4): 531-553.
Sellars, W. 1963. *Science, Perception, and Reality*. Routledge and Kegan Paul.
Setiya, K. 2011. Knowledge of Intention. In Ford, A., Hornsby, J. & Stoutland, F. (Eds.), *Essays on Anscombe's Intention*, 170-197. Harvard University Press.
Setiya, K. 2013. Epistemic Agency: Some Doubts. *Philosophical Issues: A Supplement to Noûs*, 23: 179-198.
Shah, N. & Velleman, D. 2005. Doxastic Deliberation. *Philosophical Review*, 114: 497-534.
Shah, N. & Vavova, K. 2014. Review of *On Reflection*, by Hilary Kornblith. *Ethics*: 632-636.
Shatz, M., Wellman, H., M., & Silber, S. 1983. The acquisition of Mental Verbs: A Systematic Investigation of the First Reference to Mental State. *Cognition*, 14(3): 301-321.

Sher, G. 2009. *Who Knew? Responsibility Without Awareness*. Oxford University Press.
Shoemaker, S. 1990. First-Person Access. *Action Theory and Philosophy of Mind*, 4: 187-214.
Shoemaker, S. 1994. Introspection and the Self. In Cassam, Q. (Ed.), *Self-Knowledge*. Oxford University Press.
Shoemaker, S. 1996a. *The First-Person Perspective and Other Essays*. Cambridge University Press.
Shoemaker, S. 1996b. *Self-Knowledge and Inner Sense*. Lectures I-III, in Shoemaker S. (1996a), 201-268.
Shoemaker, S. 2003. Moran on self-knowledge. *European Journal of Philosophy*, 3: 391–401.
Shoemaker, S. 2009. Self-Intimation and Second-Order Belief. *Erkenntnis*, 71(1): 35-51.
Siegel, S. 2019. Reasoning Without Reckoning. In Balcerak Jackson, B. & Balcerak Jackson, M. (Eds.), *Reasoning: New Essays on Theoretical and Practical Thinking*, 15-31. Oxford University Press.
Siewert, C. 2003. Self-Knowledge and Rationality: Shoemaker on Self-Blindness. In Gertler, B. (Ed.), *Privileged Access: Philosophical Accounts of Self-Knowledge*, 131-143. Ashgate Publishing.
Silins, N. 2012. "Judgement as a Guide to Belief" In Smithies, D. & Stoljar, D. (Eds.), *Introspection and Consciousness*. Oxford University Press.
Smith, E. R., & Miller, F. D. 1978. Limits on Perception of Cognitive Processes: A Reply to Nisbett and Wilson. Psychological Review, 85(4): 355–362.
Smith, J. & Wald, B. 2019. Collectivized Intellectualism. *Res Philosophica*, 96(2): 199-227.
Smithies, D. A Simple Theory of Introspection. 2012. In Smithies, D. & Stoljar, D. (Eds.), *Introspection and Consciousness*. Oxford University Press.
Smithies, D. and Stoljar, D. 2012. Introspection and Consciousness: An Overview. In Smithies, D. & Stoljar, D. (Eds.), *Introspection and Consciousness*, 1-27. Oxford University Press.
Sosa, E. 2007. *Apt Belief and Reflective Knowledge*. Oxford University Press.
Spitzley, T. 2009. Rationality and Self-Knowledge. *Erkenntnis*, 71(3): 73-88.
Snowdon, P. 2012. "How to Think About Phenomenal Self-Knowledge." In Coliva, A. (Ed.), *The Self and Self-Knowledge*. 243-262. Oxford University Press.
Sorgiovanni, B. 2019. The Agential Point of View. *Pacific Philosophical Quarterly*, 100(2): 549-572.
Sperber, D. & Mercier, H. 2011. Why Do Humans Reason? Arguments for an Argumentative Theory. *Behavioral and Brain Sciences*, 34: 57-111.
