Essays in the Philosophy of Motivation, Normativity, and Self-knowledge

Matthew Kinakin
M.A.

Submitted in fulfilment of the requirements for the Degree of
Doctor of Philosophy

Philosophy
School of Humanities
College of Arts
University of Glasgow

September 2023

Abstract

This thesis comprises three distinct and substantial essays concerning motivation, normativity, and self-knowledge. Generally speaking, this thesis focuses on how agents can access particular facts about themselves—e.g., *why* they acted the way they acted or *that* a particular experience is bad-for-them. In particular, I show how the motivational and normative nature of certain mental states like our motivating reasons and unpleasant pains interact in unanticipated ways with foundational questions about the *nature of phenomenal consciousness*—"what it is like" to undergo this or that experience—with questions of *self-knowledge*—how we are aware of, and come to know, the mental states we're in—and with questions concerning *moral knowledge*—how we come to know moral facts.

Take, for instance, the action of taking a painkiller or the formation of a new belief. If I asked you *why* you took a painkiller or *why* you formed the belief, you would be able to, in an immediate and direct sense, *know* why you did those things. Not only do you know what your reasons are, but you seem to know your reasons in a *special* way, again, in a sense that is immediate and direct. Furthermore, some of our mental states seem to be *bad states to be in*. Take, again, your unpleasant pain: you want to end it *because* it's bad. So, it seems, then, that you also have knowledge of the badness of your unpleasant pain.

But these claims are not without controversy. For some deny that we know our own reasons in a special way distinct from how we know of other's reasons. The first aim of this thesis is to critically assess a recent attempt to defend the claim that we do have some special, direct, and immediate knowledge of our reasons. But insofar as we construe the *nature* of why we believe and do things—i.e., our reasons—in a specific way, I argue that defending that picture likely fails.

Furthermore, turning back to our *phenomenal experiences*, it turns out that once we understand the *nature* of our phenomenal experiences—e.g., unpleasant pains—along particular representationalist lines, explaining the motivational and normative facts implicit in our painkiller-taking action becomes increasingly implausible. The second aim of the thesis argues that insofar as we are concerned with

accommodating such motivational and normative facts, we ought to abandon a particular brand of representationalism about phenomenal consciousness.

And lastly, I argue that once we fully appreciate the normative profile of a certain set of our mental states—e.g., the *badness* of unpleasant pain and the *wrongness* of an intention to lie—we are in a position to develop a novel account of moral knowledge. Specifically, I motivate taking seriously the idea that we can introspect normative facts.

Table of Contents

Acknowledgements

I must thank, first and foremost, my two supervisors David Bain and Robert Cowan. I have undoubtedly benefited a great deal from both their immense intellectual powers as well as the generosity and thoughtfulness they both displayed throughout the last four years. Each has provided invaluable philosophical contributions to this thesis, and I am grateful for the many incisive and constructive comments and conversations they have provided me with. I am no doubt a better person for it.

From my time at The University of British Columbia as both an undergraduate and a master's student, I thank all the faculty who contributed to my intellectual and personal progress. During those formative years at UBC, I was fortunate enough to have had some remarkable teachers as well as academic mentors. In particular, I am indebted to Murat Aydede, Matt Bedke, Margaret Schabas, Ori Simchen, Christopher Stephens, and Evan Thompson. They have all, in various forms, instilled a great deal of inspiration in me. I am forever grateful.

I am extremely fortunate to have been surrounded by some wonderful people while in Glasgow. Thank you to Rory Aird, Han Edgoose, Patrick Hayes, James Humphries, Ross Patrizio, Fin Reid, and Joe Slater for all being great philosophical interlocutors as well as wonderful human beings. Thank you also to Adriana Alcaraz-Sanchez for her warm friendship and her willingness to listen to my half-baked ideas when they were in their infancy.

I have had the opportunity to develop what I consider my Glasgow family. Finn McCardel, thank you for your friendship and your deep appreciation for clarity. I will always cherish our Kelvingrove Park walks as well as the many "coffee breaks" which turned into dinners which often creeped into the wee hours of the morning. Thank you to Martin Miragoli for not only being wholeheartedly himself, but for extending his love and generosity my way as if I was one of his brothers. Thank you to Pinelopi Stylianopoulou for not only geeking-out with me about the history of analytic philosophy and helping me clarify whatever jumble was in my head, but also for being one of the kindest and coolest people I know. And thank you to Laura Fearnley for, of course, our cherished friendship (and her always-second-best dance moves). But thank you also for being a guiding light through this academic world and for giving me a standard to strive for.

I owe a very special thanks to Ewa Wozniak. There was a time when I wasn't sure how motivationally feasible it was to continue on in academia. Since I have met you, you have re-ignited my passion for the pursuit of knowledge and instilled within me a profound sense of curiosity toward the world around me for which I will be forever indebted. There is nothing quite like being in your presence. Thank you and I love you.

Thank you to my friends and family back in Canada. In particular, thank you to my mother and my sister for both giving me the space and encouragement to pursue my professional goals. The love and support I have received from them throughout my life has been tremendous and truly something to behold. It goes without saying, but I could not have done this without the two of them.

And finally, I would like to end by giving a special thanks to my brother-in law David Fedorko who passed away only a few months prior to submitting this thesis. David always took a keen interest in my academic ambitions and welcomed me into his life with a big old warm bear hug. He continued to warm me and everyone else with his presence, and his absence will forever reverberate throughout my life. This thesis is dedicated to his memory.

Author's Declaration

Author's Declaration I confirm that this thesis is my own work and that I have: (i) read and understood the University of Glasgow Statement on Plagiarism, (ii) clearly referenced, in both text and the bibliography or references, all sources used in the work; (iii) fully referenced (including page numbers) and used inverted commas for all text quoted from books, journals, web, etc.; (iv) provided the sources for all tables, figures, data, etc. that are not my own work; (v) not made use of the works of any other student(s) past or present without acknowledgement. This includes any of my own works, that has been previously, or concurrently, submitted for assessment, either at this or any other educational institution; (vi) not sought or used the services of any professional agencies to produce this work; (vii) in addition, I understand that any false claim in respect of this work will result in disciplinary action in accordance with University regulations. I declare I am aware of and understand the University's policy on plagiarism and I certify that this thesis is my own work, except where indicated by referencing, and that I followed the good academic practices noted above.

Introduction

This thesis comprises three distinct essays on motivation, normativity, and self-knowledge. The thesis is not a comprehensive survey of one particular philosophical topic, nor is it unified by the defense of some broader philosophical theory. Although each chapter is distinct, they are nonetheless connected by over-arching themes. The purpose of this introduction is to describe those themes and the ways in which they unfold throughout each chapter. What the thesis concerns most directly is the following: how agents can (or cannot) access particular facts about themselves, e.g., *why* they acted the way they acted or *that* a particular experience is bad-for-them. More precisely, I show how the *motivational* and *normative* nature of certain mental states, e.g. motivating reasons and pain, interact in unanticipated ways with questions of *self-knowledge,* the *nature of phenomenal consciousness,* and *moral knowledge*. Importantly, the arguments presented in each chapter are for the most part self-contained and can be read independently of one another: nothing I discuss in one chapter hangs on a discussion in another chapter. But, as mentioned above, there are nonetheless over-arching themes. To begin to see this more clearly, let us consider more closely a theme which more-or-less underpins all three essays: self-knowledge.

1 Self-knowledge

Generally speaking, all three essays concern self-knowledge in some form or another.[1] Typically, philosophers distinguish between two kinds of self-knowledge, only one of which piques their interest.[2] Consider two examples of what would at *first glance* rightfully be described as self-knowledge:

(1): I know that I was born in New Westminster, British Columbia, Canada.
(2): I know that I have a headache.

Let us stipulate that it is true, at this very moment, that I have a headache. Plausibly, each case describes something I know *about myself*. One is a fact about where I was born and the other is a fact about how my head feels. Philosophers, however, have historically been almost

[1] 'Introspection' might also be apt here, but for the time being I stick with the intentionally broader notion of 'self-knowledge'. Later on, it will become clearer how introspection fits in with the themes of this thesis and how it might be distinct from self-knowledge.
[2] Although, see Cassam (2015) for dissent.

exclusively interested in (2) rather than (1).[3] One key reason for this is that the *kind* of knowledge which typically underpins a case like 2 is *distinctive* in the sense that it is knowledge arrived at in *a special way* and perhaps enjoys some *high level of epistemic security*. Furthermore, the sense in which my knowledge of my headache is distinctive is set against a backdrop of the ways in which we ordinarily gain knowledge of the world around us. For say I want to know where *you* were born or know whether *you* have a headache. In order for me to know those things about you, I must consult evidence, make observations, rely on testimony, and perhaps draw inferences. This is important for it seems to shed light on why (1) is not distinctive enough to interest philosophers: although (1) is knowledge I have about myself, it would seem to be knowledge I arrive at via the *same methods* I would use to arrive at knowledge of facts about you, hence it is not very distinctive. Importantly, to know that I have a headache at this very moment, I do not consult evidence, I (almost) never rely on testimony, nor do I, at least not in the usual sense, make observations[4] about myself or my behaviour.

So, we can exclude (1) as an instance of what philosophers are interested in when discussing self-knowledge: it is not distinctive enough. It becomes, then, apparent that what philosophers typically mean when they talk about self-knowledge are the ways in which it is distinctive from *other*-knowledge. To know a fact about where I was born, I must rely on the same sorts of facts, evidence, and methods I would rely on to know facts about someone *else*. Not the case for my headaches. This is admittedly a little vague, but hopefully enough to mark off the general conceptual space with which this thesis is concerned. We can, then, call the sort of self-knowledge indicative of headaches *distinctive* self-knowledge.[5]

---

[3] Classic texts are Descartes' *Meditations* (1641[1984]), Locke's *An Essay Concerning Human Understanding* (1689[1975]), and Kant's *Critique of Pure Reason* (1781, 1787[1997]). As might be obvious, cases like (1) are not of interest to philosophers since they do not involve knowledge of one's own *mental states*. More on this below. Note that Kant's own views on self-knowledge are not entirely clear. For he seemed to embrace two distinct kinds of self-knowledge: one grounded in inner sense and one grounded in rational agency. See Brook (2006) for discussion as well as Boyle (2009) for a brief albeit illuminating defense of the latter view of Kant.

[4] Note that if we do make observations about ourselves to arrive at knowledge of our headaches it isn't observation in the normal, perceptual sense. And this observational form of self-knowledge would, intuitively, be something quite special and distinct from our knowledge of other people's mental states.

[5] In the literature, sometimes the term 'privileged' is used instead of 'distinctive' (see, e.g., Gertler 2003, esp. xi-xxii; 2011: 1-86; 2021). I prefer to use distinctive since there is a sense in which privileged self-knowledge captures (1) above, yet (1) isn't something we typically want to include in the category of self-knowledge, at least not in the philosophically relevant sense. As will become clearer in Chapter 1, I use the term 'privileged' to refer to the sense in which subjects enjoy a greater degree of epistemic certainty with respect to whether they are in a particular mental state relative to others' beliefs about whether they are in that mental state.

Another feature of distinctive self-knowledge is what it is about: *our own* mental states. (2) is clearly *not* about a mental state of mine whereas (1) clearly is. So, we can further restrict the sense of distinctive self-knowledge with which we are concerned to self-knowledge of our own mental states. A picture begins to emerge. Self-knowledge is distinctive with respect to our own mental states, and our mental states are the kinds of things we either stand in some special relation to or have a special method of knowing about, or both. As Brie Gertler claims:

> To elucidate the sense in which we are authoritative about our own mental states and to explain the grounds for this authority are the principal goals of most theories of self-knowledge. These theories are especially concerned with the idea that, in grasping one's own mental states, one uses a method that no one else can use—that is, an exclusively first-personal method. (2011: 3).

Two things emerge from this picture. We can break down what I above called *distinctive* self-knowledge into two components. First, we seem to be especially *authoritative* about our own mental states in the sense that we typically are not *mistaken* about them. Second, we seem to access our mental states in an "exclusively first-personal" way. This raises an admittedly pertinent question: what sorts of things do we enjoy "authoritative" and "exclusively first-personal" knowledge of?[6]

Another way of putting the question is to ask what is the *scope* of distinctive self-knowledge? As Brie Gertler claims, the scope of distinctive self-knowledge is "at most" restricted to three things: sensations, thoughts, and occurrent attitudes (Ibid.).[7] Note two paradigmatic examples (one of which we saw above) of distinctive self-knowledge. I walk through the park on my way home from work and my attention is caught by a distinct, pleasant smell. I notice it is coming from the gardenias to the left of the footpath. I lean over, give them a good sniff, and form the belief *that gardenias smell pleasant*. Here I seem to gain some self-knowledge:

> (3): I know that I believe that gardenias smell pleasant.

[6] I say more about this distinction in Chapter 1.

[7] For those who are skeptical of distinctive self-knowledge, see the works of Wittgenstein (1953) and Ryle (1949). For those who severely restrict the scope of what we have distinctive self-knowledge of see the works of Carruthers (2010), Dennett (1991), Medina (2006), and Schwitzgebel (2008). Note that Carruthers, Dennett, and Schwitzgebel are concerned with restricting the scope of what we have *introspective* access to. For further discussion see Gertler (2011).

After a night out of heavy-drinking, I wake-up the next morning with a terrible hangover and feel the unpleasantness of my splitting headache. Here I seem to gain some more self-knowledge:

(4): I know that my headache is unpleasant.

In each case, I know something about my mental life. I know *that I believe that gardenias smell pleasant*, and I know *that my headache is unpleasant*. My knowledge of my belief and of the unpleasantness of my headache are both excellent candidates for distinctive self-knowledge. Both are mental states I seem especially authoritative about, and both are mental states I seem to access in a special way.

Now, of course, these could be resisted as paradigmatic cases of distinctive self-knowledge (especially knowledge of our beliefs)[8], but I take it that there is fairly strong pre-theoretical support for the idea that they are. And anyways, this thesis is not meant give a comprehensive survey of which mental states are or are not indicative of distinctive self-knowledge. Rather the goal here is to draw our attention to some possibly *neglected* cases of distinctive self-knowledge. So, given that (3) and (4) are plausible candidates for distinctive self-knowledge, consider the following.

Notice something important about the mental states in (3) and (4). Plausibly, each mental state can take on a *motivational* and a *normative* dimension. For example, I believe that the gardenias in the park smell lovely, and because of this belief, decide to take the long way home from work. Or consider the unpleasantness of that splitting headache which *feels so bad* that I am motivated to take a painkiller. In both cases, not only are there facts about my mental life—e.g., facts about what I believe about gardenias and facts about how my head feels—but there are importantly facts about what *moves me* (e.g., my gardenia belief and my pain) and facts about what is *bad-for-me* (e.g., my pain). Not only are these further facts *about* my mental states, but they would seem to be facts that I am especially *authoritative* of and can access in a *distinctly first-personal way*. After all, if my fiancé pressed me about why I was home later than usual, I will *just know* what my motivation is: the smell of the gardenias. And similarly for my painkiller-taking action: the unpleasantness of my pain. It's not as though I draw any inferences or engage in a sort of

[8] See footnote 7.

detective work to know these motivational and normative facts. They are apparent to me in a way that is distinctive from other-knowledge, or so it would seem.

To be clear, we are moving away from focusing *solely* on any distinctive access we might have to mental states such as *beliefs themselves*—e.g., gardenia beliefs—or *affective experiences themselves*—e.g., unpleasant headaches—and focusing instead on the motivational and normative properties those mental states might have.[9] My gardenia belief, plausibly, can play a kind of *motivating role* vis-à-vis my actions: I walk through the park *because* of my belief that gardenias smell pleasant. My unpleasant headache, plausibly, can play not only a kind of motivating role vis-à-vis my painkiller-taking actions (it definitely does that), but also is distinctly *normative*. Its badness provides me with a reason to eliminate it, and plausibly that is something I have distinctive self-knowledge of.

So, in light of the above examples, I take it that there is strong intuitive support that we have some form of self-knowledge of our reasons for ϕ-ing and of the badness of some of our mental states:

(5): I know that my gardenia belief *is my reason* for taking the long way home.

(6): I know that my unpleasant pain *is bad*.

(5) and (6) both seem to meet our two rough-and-ready requirements for distinctive self-knowledge. They both seem to be instances of self-knowledge which we have a great deal of epistemic authority over: I seem to know what my reasons for acting and believing are better than anyone else; and the badness of *my* unpleasant pain is also something which I am very much the authority over. That is, I stand in some sort of privileged position with respect to my reasons and the normative properties of some of my mental states (e.g., unpleasant pain). Furthermore, (5) and (6) both seem to be instances of self-knowledge arrived at through a distinctly first-personal method: I know my reasons and the badness of my pain *not* from observing my own behaviour and drawing an inference (as perhaps someone else would from their third-person perspective) but rather from something more intimate, something more *immediate* to my consciousness. So, not only can we say that (5) and (6) are instances of self-knowledge, but that they are, in some sense or another, *distinctive*.

[9] Of course, how we access those mental states will inform how we might go about accessing (and explaining) the motivational and normative properties of those mental states. This is especially true for Chapter 2.

This thesis is, generally speaking, concerned with the various ways in which we can account for cases like (5) and (6) within some specified framework of distinctive self-knowledge (or, more particularly, introspection; more below). In particular, a tension arises between the intuitive support for cases like (5) and (6) and particular conceptions of *what reasons are* and *what phenomenal episodes are*. But to see how those tensions arise—and to get a better sense of what the first two chapters concern—we must briefly look at how we will be understanding distinctive self-knowledge in the first place.