Sperber, D. & Mercier, H. 2012. "Reasoning as a Social Competence." In Landemore, H. and Elster, J. (Eds.), *Collective Wisdom Principles and Mechanisms*, 368–392. Cambridge University Press.
Sperber, D. & Mercier, H. 2017. *The Enigma of Reason*. Harvard University Press.
Srinivasan, A. 2015. Are We Luminous? *Philosophy and Phenomenological Research*, 90(2): 294-319.
Stanovich, K. & West, R. 2000. Individual Differences in Reasoning: Implications for the Rationality Debate. *Behavioral and Brain Sciences*, 23(5): 645-665.
Stoljar, D. 2007. Distinctions in Distinction. In Kallestrup, J. and Hohwy, J (Eds.), *Being Reduced: New Essays on Causation and Explanation in the Special Sciences.* Oxford University Press.
Stoljar, D. 2019. Evans on Transparency: A Rationalist Account. *Philosophical Studies*, 176: 2067-2085.

Strawson, P. 1974. *Freedom and Resentment and Other Essays*. Methuen.
Street, C. & Richardson, D. 2015. Descartes Versus Spinoza: Truth, Uncertainty, and Bias. *Social Cognition*, 33(3): 227-239.
Street, C. & Kingstone, A. 2017. Aligning Spinoza with Descartes: An Informed Cartesian Account of the Truth Bias. *British Journal of Psychology*, 108: 453-466.
Stein, N. & Bernas, R. 1999. The Early Emergence of Argumentative Knowledge and Skill. In Andriessen, J. and Coirer, P. (Eds.), *Foundations of Argumentative Text Processing*. Amsterdam: Amsterdam University Press.
Stueber, K. 2002. The Problem of Self-Knowledge. *Erkenntnis*, 56: 269-296.
Sullivan, K., & Winner, E. 1993. Three-Year-Olds' Understanding of Mental States: The Influence of Trickery. *Journal of Experimental Child Psychology*, 56(2): 135–148.
Sullivan, K., Zaitchik, D., & Tager-Flusber, H. 1994. Preschoolers Can Attribute Second-Order Beliefs. *Developmental Psychology*, 30(3): 395-402.
Tanesini, A. 2008. Review of Self-Knowledge and Resentment by Akeel Bilgrami. *Philosophical Books*, 49(3): 238-245.
Thompson, D. 2009. *Daniel Dennett*. Continuum Books.
Thompson, M. 2012. Anscombe's *Intention* and Practical Knowledge. In A. Ford, J. Hornsby, and F. Stoutland (Eds.) *Essays on Anscombe's* Intention. Harvard University Press.
Tooming, U. 2020. Being Familiar With What One Wants. *Pacific Philosophical Quarterly*, DOI: 10.1111/papq.12327
Turnbull, W., Carpendale J. I. M., & Racine, T. P. 2008. Relations between Mother-child Talk and 3- to 5-Year-Old Children's Understanding of Belief: Beyond Mental State Terms to Talk about the Mind. *Merrill-Palmer Quarterly*, 54(3): 367-385.
Valaris, M. 2017. What Reasoning Might Be. *Synthese*, 194: 2007-2024.
Van Woudenberg, R. & Kloosterboer, N. 2019. Three Transparency Principles Examined. *Journal of Philosophical Research*, 111-128. doi: 10.5840/jpr20191029147
Vega-Encabo, J. 2011. Self-Knowledge as Knowledge? *Teorema: Revista Internacional de Filosofía* 30(3): 35-49.
Verheggen, C. and Myers, R. 2016. *Davidson's Triangulation Argument*. Routledge.
Viedge, N. 2018. Defending Evidence-Resistant Beliefs. *Pacific Philosophical Quarterly*, 99: 517-537. DOI: 10.1111/papq.12174
Vierkant, T. & Paraskevaides, A. 2012. Mindshaping and the Intentional Control of the Mind. In F. Paglieri (Ed.), *Consciousness in Interaction: The Role of the Natural and Social Context in Shaping Consciousness*. Advances in Consciousness Research, 86, John Benjamins Publishing Company, 105-124.