2 Direct Non-inferential Access

Typically, although by no means universally, what makes distinctive self-knowledge so distinctive is *the way in which* we access our mental states.[10] Thinking, for a moment, about my access to my headache—case (2)—we can quite intuitively see that how I (typically) access the fact that I have a headache is fundamentally of a different kind than how you would access that same fact. Importantly, and as we saw echoed by Gertler, there is something first-personal about my access to my headache fact and it is this first-personal method of accessing that fact that is one of the key ways of characterizing distinctive self-knowledge.[11]

Recall that one way in which I characterized how case (5) and (6) qualified as first-personal self-knowledge was that our knowledge in those cases seems to be somewhat *immediate*. A further way of spelling that out is to say that the way in which we access some facts about our mental lives is in a manner that is importantly *direct* and *non-inferential*.[12] What exactly this entails will become clearer throughout the chapters, but for now it will suffice to point out that

[10] Here it might be worth introducing another term: namely, *introspection*. For introspection is typically used to mark off the distinct "first-personal method" we use to access our mental states and is typically seen as direct and non-inferential (Gertler 2021). The reader is welcome to think in terms of introspection here. For the time being, I eschew talk of introspection for the simple reason that my opponent in Chapter 1 does so as well. There I speak of distinctive self-knowledge and the *peculiar* way in which we access our mental states (rather than in terms of the way we introspect our mental states). The reason why my opponents eschew talk of introspection seems to boil down to the idea that they want to distance themselves from any 'observational' or 'inward looking' conception of self-knowledge. But in any case, in what follows, when I speak of distinctive self-knowledge and its direct, non-inferential nature, one is welcome to think of this in the broad sense of introspection. In Chapter 2 and 3 I speak entirely in terms of introspection rather than distinctive self-knowledge, mostly because the extant literature dealt with in those chapters speaks in those terms but also because of ease of exposition.

[11] There is no doubt more ways. See Gertler (2003: xii) for at least six different ways which, she rightly points out, need not all be present for a piece of self-knowledge to count as distinctive, or in her terminology, privileged.

[12] See, e.g., Armstrong (1963, 1993), Aydede (2003), Chalmers (2003), Fernandez (2013), Gertler (2001, 2012), and Peacocke (1999), to name but a few, for explicit endorsement of the idea that the manner in which we access some of our mental states must be direct and non-inferential. Note that I am not explicitly arguing for the claim that distinctive self-knowledge *must be* construed as directly non-inferential. Rather I am more concerned with how that popular assumption about distinctive self-knowledge interacts with other assumptions about the nature of motivating reasons, of normative properties of mental states, and of mental states themselves.

one key way of characterizing distinctive self-knowledge is that it is an instance of *direct non-inferential* knowledge. And this specification of distinctive self-knowledge as direct and non-inferential fits nicely with the picture we've painted so far; a key difference between how I know that I am having a headache and how you know I am having a headache seems to roughly track a distinction between inferential and non-inferential knowledge. There is at the very least no explicit inference made on my part and plausibly no tacit inferential support either: I am simply directly aware of the unpleasant pain of my headache. But there does seem to be some inferential process involved if you want to know whether I have a headache; observing me wince and grab my forehead, you might infer that I am undergoing some unpleasant pain in my head. Of course, I might *tell you* about my headache in which case your knowledge of my headache might plausibly be non-inferential[13] but importantly it will be too *indirect*; similarly to how my behaviour might stand in between you and my headache, my verbal report will stand in between you and my headache.

Again, this is rough-and-ready, but it serves an important purpose. Once we understand distinctive self-knowledge along these lines (which is by far the standard view), we can begin to see some tensions arise between the direct non-inferential conception of distinctive self-knowledge and cases like (5) and (6). Recall:

(5): I know that my gardenia belief *is my reason* for taking the long way home.

(6): I know that my unpleasant pain *is bad*.

Two tensions arise here in two very different ways. The first tension arises once we conceive of motivating reasons as *causal*. The second tension arises once we conceive of phenomenal states—like perceiving a red and round tomato or undergoing an unpleasant pain—as *representational states*. Each conception is in tension with the direct non-inferential nature of distinctive self-knowledge. This brings us to our third theme which we've already been introduced to: the motivational and normative features of mental states.

[13] Although this might depend on background beliefs about whether I am reliable, in which case your belief that I have a headache will depend for its justification on those antecedently held background beliefs.

3 The motivational and normative features of mental states

The first two chapters of the thesis deal directly with the apparent tensions alluded to above. To understand the first one, let us briefly consider what some people take motivating reasons to be.

3.1 Motivating reasons as causal explanations

In (5), my reason for taking the long way home is my putative belief that gardenias smell pleasant (perhaps also coupled with my desire to smell pleasant things on my walk home). This reason (e.g., this belief-desire pair)[14] *rationalizes* my action in the sense that it *makes sense* of my walking through the park from my perspective. Furthermore, a crucial fact about my gardenia belief is that not only does my belief rationalize my action, but it also *causes* my action: my reason for taking the long way home is my gardenia belief since it, in part, *causally explains* my action. Although this picture of the nature of what I will be calling 'motivating reasons'—the reasons *for which* we believe and act—is not without its detractors[15], it will be the assumed conception of a motivating reason throughout Chapter 1.

Insofar as we take a motivating reason to be a consideration which causally explains an action (or belief) like my walking through the park on my way home, a case like (5) becomes much less plausible as an instance of distinctive self-knowledge. The main reason for this—and what will be the main focus in Chapter 1—is that causal relations can only be known *inferentially*.[16] And since inferential knowledge is antithetical to the current conception of distinctive self-knowledge on offer here, we simply don't enjoy distinctive self-knowledge of our motivating reasons. So, a case like (5), although it enjoys some pre-theoretic plausibility, on a standard conception of what motivating reasons are, isn't a prime candidate for distinctive self-knowledge.

---

[14] In what follows, I drop any reference to the desire here.

[15] This picture of motivating reasons—what Davidson (1963) called 'primary' reasons—has been dubbed the 'Standard View' in the philosophy of action and in the reasons literature. As I mentioned, there are, of course, detractors. I can't begin to do justice to the vast literature on this topic, but for some fairly recent objections against the Standard View, see the collection of essays in Sandis (2009). In particular, see Sandis (2009: 1-9) for references of defenses of the Standard View as well as references of arguments against it throughout the 20th century. Those who typically argue against the Standard View don't deny that reasons are explanatory, but rather deny that they are a special kind of *causal* explanation. Insofar as one denies that motivating reasons are in part causal explanations, the tension focused on here between distinctive self-knowledge of motivating reasons and their causal nature will not arise. See Alvarez (2017) for further discussion of the various kinds of reasons, i.e., normative, motivating, and explanatory.

[16] See Hume (1772). See also Gertler (2011: 74) for this explicit argument.

This tension is the focal point of chapter 1. In short, I look at whether a recent attempt to reconcile the direct non-inferential conception of distinctive self-knowledge with the causal-explanatory conception of motivating reasons succeeds. I argue it does not.

3.2 Normative reasons and phenomenal states

Recall case (6):

> Case (6): I know that my unpleasant pain *is bad*.

Of course, when we take a painkiller to kill the unpleasantness of our pains, we are presumably acting for a reason—we have a motivating reason which is to end the unpleasantness and/or the badness of our pains. Now, if the argument about the causal nature of motivating reasons is correct, then we also fail to have distinctive self-knowledge of why we take painkillers. But there is a further tension which arises from this picture with respect to both the *kind* of normative and motivational story we can give and with respect to *how* we can further access the normative facts, and subsequently, rationally respond to them, i.e., the badness of unpleasant pain.

Some representationalist theories of phenomenal consciousness are committed to a particular picture of how we access our phenomenal states like unpleasant experiences or perceiving a red and round tomato. Strong representationalism, as I'll understand it here, is the view that sensory phenomenal episodes, like visually perceiving a red and round tomato, are identical and reducible to (exhausted by) the *representational contents* of experience. That is, the phenomenal character of experience—the what-it-is-likeness of experience, e.g., the "red-feelingness" or "reddishness" of perceiving the tomato—are identical and reducible to the representational contents of the experience. Furthermore, phenomenal character is *reducible* to representational content in the sense that phenomenal character is *explained in terms of* representational content. That is, our theory of representational content is more fundamental than phenomenal character.

Furthermore, strongly representationalist theories are committed to the claim that we *cannot directly attend* to our phenomenal experiences,[17] hence do not have direct non-inferential access to them, and that we must instead *infer* their existence from our prior awareness of what perceptually seems to be the case. In other words, our phenomenal experiences are *transparent*

[17] See, e.g., Harman (1990), Tye (1995b, 1996a 2003), Dretske (2003).

in the sense that our attention slips through any putative phenomenal character. Importantly, strong representationalism must wield an inferential picture of how we have distinctive access to our phenomenal states, a picture which is, of course, directly at odds with the picture of distinctive self-knowledge we have been considering here. And if we cannot directly attend to our own phenomenal episodes, then presumably we cannot directly attend to the badness of those episodes. So, it might simply occur to one to claim that we do not enjoy direct non-inferential self-knowledge of the badness of our phenomenal episodes since one (highly popular) view of phenomenal consciousness—strong representationalism—simply rules it out.

That would be a reasonable thing to conclude *if* the inferential picture representationalists put in place of the non-inferential one could adequately accommodate how we access the badness of our phenomenal episodes. But unfortunately it cannot. In fact, *not only* can it not accommodate our *access* to the badness of our phenomenal episodes, and hence our rational response to that badness, but strong representationalism cannot *explain* key motivational and normative *facts* about phenomenal episodes like unpleasant pain, at least not in a way that appeals to a highly attractive and naturalistic explanation: desire-based explanations.

This tension is the focal point of chapter 2. I expose a tension on the part of a number of philosophers between their representationalist conception of pains (and hence, their inferential account of how we access those pains) and their idea that the nature or badness or motivationality of pain's unpleasantness requires pains to be the targets of *experience-directed desires*. In short, what I suggest is required to accommodate those motivational and normative facts, is that we allow for the possibility that subjects have *direct non-inferential access* to their phenomenal episodes and the normative properties thereof; a kind of access that is directly at odds with the main tenets of strong representationalism.

3.3 A novel a posteriori intuitionism

Once we acknowledge that we can have direct non-inferential access to the badness of our unpleasant pains—i.e., that we can *introspect* evaluative properties—a novel possibility with respect to evaluative and moral knowledge opens up. This is the focus of Chapter 3. For there has been a recent insurgence with respect to the epistemological thesis Ethical Intuitionism:

> Ethical Intuitionism (EI): normal ethical agents have at least some non-inferentially justified first-order normative beliefs. (Cowan, 2013a).

One key way of explicating (EI) is to provide the naturalistic grounds for non-inferential justification of first-order normative beliefs.[18] But once we appreciate that some of our mental states have, broadly speaking, *normative properties*—e.g., the badness of unpleasantness— and that we can access those properties in a way that gives rise to the non-inferential justification of normative beliefs, we have the beginning of a novel a posteriori ethical intuitionism; one that is importantly rooted in our direct non-inferential introspective capacities: Introspective Intuitionism.

> Introspective Intuitionism (II): normal ethical agents can and do have non-inferentially justified first-order normative beliefs by having introspective states.

So far, there has yet to be an attempt to motivate a view like (II). In Chapter 3 I motivate the idea that we should begin to take seriously a view like (II). To anticipate, (II) depends on the truth of another view, what I call *normative introspection*:

> Normative Introspection (NISP)*:* at least some normative properties are introspectable.

Plausibly, we already see something like this with respect to our knowledge of the badness of our pains. And the most plausible naturalistic candidate for that badness seems to be introspection (rather than, say, emotion, intuition, or perception). So, Chapter 3 gives the first sustained treatment of the plausibility of (II) with respect to not only properties like badness but also deontic properties like the putative wrongness of an intention, e.g., to lie.

So, when looking at the following:

> (5): I know that my gardenia belief *is my reason* for taking the long way home.
>
> (6): I know that my unpleasant pain *is bad*.

[18] See, e.g., Cowan (2013a, 2013b), Cullison (2010), McBrayer (2010), Vayrynen (2008)

we can draw the following conclusions.

First, a case like (5) is not a prime candidate for distinctive self-knowledge because the causal nature of motivating reasons implicated in (5) precludes such reasons from being the objects of direct-non-inferential access. Since we are here concerned with distinctive self-knowledge construed along such direct non-inferential lines, we do not therefore have distinctive self-knowledge of our motivating reasons.

Second, a case like (6) seems to be a highly plausible candidate for direct non-inferential self-knowledge. However, that intuition comes under threat once we acknowledge that some accounts of the nature of phenomenal states—i.e., strong representationalism—preclude the possibility of direct non-inferential knowledge of phenomenal states. But importantly, it turns out that going in the other direction—i.e., the *indirect inferential* direction—with respect to our access of our phenomenal states proves problematic for not only accounting for a case like (6) but also for accommodating the motivational and normative features of phenomenal states themselves. Since strong representationalism is incompatible with our access to the badness of our unpleasant pains and with a highly attractive naturalistic explanation of the motivational and normative features of unpleasant pain, we should abandon strong representationalism about phenomenal consciousness.

Third, once we embrace the idea that we can directly, non-inferentially access some of the normative properties of our mental states, the conceptual space opens up for a novel a posteriori ethical intuitionism, one importantly grounded in our introspective capacities to access the normative properties of our own mental states.

Here, then, are the summaries of the chapters which will follow:

4 Chapter summaries

Chapter One:

In Chapter one of the thesis, I consider the possibility that we can have distinctive self-knowledge of our motivating reasons for belief and action. In particular, I criticize a recent attempt to argue *for* distinctive self-knowledge of motivating reasons given by Sophie Keeling (2019b). In the first part of the chapter, I present her argument against what she labels ‘the

orthodox position': the idea that we do not enjoy distinctive self-knowledge of our motivating reasons because they are causal in nature. The view under attack there is the idea that we have inferential self-knowledge of our motivating reasons. I subject Keeling's argument against an inferential account of our knowledge of our motivating reasons to a few criticisms. Keeling claims that when we engage in inference, we cannot apparently provide the putative justification which our motivating reasons are supposed to provide for the lower-order attitude (or action). I argue that Keeling's argument against the inferentialist fail. I run through various ways of interpreting her claim that engaging in inference fails to take into consideration whether something is a good reason (from the subject's perspective) for the lower-order attitude or action in question. I end the first part of Chapter 1 by claiming that an inferentialist account of our knowledge of our motivating reasons is in good standing with respect to the argument Keeling attempts to wield against it. In the second part of Chapter 1, I then present Keeling's own non-inferential account of how we gain self-knowledge of our motivating reasons, one she claims is grounded in our agentive abilities to answer 'world-directed questions' about what the good (normative) reasons are for believing or acting in a given circumstance. In short, I argue that Keeling's own non-inferential account of our knowledge of our motivating reasons does not clearly give us knowledge of the required causal relation needed to know about one's motivating reasons. At best, her account gives us knowledge of what subjects *judge* to be good (normative) reasons and not whether those judgements actually have the psychological purchase needed to be motivating reasons.

Chapter Two:

In chapter two of the thesis, I consider whether a popular theory of phenomenal consciousness—strong representationalism—can capture the motivational and normative conditions typically agreed to constrain a theory of unpleasant pain. In particular, I look to see whether strong representationalism can capture those conditions by appealing—as some strong representationalists do—to desire-based theories of motivation and normativity. For the standard account of the nature of *unpleasant* pain invokes experience-directed desires for the pain sensation to stop. But a highly popular alternative is strong representationalism, which, by contrast with the standard account, invokes *not* desires but *representational content* to capture the *phenomenology* of unpleasantness—the what-it-is-likeness of unpleasant experience. But here are two, somewhat neglected questions. What should it say about: i) the *badness* of unpleasantness and ii) our motivation for anti-unpleasantness behaviour (e.g., taking

painkillers)? Like the standard account, an attractive and highly plausible idea is to appeal to experience-directed desires for the unpleasantness to stop, in particular, *de re* desires, to explain these normative and motivational facts. In section two of the chapter, I argue that the required desires need to be *de re*. I then in section three present the theoretical commitments of strong representationalism which are incompatible with appealing to *de re* desires. Appealing to *de re* desires, I argue in section four, can't work for strong representationalism since the kind of experience-directed desires invoked are ruled out by strong representationalism's commitment to the transparency of experience, the idea, roughly: that subjects cannot directly attend to their phenomenal experiences. Hence, strong representationalism is incompatible with any view in value theory, metaethics, or moral psychology that will require *de re* desires to account for the normativity and motivationality of affect. In section five I raise some objections to my argument and give replies pointing out the difficulties which plague an inferential treatment of our access to our phenomenal states. In section six, I further elucidate the sense in which *even if* strong representationalists could explain the badness of unpleasant pain, they still struggle to accommodate our putative *access* to that badness, again, stressing the difficulties which arise from giving an inferential account of that access. In the final section, I consider where strong representationalism stands with respect to the motivationality and normativity of affect by briefly considering other, non-attitudinal-based explanations of affect's motivationality and normativity. In particular, I conclude that things look pretty grim for the strong representationalist interested in accommodating the motivational and normative features of affect.

Chapter Three:

In Chapter three of the thesis, I motivate taking seriously the claim that we enjoy direct non-inferential—i.e., introspective—access to the normative properties of our mental states. This is theoretically fruitful because it provides the novel grounds for an a posteriori ethical intuitionism. In the first three parts of the chapter, I spend some time delineating the scope of the paper, specifically by mentioning which sorts of normative properties and mental states are the most plausible candidates for introspection. In particular, I give some examples and draw on extant literature which help bolster the idea that normative properties (broadly construed to include evaluative and deontic properties) do in fact supervene mental states alone. In section four, I then motivate the idea that we have direct non-inferential introspective access not only to evaluative properties like the badness of our unpleasant pain, but also to deontic properties

like the wrongness of an intention to lie. Here I motivate what I call *normative introspection* by suggesting that we undergo particular phenomenological experiences with respect to the deontic properties of our mental states. And that introspection is at least as good as an explanation as others of our epistemic access to the putative wrongness presented in those moral experiences. In section five, I give a particular account of how introspection might plausibly provide non-inferential justification for first-order normative beliefs *given* that mental states are *never by themselves* morally right or wrong. Here I appeal to the affordance literature and suggest that even if mental states do not have deontic properties like wrongness, that they nonetheless can *afford* certain actions, and it is in virtue of the actions which they afford that they provide non-inferential justification for normative beliefs. More specifically, a mental state like an intention to lie can provide a negative affordance such as 'the intention to lie is not-to-be-acted-upon' or 'the-action of lying-is-not-to-be-done'. Notice that this is compatible with the idea that mental states are never by themselves morally right or wrong. In section six, I conclude by connecting normative introspection to introspective intuitionism (II). I consider some objections to the idea that (II) provides an epistemically independent source of justification for first-order normative beliefs. I conclude by defending (II) against those objections and by sketching how normative introspection and with it (II) might provide the grounds for the justification of our first-order normative beliefs about the *extra-mental* world.

Conclusion

That completes the introductory chapter. I now turn to Chapter one to look at whether we have direct non-inferential self-knowledge of our own motivating reasons.

CHAPTER 1: Transparency, non-inferentiality, and motivating reasons

1 Introduction

Consider the following:

1. Jack knows that he weighs 77 kilograms.
2. Sally knows that she has a headache.

The way in which Jack knows his weight and *the way in which* Sally knows she has a headache seem to be radically different. Although each piece of knowledge is about their respective self, only one, namely Sally's, is said to be an instance of what some philosophers call *distinctive self-knowledge*. Part of the self-knowledge theorist's job is to characterize what is, in some sense, *special* about Sally's knowledge relative to Jack's. One place in particular where Sally's and Jack's knowledge might be said to differ is in the *kind of access* one has to such facts. We might think that Jack's epistemic access to his weight is the same as other people's epistemic access. The fact of Jack's weight is available to Sally in the same way it is available to Jack; Sally and Jack will both have to use observational methods to determine Jack's weight, e.g., by using a scale, observing the number on the scale, and concluding that the number on the scale is how much Jack weighs. Whereas Sally's epistemic access to the fact that she has a headache seems somehow distinctive in a way that is only available to her (e.g., perhaps via introspection)[19]. That is not to say that other people cannot have knowledge of Sally's headache, but rather that the way in which other people gain such knowledge is restricted, e.g., they have to rely on Sally's verbal report or observations of Sally's physical states such as whether she winces at loud noises or rubs her forehead with some level of distress. Some philosophers, then, would say that Sally's epistemic access to the fact that she has a headache is *privileged* over other people's access to that very same fact.[20]

[19] Or at the very least, there is a way in which Sally can access that fact that is only available to her.
[20] It should be mentioned that not everyone thinks that self-knowledge is special in some way. See, for instance, Medina (2006). I bring this up to set it aside, as this paper takes it for granted that we have distinctive self-knowledge.

But notice that the description "privileged access" might equally apply to Jack's weight as well. After all, the most likely (epistemic) authority on Jack's weight will be Jack himself. Jack is presumably acquainted with information that other people aren't typically acquainted with. On this line of thought, it's not so much *how* one accesses certain information, but rather that one is suitably positioned with respect to some information. Although anyone can see what the scale reads, Jack is the one who stands in a privileged epistemic position with respect to *that* information. As Gilbert Ryle puts it, subjects will have greater "supplies of the requisite data" when it comes to data about themselves (1949: 155). We should then distinguish between what Alex Byrne (2019) has called *privileged access* on the one hand, and *peculiar access* on the other. For instance, although I can (perhaps must) use the same methods of discovery as other people to know that I become seriously irritable around young children, it might be said of me that I nonetheless enjoy a sort of privileged access to that fact since I am in a particularly privileged position with respect to the behavioural evidence needed to arrive at such conclusions. Knowledge of the conditions under which I become irritable is something I can have privileged access to. But importantly, it doesn't seem like something I can have *peculiar access* to, for again, I must use the same methods as anyone else to arrive at judgements about my own irritability.[21]

The case of Sally's headache seems importantly different. Sally's knowledge intuitively seems to be about a fact that she has *peculiar* and not just privileged access to[22]. She seems capable of using a method that only she herself can use and need not gather any sort of behavioural or perceptual evidence to arrive at such knowledge.

Consider, now, the following exchange: Jack asks Sally *why* she bought a plane ticket to Paris. She replies: 'To go on vacation.' What might we say about Sally's knowledge of *why* she bought

[21] Something which is seldom addressed is the status we should give to self-knowledge which is a mix between exploiting knowledge we have peculiar access to with knowledge we have mere privileged access to (although so called "self-interpretation accounts" (Gertler 2021) could be construed as something like a mixed account of self-knowledge. As it turns out, I think our access to our motivating reasons resembles something like a mixed picture of self-knowledge.

[22] As Byrne (2019: 10) points out, the two need not entail each other. Note also that the distinction between privileged and peculiar isn't always so sharp, and what follows is not intended to sharpen that boundary. But for clarity's sake, I take privileged access to track the extent to which one's beliefs about their own mental states are especially *secure* or prone to error; whereas peculiar access tracks the extent to which one's method is one that only they can use to arrive at beliefs about their own mental states. In what follows, I take peculiar access to be what is so distinctive of self-knowledge, but nothing I say here hangs on that fact. I expect a great deal of what we have peculiar access to to track what we have privileged access to.

a plane ticket? Intuitively, Sally is the *authority* on such matters; her belief seems especially *secure*; her self-ascriptions perhaps are even particularly *reliable*. So, Sally's self-knowledge of the reason for which she acts—i.e., her *motivating reason*—is plausibly something she has privileged access to. What about peculiar access? At first glance, it would seem that Sally's access to her motivating reason more closely resembles her access to her headache than it would her access to her weight. At least in the sense that Sally need not gather any evidence or infer her reasons from some prior information about her environment (e.g., there is no weight-scale analogue in the reasons case). The intuition, then, might be that our *epistemic access* to the reasons why we hold a particular attitude or performed a particular action more closely resembles our access to *facts about our headaches* than our access to facts about our height.

But the story isn't so straightforward. For one crucial difference between access to headache facts (H-facts) and access to reason facts (R-facts), is that the former comes with a rich phenomenological profile whereas the latter seems to lack any whatsoever, a profile which itself seems to directly present the H-fact. To speak loosely, headaches are plausibly known in a way that is more 'inwardly directed', grounded in a sort of 'inward gaze or attention' to our own internal mental world whereas that description seems wildly inappropriate for our access to R-facts. Although we seem to 'just know' both kinds of facts, the way in which we just know seems to be radically different. So although access to R-facts might more closely resemble access to facts about headaches (and hence, is a plausible candidate for distinctive self-knowledge), it is importantly different in the sense that there lacks any supposed *inward-directedness* and/or *phenomenology* which aids such access. So it would seem, then, that whatever explains our peculiar access to R-facts will be importantly different than our peculiar access to H-facts.

One highly plausible view which seems to nicely account for this difference, is expressed in the following Garth Evans quote about the self-knowledge of belief:

s

> [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: 225).

The rough and interesting idea is that unlike our access to H-facts (which, again, plausibly requires some inward-directed attention)[23], some of our distinctive self-knowledge, e.g., of our beliefs, is knowledge had on the basis of an answer to an *outward-directed* question. There is supposedly something in the answer to such questions that warrants us in self-ascribing mental states to ourselves on the basis of such answers. Setting aside whether such an account gives us privilieged access to our mental states, it does seem to give us *peculiar* access. For if I can access my mental states by answering outward-directed questions, then plausibly this is a process through which *only myself*, in the first-personal mode, can access such facts. After all, my answer to whether there will be a third world war is in no sense connected to what you might believe about the matter, *nor* should it be used as any sort of evidence for what you believe.[24] But there is something to the idea that my answer sheds light on what I do in fact believe. So, the Evans-style method seems like a plausible candidate to account for our peculiar access to mental facts which importantly do not seem to rely on inward-directed attention. Can it extend to our motivating reasons?

Recently, Sophie Keeling (2019a) has done just that. Her view is roughly the following: in order to arrive at distinctive self-knowledge of one's own motivating reasons—i.e., the reasons *for which* one acts and believes—one must first answer (like in the Evans case) an outward-directed question along the following lines: 'what are the good reasons to believe *p* or to Φ?' One then settles on an answer to the question, and thereby is warranted in self-ascribing the answer as one's motivating reason. For instance, in setting out to learn what my reason is for believing that there will be a third world war, I ask myself: what are the normative reasons to believe that there will be a third world war?' I settle on an answer: that Putin is a madman. And thereby com to gain self-knowledge of what my motivating reason is. There is supposedly *something in the answer,* or, more aptly, *in the act of giving an answer*, to such questions about the good (normative) reasons for believing *p* or Φ-ing that warrants one in self-ascribing motivating reasons to oneself on the basis of such answers. So, the seeming distinctiveness of our self-knowledge of our motivating reasons is captured by an account that piggybacks on the plausibility of an Evans-style transparency method of self-knowledge.

[23] Chapter two is dedicated to discussion of this claim.

[24] Of course, one *could* use one's own verbal or mental utterance as evidence of what someone else believes. But, as will become clearer later on, this isn't the sense in which concluding about some matter therefore entitles me to self-ascribe a belief on the basis of such a conclusion.

But some are skeptical about whether self-knowledge of our motivating reasons really is so distinctive to begin with. That is, some question whether self-knowledge of our motivating reasons enjoys any sort of privileged status (Nisbett and Wilson, 1977b) or whether it is even *in principle* something we can have peculiar access to (Gertler, 2011). We, then, have some choices. Either give up on the pre-theoretic idea that we enjoy distinctive self-knowledge of our motivating reasons, or somehow challenge that skepticism. Keeling herself is sensitive to these issues, and in fact goes to some lengths to argue in favour of distinctive self-knowledge of motivating reasons in light of the above worries. Let us first get the worries on the table. The first worry is underpinned by empirical literature (Nisbett and Wilson 1977b; Wilson 2002) which suggests that people are unreliable at knowing the reasons for which they've done something—i.e., at knowing their motivating reasons—and often times *confabulate* their motivating reasons. This undermines any privileged access we might have to our motivating reasons. The second worry is underpinned by the apparent *nature* of a motivating reason: motivating reasons are in part *causal explanations* of why some subject S believes *p* or why S Φ-ed. And as has been pointed out (Gertler, 2011), causal relations just aren't the sorts of things for which we can have distinctive self-knowledge of, since to access such causal facts, we would need to mobilize information and engage in inferential processes which are in no way peculiar to our first-personal position. This undermines any peculiar access we might have to our motivating reasons. Call the claim that motivating reason self-ascriptions aren't distinctive of self-knowledge the 'orthodox position' (OP) (Keeling, 2019a).

The aim of this paper is twofold. The first part of the paper criticizes Keeling's reasons for rejecting the orthodox position. According to Keeling, our knowledge of our motivating reasons cannot be inferential, since to give an inferential answer to why you believe *p* or why you Φ-ed would be to *disrespect* the sense in which that question demands not only an explanation but also a *justification* for believing *p* or for Φ-ing. And, so the thought goes, an account which displays such disrespect for the dual-role of the why-question cannot underpin our knowledge of our motivating reasons. Therefore, we can disregard any inferential account of our motivating reasons Here I claim that her argument against the orthodox position does not work. In short, Keeling fails to adequately prove that engaging in inference fails to 'take the why-question seriously', and hence fails to demonstrate that knowledge of our motivating reasons must be *purely* non-inferential.

The second part of the paper criticizes Keeling's own positive account of self-knowledge of our motivating reasons. Here I claim that her own *reasons transparency method* (RTM) fails to give us distinctive self-knowledge of our motivating reasons. Roughly, Keeling grounds the non-inferential justification of belief in our motivating reasons in awareness of our own agency: when we make outward-directed judgements of the sort mentioned above, i.e., when we employ her *reasons* transparency method, we are aware of making something the case, namely, our motivating reasons, and thereby are warranted in self-ascribing our motivating reasons on the basis of such Evans-style transparency questions. Keeling then sets out to support such putative warrant. I argue that Keeling doesn't give us sufficient reason to think that the kind of agency that she explicates grounds our non-inferential justification to believe that *p* is our motivating reason. In short, Keeling fails to give us reason to believe that our awareness is awareness of a *causal condition* being satisfied rather than awareness of what we judge to be a good reason to believe or Φ.

The paper proceeds as follows. First, I present Keeling's argument against the orthodox position, what she calls the 'dual role argument'. I then show that Keeling's argument against the orthodoxy does not succeed for reasons having to do with how she understands the practice of answering why-questions. Contra Keeling, I show that the orthodoxy *can* account for what she calls the 'dual role' of the question 'why?'. Second, I present Keeling's own RTM account of knowledge of our motivating reasons. I argue that Keeling's claims fail to account for her own explananda, namely, that such knowledge be *non-inferential* and *well-grounded*. At best, her account gives us knowledge of what subjects *judge* to be a good reason for holding an attitude or performing an action and not whether such judgements are causally efficacious. Subjects, therefore, only have *non-inferentially* arrived at *well-grounded* beliefs of such judgments.

2 Against the intuition that we have distinctive self-knowledge of motivating reasons

Before I critically assess how Keeling understands our distinctive self-knowledge of our motivating reasons, let me do three things. First, I want to clarify what is meant by *distinctive* self-knowledge, and second, I want to set out the orthodox position with respect to our access to our motivating reasons. In the next section, I set out some assumptions that will be crucial for understanding and analyzing Keeling's arguments.

2.1 Distinctive self-knowledge

As we saw in the opening paragraphs, *distinctive* self-knowledge is perhaps best understood as involving the following two characteristics: *privileged* access and *peculiar* access. Recall Sally and Jack. Sally's access to her headache is intuitively quite distinct from Jack's access to his weight. But that intuition is ambiguous between whether her access is privileged or peculiar. On the privileged side, Sally's access to her headache is highly epistemically secure. There doesn't seem to be a lot of sense to be made of the idea that Sally could be wrong about whether she is having a headache.[25] Compare, again, Jack's access to his weight. Even though we might take Jack to be the epistemic authority on his own weight, there is more sense to be made of the idea that he could easily be *mistaken* about such a thing. After all, weight tends to fluctuate quite a bit, and depending on how often he weighs himself and how closely he watches his diet, Jack's access to his weight might not be so epistemically secure, at least relative to Sally's access to her headache. As Byrne (2019) points out, privileged access comes in degrees. So, we will enjoy privileged access to some fact (or kinds of facts) to the extent that our beliefs about those facts are especially reliable or especially epistemically secure.

Peculiar access is, as we saw, a different story and concerns *the manner in which* we access a particular fact (or kinds of facts). Sally seems to have a peculiar kind of access to her headache, one that only she can employ—it's not as though anyone else can 'look into her mind' and directly access her headache. This is obviously not the case with Jack's weight. Jack and everyone else must use the same methods to access facts about Jack's weight. For instance, it is plausible that at some future point in Jack's life he no longer is the epistemic authority over what he weighs, for, having undergone some serious health issues as of late and subsequently had to attend many doctor's appointments and many weigh-ins, Jack's doctor has now become the epistemic authority on the matter. A nice explanation of this possibility seems to be underpinned by the idea that Jack does not enjoy any kind of peculiar access to that fact; his access is just like anyone else's, and hence his epistemic authority is more precarious than Sally's is over her headache.

[25] Of course, there are cases we could conjure up. Perhaps Sally is a neuroscientist in the making and is undergoing a scan of her own brain. She looks at the produced images and concludes (mistakenly) on the basis of the images that she must be having a headache, even though she feels no such thing. Such cases will be rare and deviate wildly from the norm. Importantly, they do not undermine the claim that we tend to display a high degree of reliability and security in introspective judgements about our some of our mental states.

When it comes to our access to the reasons for which we believe and the reasons for which we act—henceforth our *motivating reasons*—the idea that we enjoy privileged and peculiar access (henceforth PVA and PCA, respectively) lends itself quite naturally.[26] But, as noted in the introduction, some call this intuitive pre-theoretical idea into question. In what follows, I present two broad arguments that undermine *both* the PVA and PCA claims about self-knowledge of motivating reasons.[27].

2.2 The orthodox position (OP)

Contrary to first appearances, however, there seems to be some agreement amongst philosophers and scientists on the fact that we *lack* distinctive self-knowledge of our motivating reasons.[28] I will follow Keeling (2019b) in calling this the orthodox position (OP). Here I consider two arguments for the (OP), one intended to undermine PVA to motivating reasons and the other PCA to motivating reasons.

The (OP) claims that people do not have distinctive self-knowledge of their motivating reasons. The two main reasons for thinking this are that i) people are particularly bad (unreliable) at identifying their motivating reasons, and distinctive self-knowledge is standardly presumed to be an especially reliable kind of knowledge, and ii) motivating reasons are best understood as causal explanations, and causal relations are standardly thought not to be the sorts of things we can have peculiar access to. Call these arguments 'The Confabulation Argument' and 'The Knowledge of Causation Argument', respectively.

---

[26] See, e.g., Davidson (1963: 633) and more recently Boyle (2011b: 2).

[27] In what follows, I tend to drop the 'self-knowledge' part for ease of exposition except when I think it's needed. So, one can read PVA and PCA of motivating reasons as 'our privileged access to our motivating reasons' and 'our peculiar access to our motivating reasons'.

[28] See Gertler (2011: 72-75), Nichols and Stitch (2003), Nisbett and Wilson (1977), Rey (2008), Schwitzgebel (2016), and Wilson (2002). See also Byrne (2019: 13) who briefly mentions it in the context of distinguishing between distinctive self-knowledge of *attitudes* and the *explanations* of those attitudes. He claims: "If this [Nisbett and Wilson's] experimental work is correct, it chiefly impugns a subject's explanations for her attitudes, beliefs, or behaviour, not her attributions of mental states." See also Keeling (2018) whose view I will present later on, but who herself points out the rarity with which this topic is seriously discussed in the philosophical literature. A rare exception is Cox (2018).

2.2.1 The Confabulation Argument

The confabulation argument supports the (OP) by appealing to empirical considerations that suggest that people are particularly bad at tracking the causes of their attitudes and actions. Widespread error in tracking the causes of our own attitudes and actions suggests that "we lack distinctive self-knowledge of why [we] hold [our] attitudes" (Ibid.). Nisbett and Wilson (1977)[29] suggest that the causal sources of our actions are particularly opaque. According to their study, subjects were asked, upon being presented with four pairs of stockings (identical in quality), to select the one that was of the highest quality. The study suggested that the position of the stockings greatly affected the choice; the left-most stocking was selected the least often, the second left-most stocking selected the second least often, and the right-most stocking was selected the most. When asked *why* they selected the stockings they selected, the subjects, according to Nisbett and Wilson, "engaged in a *post-hoc* rationalization of their preferences, citing factors such as the superior sheerness or elasticity of the pair they chose." What the subjects cite as their motivating reason—the thing that *causally explains* their attitude—seems to be plainly made up. In other words, the subjects engage in *confabulation* when asked to give the reason for why they chose the way they did.[30]

Confabulation cases are meant to motivate the general thought that what on the surface seem to be reliable self-ascriptions—i.e., *why* we hold certain attitudes and act in specific ways—are in fact *not reliable*, and that this spells trouble for the claim that we have distinctive self-knowledge of our motivating reasons. In particular, confabulation cases undermine the claim that we enjoy PVA to our motivating reasons.