Vierkant, T. 2012a. Self-knowledge and Knowing Other Minds: The Implicit/Explicit Distinction as a Tool in Understanding Theory of Mind. *British Journal of Developmental Psychology*, 30(1): 141-155.
Vierkant, T. 2012b. Managerial Control and Free Mental Agency. In A. Clark, J. Kiverstein, and T. Vierkant (Eds.), *Decomposing the Will*, 283-297. Oxford University Press.
Vierkant, T. 2013. What Metarepresentation is for. In Beran, M., Brandl, J., Perner, J., and Proust, J. *Foundations of Metacognition*, 279-288. Oxford University Press.
Walton, K. 1984. Transparent Pictures: On the Nature of Photographic Realism. *Critical Inquiry*, 11(2): 246-277.
Walton, K. 1997. On Pictures and Photographs: Objections Answered. In Allen, R. and Smith, M. (Eds.), *Film Theory and Philosophy*. Oxford University Press, 60-75.

Weatherson, B. Luminous Margins. *Australasian Journal of Philosophy*, 82(3): 373-383.
Wei, X. 2020. The Role of Pretense in the Process of Self-Deception. *Philosophical Explorations*, DOI: 10.1080/13869795.2020.1711960
Williams, B. 2002. *Truth and Truthfulness*. Princeton University Press.
Williamson, T. 2000. Knowledge and its Limits. Oxford University Press.
Wilson, T. 2009. Know Thyself. *Perspectives on Psychological Science*, 4: 384–389.
Wilson, J. 2010. What is Hume's Dictum, and Why Believe It? *Philosophy and Phenomenological Research*, 80(3): 595-637.
Wimmer, H., & Perner, J. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13: 103-128.
Winokur, B. ms. There is Something to the Authority Thesis.
Winokur, B. Forthcoming. Inference and Self-Knowledge. *Logos & Episteme*.
Winokur, B. 2021. Critical Reasoning and the Inferential Transparency Method. *Res Philosophica*, 98(1): 23-42.
Wittgenstein, L. 1953. *Philosophical Investigations*. Wiley-Blackwell. 4th Edition, Hacker & Schulte (Eds.), printed in 2009.
Wittgenstein, L. 1967. *Zettel*. Edited by Wright, C. & Anscombe, G. E. M. University of California Press.
Wright, C. 1992. *Truth and Objectivity*. Harvard University Press.
Wright, C. 1998. 'Self-Knowledge: The Wittgensteinian Legacy.' In C. Wright, B. Smith and C. Macdonald (Eds.), *Knowing Our Minds*, 13–45. Oxford University Press.
Wright, C. 2001. *Rails to Infinity*. Harvard University Press.
Wright, C. 2014. Comments on Paul Boghossian, "What is Inference". *Philosophical Studies*, 169: 27-37.
Wright, C. 2015. "Self-knowledge: The Reality of Privileged Access." In S. C. Goldberg (Ed.), *Externalism, Self-Knowledge, and Skepticism*, pp. 49–74. Cambridge University Press.
Zaitchik, D. 1991. Is Only Seeing Really Believing? Sources of the True Belief in the False Belief Task. *Cognitive Development*, 6(1): 91–103.
Zimmerman, A. 2006. Basic Self-Knowledge: Answering Peacocke's Criticisms of Constitutivism. *Philosophical Studies*, 128(2): 337-379.
Zimmerman, A. 2008. Self-Knowledge: Rationalism vs. Empiricism. *Philosophy Compass*, 3(2): 325-352.
Zimmerman, A. 2019. Belief and Commitment: Commentary on Annalisa Coliva, *The Varieties of Self-Knowledge*, Palgrave Macmillan (2016). *Philosophia*, 47(2): 335-342.