> **Confabulation Argument (CA) against PVA**
>
> **P1** If we are especially unreliable at detecting our motivating reasons, then we lack privileged access to our motivating reasons.
>
> **P2** We are especially unreliable at detecting our motivating reasons.
>
> **C** We lack privileged access to our motivating reasons.

[29] See also Nisbett and Wilson (1978) and Wilson (2002).

[30] For a similar argument for the case of self-ignorance see Haidt (2001). In the cases Haidt is concerned with subjects are unable to account for why they hold a target attitude (importantly different than cases where subjects are in error about why they hold their attitudes). But I take it that self-ignorance cases also undermine the idea that we are especially reliable at detecting the reasons for which we have attitudes and perform actions. See Keeling (2018: 64-65, esp. fn. 2 and 3) for more references of empirical literature that undermines the reliability of our knowledge of the explanations of our attitudes and actions. I leave them to the side for purposes of space.

Premise one is just the contrapositive of PVA. Premise two is supported by the confabulation case. Therefore, we lack PVA to motivating reasons.

2.2.2 The Knowledge of Causation Argument (KCA)

The KCA gets its most explicit endorsement in Gertler (2011: 71-75). To see the argument, consider the following quote from Wilson (2002):

> My decision to get up off the couch and get something to eat, for example, feels very much like a consciously willed action, because right before standing up I had the conscious thought "A bowl of cereal with strawberries sure would taste good right now." It is possible however that my desire to eat arose nonconsciously and caused my conscious thought about cereal and my trip to the kitchen. (47)[31]

As Gertler herself notes, it seems highly plausible that subjects enjoy something like privileged access (where 'privileged' in her language means 'peculiar' in ours) to our occurrent thoughts. For instance, it seems highly plausible that a subject would have a highly reliable and peculiar way of accessing their own conscious thoughts about what they would like to eat at a given moment.[32] The structure of Wilson's thought seems to be something like this. Although I may enjoy PVA and PCA with respect to my occurrent thoughts, there is always a separate question whether the occurrent thought that preceded my action or my attitude did in fact cause the action or attitude. And to properly gain knowledge about that latter fact would seem to entail a further step in thinking—i.e., *an inference*—from one's occurrent conscious thought to the belief that it was their motivating reason—i.e., that it really was the cause of one's attitude or action.[33] Why this is problematic for distinctive self-knowledge is that it undermines the claim that we have PCA to motivating reasons. Although Gertler isn't explicit about the distinction between PVA and PCA (she often runs them together), she is to some extent aware of it. She claims:

---

[31] Also quoted in Gertler (2011, p. 73).

[32] Although even a claim like this can be questioned, at least the peculiar access part. See Gertler (2021, Sect. 3.3) where she discusses 'self-interpretation accounts' of distinctive self-knowledge (e.g., Lawlor, 2009; Carruthers, 2011; Cassam, 2015). Such accounts would construe one's belief that they have a particular desire as itself an interpretation based on more basic sensory data.

[33] What that story involves need not concern us here. But I take it that it would involve, at the very least, relying on some further background belief about the connection between one's past thoughts and judgements with the putative formation of further beliefs and the performance of actions succeeding those thoughts and judgements.

> In making this inference about the causal sources of my action, I draw on facts which I have privileged [peculiar] access—namely, the timing of my thoughts about food. But while my use of an exclusively first-personal method allows me to know that a certain thought is present, it does not allow me to know *the causal features of my thoughts*. In determining the domain of privileged access, we are concerned with knowledge directly produced by an exclusively first-personal method, or knowledge that achieves an especially high level of epistemic security. Facts that can be known only by combining such knowledge with knowledge that does not meet these conditions do not fall within the domain of privilieged [peculiar] access. So we appear to lack privileged [peculiar] access to the fact that a certain action or choice was caused by a particular mental state. (74; my emphasis).

Gertler herself seems to think that these facts undermine both PVA and PCA to motivating reasons. But for our purposes, we should focus on the latter: namely, the claim that we do not have PCA to motivating. After all, figuring out what my motivating reason is will involve employing a method which is equally available to anyone else. Here's a reconstruction of her argument:[34]

> **Knowledge of causation argument (KCA) against PCA**
>
> **P1** What explains an attitude is what caused it.
>
> **P2** Learning of causes must involve inference.
>
> **P3** *Peculiar* access (self-knowledge) doesn't involve inference.
>
> **C** Subjects lack *peculiar* access (self-knowledge) of what explains their attitudes.[35]

The assumption in the background here is that motivating reasons are themselves *causal explanations* of some target item, in our case, either an attitude or an action. The classic endorsement of this view comes from Davidson (1963), and more or less has been the orthodoxy for the last few decades.[36] Therefore, we lack PCA to motivating reasons. I take all three

[34] This argument comes from Keeling (2019a) but is slightly different since she does not distinguish between PVA and PCA.

[35] This argument leaves open the possibility that we might enjoy privileged without peculiar access to our motivating reasons, although CA above might call that into question.

[36] An important exception is Anscombe (2000) and more recently Setiya (2013). There has been an increasing tendency to view motivating reasons as non-causal explanations. Of course, it might occur to one to simply reject the causal account of motivating reasons in light of the seeming inconsistency between that account and peculiar access. This will depend on many things, e.g., how strong one's intuition is that we have peculiar access to our motivating reasons and what one's preferred theory of self-knowledge is. In any case, denying the causal role of motivating reasons is itself highly controversial and likely not seen as a welcome consequence of one's view.

premises to be fairly uncontroversial.[37] Now, one could attack this argument from many directions[38], but since the aim of this paper is to analyse an attempt to undermine premise three, I will leave things as they are.

2.3 Summary

In summary, we are too unreliable with respect to our knowledge of our own motivating reasons to enjoy anything like PVA to motivating reasons. The confabulation case supports this. Furthermore, motivating reasons are in part causal explanations of first-order attitudes and actions. And since to learn of causal properties is to engage in inference, and inference isn't a peculiar method of learning about ourselves, we lack PCA to motivating reasons.

In the next section, I present Sophie Keeling's recent attempt to dispel this picture, what I above called the orthodox position (OP). Her argument comes in two parts. First, she argues that providing anything like an inference when pressed about one's motivating reasons fails to respect an essential part of what that question asks for: *a justification* of one's attitude or action. Second, having dismissed the (OP) as plausible, she then provides her own positive account of how we might have non-inferential justification for belief of our motivating reasons, hence peculiar access to our motivating reasons.[39]

3 Preliminaries for Keeling's account

As should be obvious by now, I won't be taking a stand on what if anything best characterizes distinctive self-knowledge. My aim here is merely to consider in depth a novel view about how we might have *distinctive self-knowledge of our motivating reasons*, and Keeling (2019b) presents us with such a thing. Moving forward, the paper will comprise three parts: Sect. III will

[37] That is not to say that they cannot be challenged (see fn. 36 above). Premise three has recently been challenged by Alex Byrne (2019). His account is complex and won't be discussed in detail here. But roughly, he claims that we have peculiar self-knowledge of our first-order attitudes like belief by following an inferential rule like the following: 'If *p*, then believe that you believe *p*'. Byrne claims that following this will be both highly epistemically reliable (since it's self-verifying) and gives us peculiar access to our first-order attitudes (since it only works in the first-personal case). Interestingly, Byrne's account is inspired by the Evan's-style transparency method mentioned in the introduction to this chapter which also inspires Keeling's reasons transparency account mentioned below. There is an interesting question here which will not be pursued about the exact difference between Byrne's preferred transparency account and Keeling's.

[38] See fn. 36 and 37 for examples of claims that go against premise one and three, respectively.

[39] Note that Keeling is interested in the peculiar access claim and not so much the privileged access claim about self-knowledge. Later on, accepting the confabulation data and hence the claim that we are unreliable at knowing our motivating reasons poses problems for Keeling's view.

detail Keeling's argument for why we should reject the (OP), and some replies I give to her argument; Sect. IV will detail Keeling's preferred *reasons transparency account* (RTM) to capture knowledge of our motivating reasons. And Sect. V will raise objections and give some replies to her RTM account. But before that, it'll be important to set out some assumptions which Keeling herself accepts. Here, then, are seven preliminary remarks about motivating reasons. I do not endorse everything that follows and realize some of the remarks made here are controversial. But, after all, these are not my assumptions.

3.1 Causal explanation

Note first that the term motivating reason is a term of art. I won't venture to explicate it, but a few things should be said about it. First, as we saw above, motivating reasons will be understood as *causal*. That is, a motivating reason is, to speak vaguely, a consideration which causes a target attitude or action. For instance, my reason for taking the long way home is to smell the gardenias in the park. Were the gardenias not there (or had they not smelled so lovely) I would not have taken the long way home. Or, perhaps more appropriately, had I not *believed* that the gardenias were there, I would not have taken the long way home.

3.2 Ontologically neutral

Second, and relatedly, Keeling (2019b: 2) assumes that she can remain fairly neutral with respect to the ontology of motivating reasons when giving her argument. That is, she assumes that it won't matter whether one is a *psychologist* (e.g., Davidson, 1963; Smith, 1994; Turri, 2009) according to which motivating reasons are *mental states* or pairs of mental states, a *factualist* (e.g., Alvarez, 2010; Bittner, 2001) according to which motivating reasons are *facts*, a *disjunctivist* (e.g., Hornsby, 2008; Hyman, 1999) according to which motivating reasons are different kinds of things depending on the case at hand (e.g., a fact in non-error cases and a mental state in error cases), or a *propositoinalist* (e.g., Singh, 2019) according to which motivating reasons are propositions.[40]

[40] Note here that I have my doubts about this assumption. For it's unclear how motivating reasons could have any prima facie plausibility for peculiar access given that they were only *facts.* Of course, if motivating reasons were *psychological facts,* then perhaps that idea would be more plausible, but as far as I know, factualists take motivating reasons to almost always be *non-mental facts* (obvious exceptions being pain and other affective experiences). My speculation is that Keeling takes it that motivating reasons must be *the same kind of thing* as normative reasons, and since normative reasons are almost unquestionably facts, then it follows naturally that one would take

3.3 Rationalizing (justifying) explanation

Third, motivating reasons *rationalize* behaviour. It's not that the reasons for which a subject acts provide *mere* or *pure* causal explanations, but rather that they provide an explanation which *makes sense* from the perspective of the subject. My taking the long way home is not only *causally* explained in virtue of my appeal to the smell of the gardenias, but also *rationalized* in light of that supposed fact. One can see how taking the long way home is a *reasonable* thing to do in light of the fact that one enjoys the smell of the gardenias. The way Keeling thinks of this fact is that our motivating reasons count as potential *justifications* for our attitudes and actions, regardless of whether they actually do constitute good (normative reasons), i.e., reasons which really do count in favour of some attitude or action.

3.4 Epistemic and practical symmetry

Fourth, Keeling is mainly concerned with motivating reasons in the epistemic case, rather than the practical case. That is, she is mainly interested in the reasons for which an agent holds a particular belief. For instance, Sally believes that it is raining *for the reason* that there are grey clouds. Keeling's interest here lies in one's knowledge of the motivating reason for the belief that there are grey clouds. She, however, believes that her account of the self-knowledge of motivating reasons cuts across the epistemic/practical divide[41]. So although most of her examples are given in the epistemic case, they equally apply to the practical cases, where, for instance, Sally goes to the grocery store *for the reason* that they, say, sell mangos.

3.5 Basing relation

---

motivating reasons to be facts as well. But the assumption that each kind of reason must be the same kind of thing—'the identity thesis'—has come under attack. See Mantel (2014) who explicitly denies this identity thesis. For a defence of the idea that motivating reasons are propositions in the context of denying the identity thesis see Mantel (2017) and Singh (2019) as well as discussion of the relationship between motivating reasons and normative reasons.

[41] I won't be concerned with this, but one might question the extent to which the account Keeling gives extends adequately to the practical domain.

Fifth, Keeling notes that, at least in the epistemic realm, to have a motivating reason for some belief *q* is to *base* one's belief on some prior state.[42] Keeling endorses the following basing relation:

> **Basing Relation**
> *S's belief that q is based on her belief that p only if:*
>
> (1) the belief that *p* causally sustains the belief that *q*, and
> (2) S is disposed to take *p* to be a good reason for believing that *q*.

(1) should be required if we are trying to give an account of how we can have non-inferential access to something that is partly a causal relation. After all, it's assumed by the (OP). (2) seems to capture the sense in which motivating reasons *rationalize* actions from the subject's perspective; being disposed to take something to count on favour of a belief or action is a plausible candidate for this rationalizing role.

3.6 Well-grounded belief

Sixth, Keeling, along with many immersed in the self-knowledge literature, understands self-ascriptions that are typical of self-knowledge as "instances of *well-grounded* belief" as compared to produced from a merely reliable belief forming process. The beliefs about one's motivating reasons will *make sense* to the subject in the sense that they know very well *why* they made the self-ascription. Keeling doesn't go into details with respect to this condition, but she does say the following:

> "But supposing for a moment that reliable belief-formation suffices for warrant, it nevertheless does not suffice for *well-grounded* belief. Self-ascriptions, or at least self-ascriptions of the sort that are candidates for distinctive self-knowledge, seem to be instances of well-grounded belief. After all, self-ascribing the given motivating reason seems a *sensible* thing to do by the subject's lights" (15).

This will be important for what Keeling says later on when giving her positive account, for the kind of warrant she provides must account for the well-groundedness of such self-ascriptions.

3.7 Non-inferentiality

[42] Also, see Kurt Sylvan (2016a, 2016b) for discussion on the basing relation. Keeling thinks much of what is said about the epistemic basing relation can carry over to the practical domain.

Finally, and what is in some ways at the centre of this paper, is the notion of *inference*. As Keeling and many others claim, inference "is inimical to distinctive self-knowledge" (Ibid., 7). Again, by distinctive here we should read it as *peculiar access*, in this particular context: *direct non-inferential access*. The thought, roughly presented above, is that to use inference is to use "the same method that anyone else would use", and subsequently to strip self-knowledge of its characteristically *distinctive* status[43]. So, what Keeling is ultimately going to attack is premise two of KCA above: learning of causes must involve inference.

3.8 Summary

So, the following constraints should be clear.

> **Peculiar (direct non-inferential)[44] self-knowledge of motivating reasons:**
> *S has peculiar self-knowledge that their belief that p is their motivating reason for belief q only if:*[45]
> (a) S believes that *p.*
> (b) The belief that *p causally sustains* the belief that *q.*
> (c) S is disposed to take *p* to be a good reason for believing that *q.*
> (d) S has non-inferential justification for the belief that the belief that *p* causally sustains the belief that *q*
> (e) S's belief in (d) is well-grounded.

For instance, Sally believes that *it will rain* for the reason that *there are grey clouds*. According to the above, for Sally to have peculiar self-knowledge of her motivating reason: i) she must believe that there are grey clouds, ii) the belief that there are grey clouds must causally sustain the belief that it will rain, iii) she must be disposed to take there being grey clouds to be a good reason for believing that it will rain, iv) she must have non-inferential justification for believing that her belief that there are grey clouds causally sustains her believing that it will rain, and v) she must have her latter second-order belief be well-grounded.

---

[43] As Moran puts it (but by no means endorses it) one will have "the best-informed view of the facts in question" (2001, p. 120).
[44] In what follows, I for the most part drop the "direct non-inferential" bit.
[45] I put things in terms of an 'only if' for simplicity. One could add more constraints to arrive at a full-fledged bi-conditional, but since the main task here is to see how a supposed non-inferential account can accommodate access to a condition like (b) above, the other constraints are unnecessary.

In contrast, here is what Keeling says about the (OP):

> So the orthodoxy goes, both self and other-knowledge of motivating reasons involves some sort of inference […] Sally learns that she believes that *it will rain* for the reason that there are grey clouds because she engages in the following inference: a standard reason for believing that *it will rain* is thinking that there are grey clouds; I am a standard thinker; therefore, I believe that it will rain for the reason that there are grey clouds. (Ibid., 5-6).[46]

We can put it in the following standard form:

> **Standard Inference (SI):**
>
> **P1:** A standard reason for believing that *it will rain* is thinking that there are grey clouds
>
> **P2:** I am a standard thinker.
>
> **C:** Therefore, my reason for believing that it will rain is my thinking that there are grey clouds.[47]

Keeling's main task is twofold: first, she has to argue against why we shouldn't take too seriously the (OP) and the kind of inference she thinks is indicative of that position (SI). Second, Keeling has to give a plausible account of (d) above: defend the idea that we can learn of causal relations non-inferentially. In what follows, I present Keeling's attempt to dispel any attraction to the (OP). In doing so, Keeling takes herself to present further explananda on an account of knowledge of our motivating reasons, ones she claims the (OP) importantly fails to accommodate. After that, and after having raised some worries with her argument, I present her attempt to explain how we might non-inferentially learn of causal relations.

[46] She gets this kind of example from Cassam (2014). I have doubts about whether this is the best way to interpret someone who is sympathetic to giving an inferentialist account of our self-knowledge of our motivating reasons. In Sect. IV I present what I take to be the best (and more natural) inference that an inferentialist is likely to give of our motivating reasons.

[47] This is slightly different than Keeling's elaboration, but one which I think is more natural. I will provide a better inference later on.

4 Against the orthodox position: the dual role argument

Keeling sets out to reject the (OP) on the grounds that it cannot account for what she calls 'the dual role of the question 'why?''. She points out that when subjects are pressed about why they did something or why they believe something—i.e., when they are pressed about what their motivating reasons are—they tend to provide not only an explanation but also a putative justification for their attitude and action. With this in mind, Keeling (Ibid: 8) presents us with this argument:

> **The Dual Role Argument (DRA)**
>
> **P1***:* The question 'why?' has a dual role (either when posed by others or ourselves). The question at once requests an explanation and justification for the attitude.
>
> **P2***:* Given the characterisation of the question 'why?' in *P1*, we wouldn't be taking the question seriously if we used inference to answer it.
>
> **P3***:* We take the question 'why?' seriously.
>
> ---
>
> *Conclusion:* Our answers to the question 'why?' are not inferential.

In what follows, I want to unpack how exactly Keeling thinks DRA plausibly dispels an inferentialist conception of self-knowledge of our motivating reasons. I ultimately think it fails, and therefore, that an inferentialist picture is still on the table as a plausible account of the self-knowledge of our motivating reasons.

4.1 Unpacking the argument: take 1.

Although I think premise one can be resisted, I take premise one to be fairly intuitive. It will, however, serve useful to focus on what Keeling says in support of it to further spell out why Keeling believes our answers to why-questions are not inferential. Keeling says two things in support of premise one. First, she claims that we don't merely engage with people's answers to why-questions at the level of explanation but we also engage at the level of *justification*. That is, it is natural for us to offer "a countervailing defeater" or "undercutting" evidence for their answers (Ibid.). This is supposed to be evidence of the justificatory role of why-questions. Second, when subjects flout the request for explanation and justification they seem to be "doing something odd" and that such odd cases prove the rule of the dual role of the question 'why?'. For instance, rather than have Sally answer the question 'why believe that it will rain?' by citing

the grey clouds she sees or her belief that there are grey clouds, suppose she instead responds with: "because the perceptual mechanism detected patterns in the sky and processed them so as to result in a state of belief" (Ibid.). According to Keeling, the "peculiarity" of such an answer stems from the fact that it does not provide a *justification* for the first-order attitude, and that such answers are rightfully *criticisable* (Ibid., 9). We might first wonder whether Sally's answer in such cases is really that *peculiar*, and whether given its peculiarity, it should be explained in virtue of not providing justification. But more importantly, we might also wonder what connection, if any, exists between these kinds of answers and the use of inference. I'll come back these questions shortly.

Premise two needs further support and is the more contentious premise. But, as should be somewhat familiar from the opening remarks, our access to our motivating reasons does seem pretty seamless. It seems to be the norm in such contexts that we don't *need* to consult any evidence or make any further steps in our thinking to know why we believe something or why we've done something, at least not typically.[48] So, insofar as we take on board these norms, and acknowledge that any deviation away from those norms is odd or peculiar, and furthermore that these odd answers must either be inferential or fail to address one of the two roles of why-questions, premise two seems *somewhat* plausible. But why think we wouldn't be "taking the question seriously" if we used *inference* to answer it? Or to put it another way, why doesn't the use of inference provide answers that are explanations *and* justifications for first-order attitudes? Keeling has surprisingly little to say other than that "an inferential mechanism is not a good method to use when *providing justification for the lower-order attitude*" and that "[m]erely considering the evidence about what one's motivating reasons are *does not take into account whether they are good reasons*." (10; my emphasis). So, Keeling seems to think that the use of inference fails to capture the justificatory aspect of the why-question. And that it's this failure which ultimately provides the main support for the thought that inference is inimical to the self-knowledge of motivating reasons: inference precludes subjects from providing a seeming justification for their first-order attitude.

[48] This norm of behaviour might be the explanation for why we are so unreliable at knowing our motivating reasons.

Premise three I take it is simply the datum of common sense that we do seamlessly answer such why-questions in a "serious" manner.[49] This claim coupled with premise two entails the conclusion that we don't (and shouldn't?) use inference to arrive at distinctive self-knowledge of our motivating reasons.

There is one main question which needs to be addressed:

(1) how does an inferential mechanism fail to provide justification for lower-order attitudes (and actions)?

I think there are two ways to unpack (1). First, we can focus on the relation between Keeling's paradigm case of an inferential answer (SI) and Keeling's example of a peculiar answer which fails to engage with the justificatory role of why-questions. Second, and what I will deal with in the next sub-section, is the idea that subjects, when providing an inference, *necessarily* cannot identify with their putative reason *as a good one*. That is, inference does not leave room for the subject to identify with the motivating reason as something *they view as counting in favour* of some target belief or action.

Let's focus for a moment on Keeling's support for premise one. We're provided with an answer that fails, in certain respects, to 'take the question seriously'. One plausible reason we might think that inferential answers to why-questions fail to take the question seriously is because of the fact that *inferential answers only give purely causal explanations* like the one expressed when Sally utters the following:

> **Causal answer**: because the perceptual mechanism detected patterns in the sky and processed them so as to result in a state of belief.

[49] We might equally wonder whether non-inferential answers take the why-question seriously in the sense that they provide a *causal* explanation. An alternative might be that we must non-inferentially give our justification for believing what we believe but nonetheless engage in inference with that prior knowledge to know whether it causally explains our belief (I present something like this view below). It would then need to be explained why it doesn't seem peculiar to just immediately give your reasons when pressed about them. One explanation is just to say that we care more about the why-question norm which focuses on the demand to provide a justification rather than an explanation. When one engages in inference, one is simply flouting that norm in virtue of trying to satisfy another, namely, the norm to provide a causal explanation. This seems compatible with and perhaps even explains the empirical evidence that we are unreliable at knowing the causes of our attitudes and actions.

If causal answers provide no putative justification for the first-order attitude *and* inferential answers entail these causal premises, then it is plausible that inferential answers fail to provide the requisite justification. But Keeling hasn't provided any reason to think as much. In fact, the standard inference she gave us in the beginning is an instance of an inference *without* any causal premise. Recall:

> **Standard Inference (SI):**
> **P1:** A standard reason for believing that *it will rain* is thinking that there are grey clouds
> **P2:** I am a standard thinker.
> **C:** Therefore, my reason for believing that it will rain is my thinking that there are grey clouds

There is no obvious causal premise in (SI). Perhaps P2 is supposed to play that role—being a standard thinker *causes* one to believe for a specific reason. But the answer it produces isn't purely causal. And it's unclear why containing *one* causal premise apparently fails to provide a justification. Isn't premise two providing that support? P1 above seems to offer something by way of a justification. After all, in engaging in a genuine inference, one must *endorse the content* of the premises which would include endorsing the thought that grey clouds count in favour of the proposition *that it will rain*.[50]

The problem, I think, is more serious for Keeling. We can substitute part of the premise in "a standard reason for believing that *it will rain* is thinking that there are grey clouds" for a *causal proposition*. We get, with some tinkering:

> **Standard Causal Inference (SCI)**
> **P1:** [A] standard reason for believing that *it will rain* is *that the perceptual mechanism detected patterns in the sky and processed them so as to result in a state of belief.*
> **P2:** I am a standard thinker.
> **C:** Therefore, I believe that it will rain for the reason that *the perceptual mechanism detected patterns in the sky and processed them so as to result in a state of belief.*

[50] I come back to this point in the next subsection.

Here is an inferential answer with *two* apparently purely causal premises (let's suppose). What is wrong with this inference? The purported peculiarity of citing one's perceptual mechanism as a reason for belief, I take it, comes from, again, its failure to offer *a justification* for the attitude or action. So, we can see how *if* one thought that this is how the structure of an inference *must go* when answering why-questions, that it would then be plausible to conclude that inferential answers fail to provide justification. But even in this case I think this conclusion can be resisted. Doesn't citing perceptual mechanisms and pointing out the fact that they *'detect'* clouds lend itself as something of a *justification* for the attitude? As Keeling points out, citing perceptual mechanisms might be fine in scientific contexts. But surely scientific answers are appropriate *outside* of scientific contexts, and not just restricted to them. We might even thank someone for providing a more scientific answer in non-scientific contexts, however initially odd it might be. If someone gave me the conclusion of (SCI) above as an answer to a why-question about their belief that it will rain, I might initially think it odd, but would ultimately understand why they believed that: "oh, it's because they see grey clouds".

Furthermore, once we see this structure, two things become apparent. Our apparent "non-inferential" answers might be *elliptical* for two things: i) a purely causal answer, and ii) an inference. When Sally replies quite automatically to the question why she believes it will rain by saying '*that there are grey clouds'* this plausibly is elliptical for something like the following: because I am having a *perceptual experience* as of there being grey clouds. Now, I take it this answer is kosher for Keeling. It seems to provide a *causal* explanation and gives a *justification* for why one would believe that it's going to rain. After all, it seems to nicely rationalize one's first-order attitude from the perspective of the subject. But the answer about one's perceptual experience could plausibly be elliptical for the answer citing a perceptual *mechanism* in which case it's unclear why a mere *physical elaboration* of one's perceptual experience of grey clouds precludes the possibility of it giving a rationalization of one's belief. Moreover, notice that Sally's answer, although provided in an automatic and non-deliberative manner, need not preclude that the answer is *inferential*. Although when she gives the answer to her interlocutor, she does not *engage in inference*, the answer nonetheless is elliptical for these other answers. And importantly, I don't see how these answers fail to capture the request for justification of one's attitudes and actions.

Perhaps intuitions surrounding such answers differ. Even so, granted that we find such 'purely causal', or rather, 'scientific' answers peculiar by no means entails that they fail to provide a justificatory answer. Surely there are other explanations as to the peculiarity of such answers that do not exclude them as justifications. To explain the peculiarity of Sally's answer we might appeal to the simple fact that people do not tend to cite such processes when giving their answers to why-questions, or that the context does not make such an answer expected. And this need not entail that Sally's answer *lacks* justification for her belief that it will rain. In such a context, it would make sense to ask Sally 'what *patterns* did your perceptual mechanisms detect?'. Her reply could very well be 'my perceptual mechanisms detected *grey clouds in the sky*', and from there we can engage with her answer at the level of justification, not merely explanation, by offering 'countervailing defeaters' or by 'undercutting' her claim; perhaps by reminding her that she recently ingested a psychedelic substance and has good reason not to trust her perceptual faculties. And this engagement with the answer *as* a justification is, according to Keeling, indicative of capturing the dual role of the question 'why?'

Three things emerge from this discussion. First, it's unclear that what Keeling calls 'purely causal answers' fail to provide a putative justification for a first-order attitude. Second, it's perhaps even more unclear what the connection is between causal answers and inference such that the failure of the former to provide justification for first-order attitudes somehow infects the latter with that same failure. Third, a more general worry about the dual role argument concerns the suggestion that we can disregard a philosophical view about how we come to know our motivating reasons based on what it means to 'take a question seriously', that is, about what does and does not seem *odd* in social contexts. In other words: why think that the *form* our answers take to why-questions gives us evidence for what is *required* to know what is being asked? Surely there is a distinction between the *form* my answer takes and *the epistemic status of my answer*. Let me explain**.** For instance**,** I say (upon observing a vapour trail in a cloud chamber): "there goes a proton". You say: "why do you believe that?" I reply: "because I can *see* the vapour trail". Seeing a vapour trail, and simply giving this as my response to the question 'why?' seems non-inferential in the sense Keeling is concerned with. I do not engage in any inference. But even if the *answer* I give is non-inferential, the justification for the belief that *there goes a proton* will be importantly *inferential*. What explains and *justifies* the belief that *there goes a proton*? Presumably, *a whole set of background beliefs* about vapour trails and protons which ultimately provide the epistemic justification for the belief that there is a proton.

These background beliefs might *cognitively penetrate* my experience such that they alter the content of what my perceptual experience comes to represent[51]: it represents a proton which then allows me to non-inferentially believe—in the sense that I need not *perform an actual inference*—that there is a proton there. Or perhaps I merely have a more 'low-level' perceptual experience of colours and shapes and movement and it's this experience in conjunction with my background beliefs about protons, vapour trails, and cloud chambers which combine to produce my proton belief automatically and implicitly. In any case, the justification for the proton belief will ultimately be *mediated* by those background beliefs even though, again, I may not explicate this inferential support.

Note that the *form* one's answer can take in the proton case is, let's say, non-inferential in the sense that one does not *perform* an explicit inference in front of one's interlocutor. They do not, for instance, say: I see a vapour trail; vapour trails are signs of protons; therefore, I believe *that* is a proton. Instead, they automatically form the belief. But the *justification* for the belief *that is a proton* is importantly *dependent* on one's prior background beliefs. What this example shows is that although our answers might be automatic and non-inferential in the psychological sense, they might very well be *epistemically dependent*. We give our belief non-inferentially which depends for its justification on prior background beliefs. That is, how we are ultimately *justified* in holding the proton belief is because we have justification for other background beliefs (e.g., justification for the belief that vapour trails in cloud chambers reliably indicates the presence of protons). Crucially, Keeling's dual role argument does not show that our beliefs about our motivating reasons are non-inferential in this latter sense.

So, even though an answer to a why-question might take the *form* of a non-inferential answer (perhaps it's even the norm) that does not mean that the epistemic support for that belief is non-inferential. So, when we ask Sally why she believes it is going to rain and she replies, "that there are some grey clouds nearby" even though her answer here is non-inferential in form—she doesn't make an explicit *psychological inference*—it is nonetheless possible that what ultimately *justifies her belief* that her judgement that there are grey clouds is her motivating reason is some prior background belief (e.g., beliefs about judgements causally effecting one's attitudes). Keeling hasn't given us reason to think that this isn't the case. As the preceding

[51] For discussion of cognitive penetration see Plylyshyn (1999), Siegel (2012), McPherson (2012), and Cowan (2013a, 2013b, 2015).

discussion shows, Sally's answer "because there are grey clouds" might just be elliptical or a placeholder for something more causal or even inferential. Why, we might ask, would performing that inference be so problematic?

Importantly, a lesson we can draw from this is that *if* one's justification for a belief is epistemically *mediate*[52] in the sense that it depends for its justification on the justification of prior background beliefs, then there *should be nothing wrong* with making that epistemic support *explicit*. In other words, there should be nothing epistemically wrong with *psychologically inferring* one's belief that *p given that* the belief that *p* is epistemically dependent on backgrounds beliefs, say, belief that *q* and belief that *r*. So, our beliefs about our *motivating reasons* shouldn't be, according to Keeling, epistemically mediate.

So, I think it follows that *if* performing an explicit inference to one's belief about what their motivating reasons are is epistemically problematic (as Keeling suggests via the dual role argument), then it must be the case that the epistemic support for that belief *is not* dependent on the justification of prior background beliefs. Otherwise, it would be baffling why performing an inference is so problematic to begin with.[53]

4.1.1 Summary

In any case, it doesn't look as though Keeling has adequately discounted the (OP), i.e., the orthodox position whereby in order to know what my motivating reason is I must "engage" in inference or perhaps have background beliefs in place which epistemically support my judgements about what my motivating reason is. In the next section, I try to further unpack Keeling's argument and see whether better sense can be made of it. I will then present an improvement to the inferentialist picture Keeling presented us with; one I think is quite intuitive and clearly avoids the worries raised by Keeling herself.

4.2 Unpacking the argument: take 2

Recall:

> ***P2****:* Given the characterisation of the question 'why?' in *P1*, we wouldn't be taking the question seriously if we used inference to answer it.

[52] For an in-depth discussion of epistemic mediacy see Cowan (2013b)
[53] This will be important to keep in mind for Sect. 5

Keeling then claims:

> "An inferential mechanism is not a good method to use when *providing justification for the lower-order attitude* [and that] [m]erely considering the evidence about what one's motivating reasons are *does not take into account whether they are good reasons*." (Ibid.)

Now recall our standard inference (the one Keeling thinks is indicative of the (OP)):

> **Standard Inference (SI):**
>
> **P1:** A standard reason for believing that *it will rain* is thinking that there are grey clouds
>
> **P2:** I am a standard thinker.
>
> **C:** Therefore, my reason for believing that it will rain is my thinking that there are grey clouds

As mentioned above, P1 seems to "provide justification for the lower-order attitude". But something (SI) importantly leaves out is whether the subject engaging in (SI) *takes the grey clouds* to *count in favour of* the proposition that it will rain. It doesn't show that, from the subject's perspective, they actually take there to be *good (normative) reason* to hold the belief. In other words, it cannot provide a rationalization from their perspective. Rather the belief that their motivating reason has something to do with the fact that there are grey clouds has nothing to do with what the subject actually *endorses* as being a good reason. Above I claimed that a subject who genuinely performs (SI) must *endorse the content* of the inference; they must endorse P1, and that seems like it will entail endorsing that thinking there are grey clouds counts in favour of believing it will rain. But importantly, that isn't quite true since the inference doesn't mention good reasons but rather 'standard' reasons. What I take Keeling to mean when she says that inference cannot provide *a justification* for the lower-order attitude is that when a subject engages in inference, they cannot view their self-ascribed motivating reason as something *they actually endorse*, that is, as something they see as providing justification for the lower-order attitude. And that's how inference fails to respect the dual-role of the why-question. It can't capture the sense in which *from the subject's perspective* they take *q* to count in favour of *p*. (SI) leaves it open whether Sally endorses the supposed fact that there are grey clouds as a good reason for believing that it will rain.

I think this is a better interpretation of her argument and one that has some past textual support.[54] But I still think it is open to objections. If we are to defend the (OP), we need to provide an inference which shows that the subject endorses that some proposition *q* counts in favour of *p*—e.g., that there are grey clouds counts in favour of believing that it will rain. I think one can be given. Consider the following inference:

Endorsement Inference (EI)

E1: I judge that *q* is a good (normative) reason for believing *p*. **[N-judgement]**

E2: When I make an N-judgement, my N-judgement tends to cause my believing *p*. [empirical premise]

C: Therefore, my N-judgement is my motivating reason. [M-judgement]

Here we have an inference, I argue, which captures *both* the sense in which a subject *endorses* some proposition as counting in favour of another—what I call one's *'N-judgement'—and* the sense in which such endorsements can be causally efficacious. Note that if pressed about my reason for believing *p* I could simply reply '*q'*. That is an answer which, according to Keeling, respects the dual-role of the why-question *even though* it might very well *epistemically depend* on my background belief, namely, the background belief expressed in E2, a belief which importantly provides the support for thinking that my N-judgements do in fact causally sustain, and hence explain the first-order attitudes they concern. I don't see in what sense this inference fails to provide a justification, from the subject's own perspective, for the lower-order attitude. Hence, I think we have a case of inferring our motivating reason which does provide justification for the lower-order attitude.

Keeling might respond as follows. She might say that it's one thing for someone to *endorse their N-judgement* and another thing entirely for a *subject to judge* that *q* is a good reason, i.e., *make* an N-judgement. The point here is that when subjects consider premises and make explicit inferences, what they are judging in that case is not whether *q* itself is a good reason, but rather whether the premise—in our case the N-judgement *itself*—is a good reason, and in such cases, *q* itself is not being evaluated as a good reason; the *judgement* that *q* is a good reason is being

[54] See Keeling (2018). There she is explicitly interested in accounting for our attitudes and reasons in terms of the connection they bear to our rational agency, as something we in part *determine*.

evaluated as a good reason in support of the conclusion that it is one's motivating reason. In other words, in engaging in an inference, we are shifting our focus from an endorsement of the content [that *q is a good reason for p*] to the content [*I judge* [that *q is a good reason for p*]]. Although our N-judgement has as its propositional content [that *q* is a good reason for *p*], the judgement one makes when making the inference—call this our *second-order judgement*—does not itself include any *direct* endorsement about whether *q* is a good reason for *p*. Rather the second-order judgement one makes when making the relevant inference takes our N-judgement *as itself a good* reason in support of the claim that *q* is one's motivating reason. Keeling's point, I take it, is that when we make an inference like the one above, we cannot *at the same time* take our N-judgement to support C *and* make the judgement that *q* is a good reason for *p*: failing to make the latter judgement is a failure to provide what one takes to be "justification for the lower-order attitude". If, on the other hand, one merely makes the *judgement* that *q* is a good reason to believe *p*—simply makes the N-judgement—without making any further inference[55], then necessarily (or so I take Keeling to imply this) one has "taken into account whether they [some considerations/propositions] are good reasons".

Perhaps the point can be put as follows. In finding out that *q* is your motivating reason for belief *p*, you must somehow *endorse* that *q* is a good reason to believing that *p*. Inferences of the kind mentioned above apparently *do not involve* such endorsements. The inference from E1 and E2 to C involves *a different kind of endorsement*, namely, an endorsement about one's *judgements about q being a good reason to believe p*—what I above called a second-order judgement. The relevant endorsement needs to be first-order and not second-order, something along the lines of *I judge that p is a good reason to believe q* and not *I judge that 'my judgement that q is a good reason to believe p' is a good reason*. And it's this distancing from one's first-order judgement which entails that one no longer provides justification for the lower-order attitude. Rather it's the mental state itself that acts as non-inferential grounds for the self-ascription of the motivating reason. So, although my second-order judgement about my N-judgement is *about* my judging that *q* is a good reason, endorsing my N-judgement itself as part of the inference doesn't itself involve my endorsing that *q* is a good reason, i.e., evaluating *q* as a good reason.

[55] One might naturally wonder here how we then are able to arrive at a conclusion about a causal relation. This is Keeling's task in what will come below.

But why think that shifting focus from the content of one's N-judgement—that *q* is a good reason fo*r p*—to the content of the second-order judgement which contains not only the N-judgement content but the N-judgement *itself* entails that one no longer endorses the content of the N-judgement? It's not as though we lose our sense of what we take to be a good reason for *p* when we become aware of that judgement; it's not clear why endorsing our N-judgement as something that counts in favour of believing that it is our motivating reason precludes the N-judgement from retaining its normative endorsement. I don't see any good reason to think that I can't make an N-judgement, then use my knowledge of my N-judgement *as a premise* in a bit of reasoning. In fact, we can even concede a point to Keeling: namely, that providing a *purely* inferential method to arrive at the justificatory feature of a motivating reason—what one *takes to be a normative reason*—cannot provide the relevant justification from the subject's perspective. But importantly, providing what one takes to be a justification for a lower-order attitude or an action is not the whole story of what is involved in knowing one's reasons. There is a separate question whether that first-order endorsement—again, the N-judgement—does in fact have the psychological purchase needed to be one's motivating reason.

We can grant that I non-inferentially know my N-judgement, and hence can answer the justificatory part of the why-question non-inferentially: it is just my N-judgment! And perhaps I have also given you a *causal explanation* by merely giving you my simple answer. But importantly, whether that answer is *justified* will depend on whether I am justified in believing that my N-judgements cause my lower-order attitudes and actions. And explicitly engaging in that inference does not preclude any sort of knowledge of what my motivating reason is. I might even, metaphorically speaking or perhaps quite literally *mentally point* to my N-judgement and say, "*that's* my motivating reason!" and I might be justified in making that judgement because I have justification to think that it—my N-judgement—occupies the right causal role with respect to a target belief or action of mine.

Thinking about the practical case for a moment,[56] I tell you my reason for taking the long way home: the smell of the gardenias. You can press me about how exactly I know that's my reason. I then perform the following inference:

[56] Recall, Keeling thinks her argument applies to both the epistemic and practical domain.

G1: I judge that a good reason to take the long way home is to enjoy the smell of the gardenias.

G2: When I judge that something is a good reason for something else, that judgement typically becomes my motivating reason.

C: My judgement that the smell of the gardenias is a good reason to take the long way home is my motivating reason.

Notice I could have just said: 'I enjoy the smell of the gardenias' without performing the inference. But importantly, performing this inference seems epistemically kosher. In fact, it seems like a fairly good inference, so long as one has evidence for G2. But for our purposes, note that there is nothing puzzling about this inference; it is not as though our interlocutor would continue to press us and say: "yes, but what about your *justification* for taking the long way home". Him and I will equally understand what I am providing when I explicate the inference: an explanation *and* a justification.

I want to end this section by briefly mentioning a point which gets brought up in the context of self-knowledge. In particular, it is an objection which some, who go by the label 'agentialists'[57], mount against those who think that self-knowledge can be inferential. So, it is a possible objection against the view I have defended here, and it typically claims that inferential self-knowledge leads to 'alienated' self-knowledge (Moran 2001; Boyle 2011b, 2019). The idea is a similar one to the one we characterized Keeling as endorsing above. I don't want to spend too much time on it but want to bring it up in order to say a few things with respect to alienation and motivating reasons. Here is Matthew Boyle (2011b) expressing what he takes to be the fundamental difference between his agentialist (what he calls *reflectivist*) approach to self-knowledge and other *epistemic* approaches:

> "The reflectivist rejects an explanatory demand that many theorists of self-knowledge accept. He 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, and consequently denies that the task of a theory of self-knowledge is to explain how these conditions come to stand in a relation that makes the latter knowledge of the former…Reflectivists reject all versions of the epistemic approach. This does not mean that they offer no account of how self-

[57] For uses of this term, see, e.g., Gertler (2021), Keeling (2018), and Sorgiovanni (2020). Typical agentialists are Moran (2001, 2003, 2012) and Boyle (2011a, 2011b, 2015, 2019).

> knowledge is possible. They offer a different sort of account, one that is primarily metaphysical rather than epistemological… The reflectivist's task is to explain the nature of various mental states in a way that clarifies why their existence implies that their subject has tacit knowledge of them, and what this tacit knowledge can amount to." (235).

So, according to agentialism, or at least Boyle's characterization of it, the mere having of a given mental state M will entail that one is in a position to know that one has M simply because of the very nature of the state itself. Boyle continues to elucidate agentialism by claiming that some mental states are states which we *determine* through our rational capacities. Boyle explicates this further in the case of belief:

> Briefly, the idea would be that, for a rational creature, believing *P* just is being in a condition of actively *holding P* to be true. The self-determined character of this condition is especially *evident* where a person consciously considers whether *P* and 'makes up his mind'… One indication of this is that we expect a rational subject who believes *P* to be able to address the question what convinces him that *P* is true, whether he has consciously deliberated or not. A person will not, of course, always have specific grounds for holding a given belief, but the interesting thing is that, even when he admits to lacking grounds, *he accepts the presupposition of the question*—that he is in a position to speak for whatever grounds he has…What these observations suggest, I think, is that *all* our normal, non-alienated beliefs are, in a perfectly good sense, acts of our capacity to make up our minds. They are all enduring actualizations of our power to evaluate propositions as true, in the light of such grounds as we deem relevant. This evaluation is not an act one performs to *produce* a belief in oneself; it is one's belief itself." (236; his italics).

Boyle is concerned with belief, but I think the point can carry over to motivating reasons. Our self-knowledge of our motivating reasons will in part be constituted by some *knowing act* we perform—in our case, the power to knowingly evaluate some proposition to be true, e.g., that *q* is a good reason for believing *p*. And the knowing act is, importantly, not something which we merely do to "produce" a motivating reason in us but *is the motivating reason itself*. So, we make a judgement of the sort we saw above: I judge that *q* is a good reason for believing *p* and it is in virtue of this capacity to 'make up my mind' which entitles me to the knowledge that that judgement is my reason.

The point, I take it, that agentialists think is key to self-knowledge is that we do not employ any distinctive *epistemic powers* in coming to learn of our mental states. Rather our access to our

attitudes (and our motivating reasons) is importantly determined through our rational capacities to make up our minds. Our epistemic *entitlement* to self-knowledge simply falls out of that metaphysical fact: in determining an attitude in virtue of exercising our rational capacities we thereby come to know of our attitudes and our reasons. Exercising my rational agency with respect to a proposition *p*—i.e., determining whether *p* is true in light of my sense of the reasons in favour of *p*—is *constitutive* of what it is to actively believe *p* and further to believe for a reason. And since I am aware of what I am doing while exercising *that* capacity, I am thereby aware of what I believe: it's the very thing I am doing when exercising my rational capacities, i.e., knowingly evaluating that *q* counts in favour of *p*. To think otherwise would be to fail to see one's beliefs, and hence reasons as in some sense 'up to them'. In other words, it would be to see one's belief and reasons as mere objects of discovery and not as something expressive of one's rational stance on the world.

For our purposes, to engage in inference is to call that picture into question. By *not* taking our N-judgement at face value—i.e., as expressive of what our reason actually *is*—we are somehow undermining or turning away from an important fact about ourselves: that we are rational creatures who get to determine our attitudes through the mobilization of our reasons. When we fail to take that story at face value (e.g., when we engage in inference and see our reasons not as something determined by us but rather as *discoverable* by us), we thereby take a perspective on ourselves—an alienated one—which necessarily precludes the more fundamental one: the picture which paints us a rational beings.

But what that picture fails to appreciate is that not only are we rational creatures who attempt to determine our attitudes by mobilizing our reasons, but that we are also *empirical* beings whose attempts at rational determination are subject to a kind of failure on behalf of that fact—the empirical one. It's no less a denial of what we are capable of—rational self-determination—then it is an acceptance of our situation in the world—a situation which means that our attempts to express the one nature might fail because of the other.

In other words, there is nothing problematic about having a sense of what one takes to be a reason for something else—i.e., a sense of what one *views* as a normative reason—and at the same time wondering whether one's sense of what one views as a normative reason ends up

playing the right causal role required to have one's attitudes determined by that sense. We can, that is, wonder about the causal bit while simultaneously engaging with one's sense of the normative reasons as items the subject is responsible for, as items that can be criticized, and as items which express some particular agentive perspective on the world. It is a separate question entirely whether those items have any psychological purchase.

4.3. Summary

We saw that Keeling's dual-role argument can be resisted. That is, there is no good reason to outright dismiss the idea that our knowledge of our motivating reasons is inferential. I went through various interpretations of Keeling's argument, and we found them all wanting. I then pressed a line of attack from agentialists who claim that engaging in inference is alienating in the sense that we seem to be turning our backs away from a fundamental fact about how we relate to some of our mental states. I questioned that objection. It might turn out that in the case of our relation to our own beliefs, agentialists are on to something, but in the case of our motivating reasons, so long as we understand them as causal explanations, the idea that inference plays a key role in our knowledge of them will remain an attractive idea.

Before turning to Keeling's own positive account of our access to our motivating reasons, I want to briefly remark on the connection between why-questions and confabulation. Keeling assumed that why-questions ask for a justification and an explanation. We saw that inferences putatively fail to provide justification for the lower-order attitudes. This was questionable but let's grant her that. Notice how a proponent of the orthodoxy might respond. They might point out that the answer one gives to the question 'why?' is not necessarily their motivating reason, since the *key takeaway* from the empirical results of experiments like that of the stockings and confabulation was that our answers to why-questions about our actions systematically failed to account for a *causal explanation*. We can turn the dual role argument against Keeling herself. The point being that whether one knows that *q* is one's motivating reason for *p* is not as simple as merely reflecting on one's non-inferential answers to why-questions (as empirical results suggest otherwise). A proponent of the orthodoxy would reject the dual role argument on the grounds that such linguistic evidence isn't clearly evidence bearing on the issue. They would demand that Keeling provide reasons to think that we should trust the social-linguistic evidence for capturing what is supposedly *causing* our attitudes. Regardless of whether our answers "must"

be non-inferential in such contexts, there is still a separate and distinct question about whether those answers are in fact our motivating reasons, and how we might come to know whether they are. Given the empirical data about confabulation, one might expect us *not* to respect the manner in which we typically answer such questions, for clearly, we are doing something wrong. So, it's not clear that accounting for that practice should be such a desideratum on a theory of self-knowledge of motivating reasons.

For the rest of the paper, I will consider Keeling's own positive, non-inferential account of our knowledge of our motivating reasons. Keeling uses the dual role of the question 'why?' as an explanandum for her own argument; that is, an account of our knowledge of our motivating reasons has to account for the fact that they explain and justify our attitudes and actions. As was argued for by Keeling, inferential answers fail to capture the justificatory aspect of the question 'why?', and therefore we should reject it. What follows is her non-inferential account.

5 Keeling's RTM account

Keeling's alternative to the (OP) is what she calls the *reasons transparency method* (RTM) (Keeling, 2019b: 10). She claims that RTM is a better alternative than the (OP) because it ultimately accounts for the dual role of why-questions—it accounts for the explanatory and justificatory aspects of the question 'why?'.[58] The argument for RTM consists in a number of subtly argued for and motivated points. In providing the details of her account, Keeling presents us with the general *method* used for learning of one's motivating reasons as well as the general *warrant* underpinning such a method. What follows is Keeling's RTM account. Later I raise some objections to her view.

5.1 The transparency method

Transparency accounts of self-knowledge get their name from a famous passage in Evans (1982):

> In making a self-attribution 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?" (225).

[58] Since I've defended the (OP) above, Keeling's argument is *at best* a competitor to the inferentialist picture.

Evans thought that our beliefs about some particular proposition *p* were 'transparent' in the sense that when undertaking to know what we believe we seem to 'look through' any putative belief of ours and instead *directly* consider the subject matter they are about. More precisely, questions about whether we believe that *p* are transparent to questions about whether *p* is the case. If I answer in the affirmative 'yes' to the question whether there will be a third world war, then I am, according to some, warranted in self-ascribing the belief that *p*. Call this *doxastic transparency*.[59]What makes the recognition of one's belief 'transparent' to oneself is either due to the fact that one "looks through" one's attitude in answering the world-directed question or is due to the fact that one's attitude *becomes* transparent to one after considering the question and settling on an answer[60]. Transparency accounts[61] ask subjects to focus their attention on the object of their attitude; for instance, in the above example one considers what is called the 'world-directed question', and often, although not always, considers the various reasons that count in favour of believing some proposition, in this case whether there will be a third world war. Importantly, subjects do not consider evidence about themselves that give them reason to believe that they hold a particular attitude. In answering the question 'do you believe that *p?'*, one does not consider evidence about themselves that would count in favour of forming the belief that they believe *p*. Rather subjects consider *directly* whether *p* is true. Transparency accounts are typically directed at determining first-order attitudes. [62]

Keeling's account breaks slightly from this tradition in that she applies the transparency account to the *motivating reasons* for which one believes some proposition[63]. The main difference between Keeling's account and other transparency accounts seems to lie in the difference of the

---

[59] This should not be confused with what is sometimes called 'the transparency of experience' which I discuss in the next chapter. See, e.g., Gilbert Harman (1990), Michael Tye (1995b, 1996a, 2003) and Fred Dretske (1995, 2003).

[60] I make this distinction because often times when answering the world-directed question the metaphor of "looking through" one's belief will not be appropriate due to the fact that the relevant belief has not been formed, and that answering the world-directed question actually generates the belief itself. This is similar to Boyle's views we briefly considered above.

[61] Notable transparency theorists are Boyle (2009, 2011a, 2011b, 2019), Byrne (2005, 2019) Fernandez (2013), and Moran (2001). Transparency accounts are closely associated with a 'deliberative' question (rather than a 'theoretical' question) for learning of one's attitudes. See Moran (2001, esp. p. 55-65).

[62] The Transparency Method (TM) isn't clearly a non-inferential method. After all, one seems to have to *reason* from some claims about what reasons there are for having a given attitude to what one's attitude actually is. See Cassam (2014) for why TM is inferential. See Boyle (2015) for a reply. Byrne, who embraces a (TM) account, explicitly states that it is inferential. See Boyle (2019) for a reply.

[63] See Keeling (2019a) for a discussion on the application of Boyle's (2011a) account of motivating reasons. Keeling's RTM account gains inspiration from Boyle's.

question considered for determining one's motivating reason. Evans' question was one concerned about the truth of some proposition *p*, whereas for Keeling it will be about whether some consider *p* is a good reason for believing *q* or ϕ-ing. Here is Keeling's account.

5.2 The reasons transparency method

Keeling claims:

> RAIN: I believe that *it will rain* and you ask me 'why?'. I consider what justifies believing that *it will rain* and conclude: *the grey clouds*. I can tell you that my motivating reason is that there are grey clouds.
>
> SEMINAR: I intend to go to the seminar today and you ask me 'why?'. I consider what justifies going to it and conclude: *the seminar will be interesting.* I can tell you that my motivating reason is that the seminar will be interesting.

According to Keeling, in order to know what my motivating reason is, I must consider the 'world-directed' question to which my motivating reason answers. I, in a sense, direct my attention to the world. In doing so, my motivating reason becomes *transparent* to me. This is similar to the way one directs their attention outward when considering whether they believe that there will be a third world war; I consider whether *p* is true and in virtue of doing this come to know whether I believe *p*. Similarly for the RTM account, I consider what the *normative reasons* are for believing *q;* I settle on some answer, say, *p*, and by doing so, *in part*, make it the case (and know) that I have motivating reason *p*.

In the Evans example, the appropriate 'world-directed question' for learning of one's belief was 'is *p* true?'. In answering in either the affirmative or negative, one can learn of what they believe. For Keeling, the appropriate 'world-directed question' for learning of one's motivating reason for a belief *q* is 'what are the good reasons for believing *q*?'. In answering the question, one can learn of what their motivating reason for a particular belief is; they conclude that *p* is a good reason for believing that *q*; based on this conclusion, one can then self-ascribe *p* as their motivating reason, similarly to how one can self-ascribe *p* as one of their beliefs after considering the world-directed question 'is *p* true?'.

So, Keeling's account is an elaboration on the more familiar transparency method (TM). More needs to be said, however, in favor of the transition from one's judgement that '*p* is a good reason for believing *q*' to '*p* is my motivating reason for belief *q*'. That is, how is it rational to transition from a seeming judgement about what one takes to be good reasons to a judgement about one's psychological life? A *key* task for those attracted to (TM) as a distinctive method for arriving at self-knowledge is to explain how in answering a *question about the world*, I thereby am warranted in giving an answer about *my psychological state*. In other words, how is it rational for a subject to transition from a *first-order judgement* about some proposition *p* to a *second-order judgement* about whether *p* is their motivating reason? As Boyle (2011b) puts it: "The puzzling thing about this transparency is how the world-oriented reflection can bear on the question about my own mental state." (226).[64] Or, more recently: "It is the psychological knowledge whose warrant is in question, and the *problem of transparency* is that nothing in my apparent basis seems to supply a ground for it." (his emphasis; 2019: 1014).

So, Keeling owes us a story about how it is rational to transition from 'p is a normative reason for q' to 'p is my motivating reason for q'. Ultimately, what Keeling suggests is that a subject will be warranted in self-ascribing *p* as their motivating reason for *q* in virtue of the fact that i) they are non-inferentially agent-aware of the relevant *normative* judgement *as* their motivating reason, and ii) that they fulfill a necessary condition (through their normative judgement) on something being a motivating reason. These two claims together imply that one non-inferentially knows their motivating reason. I will come back to this last point later on.

Before we see how Keeling attempts to justify that transition, let us first consider, again, some important things Keeling says about RTM.

5.3 From 'p is a normative reason for q', to 'p is my motivating reason for q'

To reiterate where we are at, the social-linguistic fact of how people generally treat 'why?' questions for their attitudes and actions supposedly provides support for RTM. We direct our attention to the world, or as in cases like SEMINAR, we direct our attention inward, to (supposed) facts about ourselves (that we'd find the seminar interesting).[65] Importantly, when

[64] For other recent formulations of the puzzle of transparency see, e.g., Byrne (2011, 2019, esp. Ch. 1 and Ch. 4), Braz (2019), Cassam (2014), and Moran (2012).

[65] Note that this last claim is suspect, for most who are interested in the transparency method, or some form of it, hold that we cannot direct our attention inward to our attitudes. In other words, they put a restriction on attention.

subjects consider the world-directed question for figuring out their motivating reasons, they do not use *evidence* in order to do so (Ibid., 11). RTM is apparently consistent with this phenomenon.

More importantly, when subjects consider why they believe what they believe (or act the way they act), they, in the first instance, consider what the good reasons are for the belief they hold (Ibid.). For instance, when Sally considers why she believes that it will rain, or whether it will rain, she considers the normative question: 'what are the normative reasons for believing that it will rain?'. More generally, subjects, when employing RTM, move from the world-directed question 'what are the normative reasons for believing q?' to giving some answer, 'p is a normative reason for believing q', to 'p is my motivating reason for believing q'. What justifies this transition? The move from 'p is a normative reason for believing q' to 'p is my motivating reason for believing q' is obviously not a rational inference, as Keeling herself notes.[66] Keeling's task, then, is to characterize how exactly such a transition is rational and non-inferential.

Before considering the general warrant underpinning the RTM account, Keeling briefly considers how RTM is "sufficiently reliable to issue knowledgeable beliefs" and how "our answers to the question 'why hold that attitude?' match up reliably to our actual motivating reasons." She claims:

> It seems plausible that, as an empirical matter, our answer to the question 'what are the normative reasons for holding that attitude?' reflects which reasons come to mind most strongly and easily, i.e., which are most *vivid* and *available*. Our conclusions are often influenced by such factors (e.g., Mele, 2000). And in these cases, our motivating reasons will tend to be the most vivid and available facts. (13).

As Braz (2019) puts it: "So, the puzzle of transparency might be described as the challenge of explaining why it is rational to proceed from a judgment about the external world to a judgment about one's mind, *given that we lack any capacity to look inside and directly observe our mental representations*" (7; his italics).

[66] Note two things here. First, the mere fact that we seem to make a transition between these two *judgements* would suggest that the transition is *inferential*. For some think that for something to be non-inferential it must be justified by a non-doxastic state. Second, the fact that the two judgements have such distinct contents would seem to heavily suggest that there is a background connecting belief involved supplying the justification from the normative judgement that *p* is a good reason to believe *q* to the judgement that *p* is one's motivating reason. As will emerge below, Keeling herself does attempt to ground the justification in what she calls 'agent's awareness', a supposed non-doxastic state.

In order for RTM to issue in knowledgeable beliefs about one's motivating reasons, it needs to be the case that the judgements one forms after using RTM correspond *reasonably frequently* with one's actual motivating reasons. Keeling is suggesting that this will be the case in virtue of the fact that the considerations that come to mind most frequently will tend to be the considerations that help constitute our normative judgements. And these same considerations will also tend to be our motivating reasons. The reasons why RTM is reliable, then, is due to the fact that one's normative judgements and one's motivating reasons are causally influenced by the *same kinds of considerations*; ones that play a more "active role in one's cognition will presumably make the consideration more 'present' to the subject" (Ibid., 16).

It is of course *empirically possible* that no such reliable connection obtains. In fact, it seems as if this is exactly the claim that Keeling needs to argue for in order for her account to work. For she is saying that our answers to 'world-directed questions' will be biased by such 'vivid' and 'available' considerations, and that these same considerations will become, or already be, our motivating reasons. But the claim that these types of considerations *become* or *track* our motivating reasons is exactly the main claim Keeling needs to establish. Saying that it is empirically plausible does not suffice to ground this claim.

Furthermore, is this plausible? A response here is to appeal to the very confabulation data mentioned above which motivates the claim that our motivating reason self-ascriptions are *unreliable*. Of course, the answers to questions about normative reasons are importantly different than the answers subjects might give in an experimental setting about why they think they did what they did. But it's unclear whether had subjects *instead* answered a question about what their normative reasons were they would somehow avoid confabulating their motivating reasons. It seems to me that answering such a normatively charged question and concluding on the basis of such answers what one's motivating reason is would lead to confabulation just the same.

But thankfully, Keeling wants to move away from a reliabilist account of the warrant underpinning RTM like the one briefly expressed above. She claims:

> We should avoid a reliabilist account of the warrant. After all, one option would be to say that the subject is warranted in using RTM simply because a reliable connection holds between her judging that *p* is a normative reason and *p* being her

> motivating reason. But supposing for a moment that reliable belief-formation suffices for warrant, it nevertheless does not suffice for *well-grounded* belief. Self-ascriptions, or at least self-ascriptions of the sort that are candidates for distinctive self-knowledge, seem to be instances of well-grounded belief. After all, self-ascribing the given motivating reason seems a *sensible* thing to do by the subject's lights when they employ RTM. (15).

Here we get a further glimpse into how Keeling understands distinctive self-knowledge: a distinctive feature of self-knowledge is that the beliefs are *well-grounded*. This is condition (d) above. Reliablist warrant does not suffice for well-grounded belief. The beliefs that RTM outputs are well-grounded. Hence, a reliablist warrant for RTM will not suffice. Keeling doesn't say too much about what she means by 'well-grounded', but I take it we can interpret her here to mean that the subject is aware, in some capacity, of a key feature of her motivating reason.

Recall the preliminaries:

> **Peculiar self-knowledge of motivating reasons:**
> *S has peculiar self-knowledge that their belief that p is their motivating reason for belief q only if:*
> (a) S believes that *p*.
> (b) The belief that *p causally sustains* the belief that *q*.
> (c) S is disposed to take *p* to be a good reason for believing that *q*.
> (d) S has non-inferential justification for the belief that the belief that *p* causally sustains the belief that *q*
> (e) S's belief in (d) is well-grounded.

Importantly, again, Keeling owes us a story about how we rationally transition from our judgement *that p is a good reason for believing q*—what I will again call our N-judgement—to the judgement *that p is my motivating reason for q*—what I will call our M-judgement, all the while keeping in mind the above conditions.

Keeling's preferred account of the warrant is what she calls *agent's awareness*. Our N-judgements and M-judgements, for Keeling, are *mediated* by a conscious experience *"*as of having motivating reasons which then warrants the self-ascription" (Keeling, 2019a: 18). To anticipate, Keeling claims that subjects make an N-judgement (by engaging in RTM), and through this N-judgement become aware *not only* of the fact that they have made the N-

judgement but also aware of their N-judgement *as their motivating reason*. This agentive awareness, coupled with further facts about subjects satisfying a necessary condition on them having a motivating reason, warrants the further M-judgement self-ascription.

What follows then is Keeling's attempt to justify the transition from the N-judgement—that *p* is a good reason to believe *q*—to M-judgement—that *p* is my motivating reason for *q*. I then move on to consider objections to her view.

5.3.1 The warrant: awareness of judging that p

To judge that *p* is a normative reason for q, one *partly makes it the case* that p is their motivating reason for q. How so? According to Keeling, a plausible story about the warrant involves what she calls 'agent awareness' or 'agentive phenomenology'(Ibid.). Borrowing from O'Brien (2007), Keeling claims that "subjects have a unique non-conceptual 'agent's awareness' of judgments." It's in virtue of *doing something*, in this case of judging, that agents are aware of what they judge, and ultimately are warranted in concluding that they judge that p[67].

Agent's awareness arises from relevant *background beliefs* that penetrate the agent's experience of judging. The penetration in question, or how Keeling understands it, is akin to how particular background beliefs about Royal Gala apples can penetrate one's experience such that one is experientially aware of a Royal Gala apple rather than a round red object (Ibid., 16). The background beliefs that penetrate the agent's experience come in two parts: first, there are the background beliefs which penetrate the subject's experience such that they become aware of their N-judgement; and second, there are the background beliefs which penetrate one's experience further such that they then become agent aware of their N-judgement *as their motivating reason*. Let us focus on this second kind of cognitive penetration.

As Keeling claims, in the case of *making a normative judgement*—N-judgement—the background beliefs which cognitively penetrate our experience will be: "that I generally succeed in determining my motivating reasons" and/or "that when [I] take *p* to be a normative reason it

[67] Keeling draws an analogy to intentional action. Similar to how one knows the intention with which they act (in virtue of the fact that they are doing it), one knows that they judge *that p*. Note that this picture is so far compatible with an inferentialist picture of our knowledge of our motivating reasons.

generally becomes [my] motivating reason."[68] These background beliefs, I take it, non-inferentially underpin the awareness of one's normative judgement that *p as one's motivating reason that p*. The awareness in question involves "a conscious experience as of having motivating reasons" and this awareness (supposedly) will partially warrant the self-ascription of the content of *p* as one's motivating reason.

But notice this is yet to give us a *warranted* self-ascription, for having a conscious experience of one's normative judgement that *p* for belief *q* as one's motivating reason for belief *q*, does not tell us what *justifies* the further belief that *p* is actually one's motivating reason. It's one thing to *experience* one's N-judgement as being one's motivating reason, and another to be justified in believing that it is one's motivating reason.[69] What follows is Keeling's claims about how beliefs formed on the basis of such judgements (and the accompanying agent's awareness) are justified.

5.3.2 Partly making it the case

Recall two conditions on having a motivating reason concerning the basing relation. For her, the basing relation consists in: "(1) the belief that *p* must causally sustain the belief that *q,* and (2) S must be disposed to take *p* to be a good reason for believing *q*." (Ibid., 2). She claims:

> When subjects employ RTM, and take *p* to be a normative reason for holding attitude *A*, when all is going right, subjects *partly make it the case* that they hold *A* for the reason that *p*. This results, I take it, from a relevant constitution relation. Necessarily, when one holds the attitude *A* for the reason that *p*, one will be prepared to take *p* to be a good reason for holding *A*.

Simply put, in order to hold an attitude for a reason it is necessary that one 'be prepared' to take the motivating reason to be a good reason. In other words, one cannot hold attitude A for reason *p* if S is not 'prepared to take' *p* as a good reason for A. No preparedness, no reason. Why mention 'preparedness'? For Keeling, being prepared to take *p* to be a normative reason is

[68] Ibid., p. 19. Why think that subjects have these background beliefs? Keeling again points to the fact that subjects' answers to 'why?'-questions are "epistemically criticisable" and that the general practice of engaging in criticism of one's answers to 'why?'-questions "seems to presuppose that [one] sees normative reasons as potential motivating reasons." (see p. 18 in particular).

[69] A perhaps simpler story to tell at this stage, is to allow the background beliefs to confer justification to the agentive awareness of one's N-judgement as their motivating reason, and have one's motivating reason belief be based on this non-doxastic agent's awareness, but nonetheless epistemically dependent on the prior background beliefs.

constitutive of the basing relation – it's a necessary condition. Subjects take a consideration, *p*, to be a normative reason and—*by fulfilling a necessary condition on basing*—partly make it the case that *p* is their motivating reason (that *q* is based on *p*). One's self-ascription that *p* is their motivating reason for belief *q* is warranted in virtue of the fact that agents partly make it the case that they have motivating reason *p* when they judge that *p,* by exhibiting the disposition to take *p* as a good reason for believing *q*. That is, when subjects make an N-judgement, the N-judgement is itself part of what it means to have a motivating reason.

We are now in a position to see how Keeling's account gives us peculiar (direct non-inferential) self-knowledge of motivating reasons. Here's an attempt at filling in the picture: Sally believes that *it will rain*. She's asked 'why?' she believes this. On the RTM account, here's what happens. Sally considers the world-directed question 'why believe that it will rain?'. She deliberates[70] about the question, settles on an answer, namely, that *there are grey clouds*. Sally gains knowledge that *that there are grey clouds* is her motivating reason as follows: Sally becomes agent-aware of the reason for the belief that *it will rain* in virtue of judging the supposed fact *that there are grey clouds* to be a normative reason for the belief *that it will rain*—her N-judgement. A necessary condition on having a motivating reason *p* for believing *q* is that "one be disposed to take *p* to be a good reason for *q*". So, in making an N-judgement one "partly makes it the case" that *p* is their motivating reason; they satisfy a necessary condition on something's being a motivating reason for them. We are then agent-aware of our N-judgements, and aware of a necessary condition, although not in those terms. This awareness is in turn then cognitively penetrated by a background belief turning the content of one's awareness of their N-judgements into an awareness with content having to do with what their motivating reason is. The belief Keeling appeals to in this case to cognitively penetrate the N-judgement awareness is the following: the belief *that I generally succeed in determining my motivating reasons*. One is then agentially aware *that p is their motivating reason* through such cognitively penetrated conscious experiences and the fact that they have in part determined their reason, and then is able to self-ascribe their motivating reason on the basis of this awareness. The transition from our N-judgements to our M-judgements are therefore justified in the above way and yield distinctive self-knowledge.

[70] As Keeling mentions, deliberation is not necessary. Sometimes agents will simply conclude something is the case (e.g., that the clouds are grey), and in doing so, partly make it the case that it is their motivating reason.

5.4 Summary

If what Keeling says is correct, and none of the above involves any inferential processes, then we can learn of our motivating reasons non-inferentially. Importantly, Keeling's view can account for the dual role of the question 'why?'.[71] Sally takes the why-question seriously when she answers the world-directed question, judges that *p* is a normative reason for believing *q*, and becomes agentially aware that *p* is her motivating reasons for *q*. Keeling's RTM and the warrant underpinning it, apparently accounts for the following:

(a) S believes that *p.*
(b) The belief that *p causally sustains* the belief that *q.*
(c) S is disposed to take *p* to be a good reason for believing that *q.*
(d) S has non-inferential justification for the belief that the belief that *p* causally sustains the belief that *q.*
(e) S's belief in (d) is well-grounded.

When one employs RTM and all is working well, (a)-(e) will apparently all be met. Keeling's RTM account, however, faces challenges of its own. RTM seems to capture the justificatory feature of the question 'why?', but not the explanatory feature. That is, when agents offer what they take to be their motivating reason for a belief or action (under RTM), nothing about RTM makes the agent aware that *what* they have judged to be a normative reason for a belief or action *in fact caused* the relevant belief or action. And if one of Keeling's explananda is that such beliefs be *well-grounded*, it isn't entirely clear that we get such well-grounded beliefs about our motivating reasons. Rather we simply get well-grounded belief about what we judge or are prepared to take to be a good reason for a belief; that is, a well-grounded belief about our N-judgement.

In the next section, I develop this objection in more detail. I argue that RTM fails to account for how subjects become non-inferentially aware that their normative judgement made via RTM is in fact the cause of the target first-order attitude. Absent such an account, we get at best a *reliablist* rather than a *well-grounded* picture of how RTM provides knowledge of motivating

[71] Note that I think the orthodoxy *can* account for the dual role of the question 'why?' for reasons I mentioned above. So at best we have two competing accounts of how one might have self-knowledge of motivating reasons—one inferential, the other non-inferential. The accounts are similar in the sense that subjects will come to know of their N-judgement *non-inferentially*.

reasons, failing to capture one of Keeling's explananda for a theory of self-knowledge of motivating reasons. But, as we saw, we seem to have good reason to reject a reliabilist account of the warrant underpinning RTM, since the confabulation data mentioned at the outset raised doubts about our reliability. But before that, I want to raise a different point: whether or not Keeling's RTM account is indeed non-inferential. For it seems as though her appeal to background beliefs to "modulate" our experiences of our judgements makes it plausible that there in fact are background beliefs providing justification.

6 Objections against the RTM

I consider two main things in this section. The first thing I consider is that Keeling's RTM account isn't so clearly a non-inferential account. Second, I argue that Keeling has not provided us with reason to think that RTM gives us non-inferential knowledge of the fact that our N-judgements (or the beliefs implicated in them) causally sustain the first-order attitudes they target.

6.2 Is RTM really non-inferential?

Recall that Keeling claims that subjects become agent aware of their N-judgement as their motivating reason because subjects typically have background beliefs which cognitively penetrate the subject's experience. The relevant background belief, according to Keeling, was one of the following:

> Belief (1): that I generally succeed in determining my motivating reasons, or
> Belief (2): when I take *p* to be a normative reason it generally becomes my motivating reason.

Keeling claims these background beliefs merely "modulate" our experience to cognitively penetrate them in a way that alters the content of what we are conscious of. We go from an experience of a judgement about normative reasons to an experience of a judgement about what our motivating reason is. That is, one goes from an *experience* with the content [I judge that *p* is a good reason for believing *q*] to an *experience* with the content [*p* is my motivating reason for believing *q*]. Again, belief (1) or belief (2) merely modulates that alteration of content.

But I think this claim is dubious. For we aren't given much reason to think that these background beliefs do not mediate the justification between one's agent awareness of their N-judgement as their motivating reason, and their belief that their N-judgement is their motivating reason. Keeling stresses that the background beliefs involved in one's cognitively penetrated agent's awareness of their motivating reason only "modulate one's experience". But if such beliefs are *required*, then one's agential awareness of what one's motivating reason is seems to depend on the justification of these background beliefs, which entails that the belief formed on the basis of such experience is epistemically *mediate*. Keeling isn't careful to distinguish between a belief being psychologically non-inferential and epistemically independent. Clearly, she aims to make her reasons transparency method at least *psychologically non-inferential*. But it's unclear that that's what is at issue when proponents of the (OP) claim that knowledge of our motivating reasons is inferential.

Recall, also, from Sect. IV, that *if* inference is epistemically problematic in the psychological sense—i.e., in the sense that one explicitly engages in an inference—then the belief one forms with respect to their motivating reason must not itself be *epistemically mediate*. The key thought there was that were something to be epistemically mediate in the sense that it relies on the justification of background beliefs for its own justification—e.g., think of the belief that there is a proton present—then there should not be anything epistemically wrong with making that justificatory support explicit. In other words, there shouldn't be anything wrong with making that inferential support explicit. So, the justification for beliefs about one's motivating reason should not depend *epistemically* on further background beliefs. Now, appealing to background beliefs at least makes it prima facie plausible that one's motivating reason beliefs are epistemically dependent.

Of course, Keeling can stick to her guns and simply insist that belief (1) or (2) merely modulates one's experience without playing a significant epistemic role vis-à-vis the justification of one's motivating reason beliefs.[72] That is a possibility, but we might wonder why subjects hold these beliefs in the first place. It seems quite convenient for Keeling's account that subjects a) happen to have the beliefs required to give rise to the necessary experience, and b) that those beliefs do

[72] Recall Keeling appeals to a cognitively penetrated experience of seeing an object as a Royal Gala apple rather than a red and round object. But the apple case seems importantly different than the motivating reason case.

not play a justifying role vis-à-vis the subject's motivating reason belief. We are at least owed an explanation for how all subjects come to hold these required background beliefs. One plausible explanation for why subjects hold such background beliefs is because they engage in implicit and explicit theorizing about their reasons, and over time come to develop such background beliefs (e.g., taking note of the timing of certain thoughts with the formation of new beliefs and the performance of actions). This explanation seems to lend support for the idea that the background beliefs provide justification for one's belief about their motivating reason.

Keeling will of course deny this picture. But she needs to say more about these background beliefs, and why we should think that they merely modulate experiences—i.e., alter the content of what we experience—without further providing the justification for the beliefs based off such experiences.

An important question which needs to be addressed in this context is what role do such background beliefs like (1) and (2) play vis-a-vis not only the content of one's agentive experience—what Keeling calls agent awareness of one's normative judgement as one's motivating reason—but also with respect to the justification conferring powers of that awareness/experience? That is, given that cognitively penetrating states are required for subjects to have the relevant experience of something as their motivating reason, and hence to have justification for motivating reason beliefs, do those cognitively penetrating states confer justification to one's agentive awareness? Again, Keeling will respond in the negative. But if that's the case, then for subjects to have the relevant cognitively penetrated experience—agentive experience *as of* their motivating reason—provide justification for motivating reason beliefs, not only will the *content* of one's experience need to be altered to include motivating reason content, but subjects will also need to stand in some sort of relation to the causal relation their experience is putatively *about*, the causal relation their background penetrating beliefs *make them aware of*.

Notice Keeling's story hasn't said anything about whether such a causal relation will obtain between our normative judgement—e.g., when we judge that the grey clouds count in favour of rain—and our beliefs and actions—e.g., our belief that it will rain. But then we might plausibly ask: where does the justification come from if not the background beliefs? Well, one story which could be told is that subjects *partly make it the case* that something is their motivating reason,

and it is *this fact* which the cognitively penetrating states (e.g., belief (1) and (2)) make the subject aware of. This is precisely the story Keeling seems to want to give. But I don't think it makes subjects aware of the *proper condition* on having a motivating reason; that is, what subjects are aware of when they "partly make it the case" that they have a motivating reason—i.e., when they make a normative judgement—*is not* that they have made a causal relation obtain but rather merely that they are disposed, e.g., to take *p* to be a good reason for belief that *q*.

I now turn to this worry.

6.3 Knowledge of the causal condition

This brings me to the main worry with Keeling's RTM account. Recall that in order for a belief to be based on another belief, and for something *to be* one's motivating reason: "(1) the belief that *p* must causally sustain the belief that *q,* and (2) S must be disposed to take *p* to be a good reason for believing *q*." (Ibid., 2). RTM gives us an account of how condition (2) is satisfied. When agent's judge that *p* is a good reason for believing *q,* they will be *prepared* to take *p* to be a good reason for believing *q*. Agents satisfy a condition on having a motivating reason by satisfying condition (2)—by being prepared to take *p* to be a good reason for belief *q*. What about condition (1)? Keeling hasn't said anything about the agent's relationship to condition (1). For all agents do in judging that *p* is a normative reason for *q* is make condition (2) obtain.

So, although subjects have an experience *as of* their normative judgement being their motivating reason, they don't seem to have non-inferential justification for the belief that their normative judgement *is* their motivating reason, since what they are putatively aware of when making a normative judgement is condition (2)—that's what actively judging *does*—and importantly *not* condition (1) about causal relations. In other words, given that one's experience justifies in virtue of them *making something the case*, they only have non-inferential justification of what they *judge* to be a good reason for believing, and not whether their judgement actually appropriately causes their lower-order attitude or action.

In other words, RTM doesn't explain how answering 'word-directed questions'—'what are the normative reasons for believing that *q'*—and thereby self-ascribing the corresponding belief that *p* which answers such questions, gives *non-inferential self-knowledge of the fact that the belief that p causally sustains the belief that q*, or for that matter, of *any* causal relation. For

nothing we have been told about the structure of our agent's awareness of our N-judgement as our motivating reason suggests that we are aware of a causal relation obtaining between our N-judgement (or the beliefs implicated in them) and the target first-order attitude. Recall:

> **Peculiar self-knowledge of motivating reasons:**
> *S has peculiar self-knowledge that their belief that p is their motivating reason for belief q only if:*
> (a) S believes that *p.*
> (b) The belief that *p causally sustains* the belief that *q.*
> (c) S is disposed to take *p* to be a good reason for believing that *q.*
> (d) S has non-inferential justification for the belief that the belief that *p* causally sustains the belief that *q*
> (e) S's belief in (d) is well-grounded.

Condition (b) and (c) are the two necessary conditions found in the version of the basing relation that Keeling gave us. In order to have *distinctive self-knowledge* of my motivating reason, (b) and (c) need to obtain, and I then need to make a judgement that *p* is a good reason for believing that *q*. This judgement *itself* satisfies condition (c), and I thereby I am supposed to have distinctive self-knowledge of my motivating reason; importantly, this knowledge is apparently *well-grounded* since I am aware of condition (c); this is presumably why it *makes sense* to self-ascribe the motivating reason.

But here is the worry. The worry raised above targets condition (d). In employing RTM, I seem to gain non-inferential justification for the belief that I am disposed to take *p* to be a good reason for believing *q* or perhaps that I have made a particular N-judgement. But surely, we need justification for the claim that our N-judgement causally sustains belief in *q*. RTM doesn't account for how agents know or have non-inferential justification for belief in (b), hence *(d)* is not met; and knowing whether this condition—(b)—obtains seems crucial for whether an agent knows, in a particular case, what their motivating reason is. This gap might be surprising since a key source of doubt that we have non-inferential knowledge of our motivating reasons, is that we do not have non-inferential justification of causal relations.

To reiterate, one's agent awareness of their N-judgement as their motivating reason only puts the subject in touch with *part* of a necessary condition on having a motivating reason. Although subjects do something when they answer world-directed questions—they express what they are disposed to take to be good reasons in favour of some belief—what they *do not do*, at least on the picture Keeling has offered us, *is make it the case that their judgement about the good reasons for believing something causally sustain believing that thing*. In short, she has not shown that we can have non-inferential self-knowledge of a causal relation, and therefore of our motivating reasons.

Keeling might object as follows: S *need not know* that condition (b) obtains or be aware of it in any way; it only needs to be the case that (b) obtains. When conditions (a)-(e) obtain, then S will know that belief *p* is their motivating reason. She might further respond that knowing that (a)-(e) obtain is only required to *know that one knows* their motivating reason, but not for simply knowing their motivating reason. So long as subjects have the sorts of agentive experiences—i.e., experiences of their normative judgements *as being their motivating reason*—and those experiences produce beliefs which *reliably track* the fact that their N-judgements causally sustain the target lower-order attitude, then subjects will have distinctive self-knowledge of their motivating reasons.

But this reply raises worries of its own: First, this sounds like a merely reliabilist story of the warrant underpinning our self-ascriptions (something Keeling does not want); and second, if Keeling wants to retreat back to agent's awareness, we need to be told why it is condition (c) that is important for warranting the relevant self-ascriptions rather than condition (b). That is, we need to be told how an agent's experience of their normative judgement *as* their motivating reason *non-inferentially* justifies the belief that their normative judgement causally sustains the target first-order attitude. I will take both points in turn.

Recall that a plausible story about the warrant for RTM was that RTM is a reliable method for knowledgeable beliefs about motivating reasons. This was backed-up by considerations having to do with the empirical plausibility that our answers to the question 'what are the normative reasons for believing *q*?' will typically be considerations that are most 'vivid' and 'available'. Furthermore, for Keeling, it seems empirically plausible that these considerations will become our motivating reasons; that is, the most *salient* considerations will tend to be our answers to

the 'world-directed question' as well as our actual motivating reasons. But as I mentioned when first considering this point, to merely point out that it is empirically plausible that our answers to 'what are the normative reasons for believing *q*?' reliably track our motivating reasons is not enough to properly establish RTM as a facilitator of knowledge. And as we saw at the outset of this chapter, we have what seem to be good empirical reasons—e.g., the confabulation data—to think that subjects are quite unreliable at knowing the causal sources of their actions and attitudes.[73]

In any case, Keeling herself wants to *reject* a reliabilist account of the warrant in light of the fact that it does not suffice for *well-grounded beliefs*. After all, when subjects self-ascribe their motivating reasons, it seems a sensible thing to do in the sense that they have a particular kind of experience which illuminates the fact that something is their motivating reason. But it seems that for Keeling to account for well-grounded belief, she needs to appeal to the fact that subjects *make it the case* that they have a motivating reason. There are *two* necessary conditions on having a motivating reason: (b) and (c). So, perhaps Keeling can lean on the fact that being agent aware of one's N-judgement as their motivating reason is enough to provide non-inferential justification for belief in one's motivating reason because it's enough that one *partly* makes it the case that they have a motivating reason.

In this case, Keeling must then appeal to the fact that subjects are *aware* of the necessary conditions on having a motivating reason. This, after all, seems to get us *well-grounded* belief. But it still seems a mystery how such awareness manages to give the agent in question a well-grounded belief of their motivating reason when the only thing a subject is agent-aware of is *one* of the necessary conditions on having a motivating reason. Agent's awareness gives us a story of how we *partly* make it the case that some consideration is our motivating reason: when we make a normative judgement, we will be *prepared* to take it to be a good reason (satisfying condition (2) on basing). But condition (1) of the basing relation, or condition (b) above, is left unaccounted for. In fact, *if* well-grounded belief is one of our explananda, and we want to construe this well-groundedness in terms of *agent's awareness*, then wouldn't it be more appropriate to understand that awareness as an awareness of the causal condition rather than the preparedness condition? After all, it's this particular *causal* understanding of a motivating

[73] There is also a question, which I leave to the side, about whether the justification for a reliably produced belief is inferential or not.

reason that is motivating the orthodox position. In other words, it's unclear how one can have agent's awareness, and hence well-grounded belief, that something is their motivating reason if they do not have such awareness of their normative judgement causally sustaining the target belief. At best, then, we are presented with an account of how we could be aware of (c) obtaining. But given that that's only a necessary and not a sufficient condition, it is unclear how it would turn out to be sensible, by the subject's own lights, to self-ascribe the motivating reason. (c) on its own does not suffice to warrant a knowledgeable belief about one's motivating reasons. If all this is correct, then Keeling has not given us an account of how we can have non-inferential, well-grounded knowledge of our motivating reasons. Importantly, a key—perhaps *the key* condition she set out to accommodate—namely (d), that we have non-inferential justification for believing that a causal relation obtains, is not captured by her account.

Conclusion

The orthodox position claims that we can only have knowledge of causal relations via inference. Motivating reasons are best understood as the causes of attitudes and actions. Therefore, we can only have inferential knowledge of our motivating reasons. Furthermore, inference is inimical to self-knowledge. Therefore, we do not have *distinctive* (what I have been calling *peculiar*) self-knowledge of our motivating reasons. However, inference apparently cannot provide justification for our first-order attitudes, and thus knowledge of our motivating reasons is non-inferential. I gave reasons to doubt this claim in Sect IV. The picture I presented was one in which we make a normative judgement about whether *p* counts in favour of believing *q* and then from our (perhaps non-inferential) knowledge of *this* judgement infer that it is our motivating reason. Importantly, this inference relied on a belief about the fact that N-judgements tend to be the causes of first-order attitudes. So, we had good reason *not* to reject an inferentialist account of our knowledge of our motivating reasons.

Keeling then attempts to give us a picture of how knowledge of our motivating reasons is obtained in a *non-inferential, well-grounded* way. Her account, I have argued, fails to provide us with the appropriate knowledge. She can either give up on well-groundedness as an explanandum and opt instead for a reliablist picture of the warrant or give us a more detailed account of how exactly subjects come to have non-inferential knowledge of the relevant causal relations such that self-ascribing a motivating reason seems a *sensible* thing to do in the eyes of the subject. It seems she can't (and doesn't) want to go in for reliabilism since it seems to be the

very thing undermined by the empirical data. So, she's left with having to provide a better picture of the well-grounded warrant, a picture which importantly must remain non-inferential.

In light of the difficulties which a non-inferential account of our access to our motivating reasons presents, I think it best to temper any hope that we can non-inferentially know our motivating reasons.

Interlude

We saw that what posed a problem for the idea that we enjoy distinctive self-knowledge of our motivating reasons was the further idea that such entities are in part *causal entities*. The causal nature of motivating reasons threatens their status as candidates for distinctive self-knowledge: they must be known *inferentially*. So, when we turn to look at the *normative* properties of our mental states—something like the *badness* of our unpleasant pains or the *normative reasons* they provide—the prospect for distinctive self-knowledge seems highly plausible, even likely. For whether something is bad-for-you or a normative reason need not, at least on the surface, involve anything to do with whether something has caused anything else. Recall my splitting headache from the introduction. My headache is so unpleasant that I head to the medicine cabinet and pop a painkiller. Now, *why* I head to the medicine cabinet to take a painkiller—presumably, *because* of the badness of my unpleasant pain—is something I apparently can't have distinctive self-knowledge of. But notice something about this picture which seems *immune* from that claim: my knowledge that *it*—the unpleasant pain—*is bad*. In other words, what seems like an excellent candidate for distinctive—hence direct non-inferential—self-knowledge are the normative properties of our mental states, e.g., the badness of unpleasant pain.

But, of course, things aren't so simple. For there is once again a serious incompatibility here. Previously, the incompatibility was between the causal nature of motivating reasons and direct non-inferential self-knowledge of motivating reasons. We saw that those two things just don't seem to be reconcilable, so we gave up on the latter. But now the incompatibility is slightly different. The incompatibility is between three things: the *normativity* and *motivationality* of affective mental states—e.g., the badness of unpleasant pain—the *representational nature* of those states, and direct non-inferential access to those states. However, rather than abandon the idea that we enjoy direct non-inferential access to those states, and hence to the normative

properties of them, I instead argue that we should abandon the *representationalist conception* of those states since the particular *inferential* account representationalists must give of our access to our phenomenal states is itself seriously incompatible with the normativity and motivationality of affect.

I now turn to these issues.

CHAPTER 2: Representationalism, transparency, and the problem of de re desire

1 Introduction

Most of our pains are unpleasant[74]. And most of our unpleasant pains motivate, for instance, when we remove our hand from an intensely hot stove. The standard view about the nature of unpleasant pain—sometimes called the 'second-order desire' view—has it that pain's unpleasantness and its motivationality are constituted by the frustration of an anti-pain desire, namely, an *experience-directed desire* not to be undergoing the pain sensation one is currently experiencing. What is unpleasant about your pain is that you want not to be experiencing *it*. Furthermore, what *motivates* you to do something about your pain is your desire not to be experiencing *the pain*.

But that's not all we can say about unpleasant pain. For instance, all our unpleasant pains are pro tanto *bad* for us. And most of us respond to our unpleasant pains, for instance, when we take painkillers. Although somewhat neglected facts, philosophers have attempted to capture these further claims of unpleasant pain by appealing to desires. Fellow evaluativists about unpleasantness—those who explain unpleasant phenomenology in terms of *representational* content—such as Brian Cutter and Michael Tye, and David Bain do invoke such desires. For instance, Brian Cutter and Michael Tye (2014) invoke desires to explain the *badness* of unpleasant pain, and David Bain (2017) invokes desires to explain our *anti-unpleasantness motivation*, e.g., to take painkillers. What is so *bad* about my unpleasant pain, according to Cutter and Tye, is that I desire not to be experiencing it. And why I head to the medicine cabinet and reach for the painkillers, according to Bain, is because, again, I desire not to be experiencing *that* unpleasantness. These further normative and motivational challenges can be met by appealing to desires.

Notice, for our purposes, what the desires are directed at. In both cases the desire is directed at the unpleasantness itself: *a phenomenal episode*. Capturing unpleasant pain's badness and our anti-unpleasantness motivation is, for some, a matter of directing a desire at the affective

[74] The case of pain asymbolia, where subjects report being in pain but do not seem bothered by it, suggests a separation between the sensory dimension of pain and the affective, 'awful' or 'bad' dimension of pain. See Grahek (2007) for discussion. I follow what is commonly taken to be practice by distinguishing between these two components of pain.

component of pain. It would seem, then, that underlying these desire-based strategies is the thought that it is possible to direct desires at phenomenal episodes, in particular, *de re* desires that are *directly* about the unpleasantness in question.

But this is no innocent claim, or so I want to argue, since it would seem that for subjects to hold the *relevant de re* desire, they need to be directly aware of their phenomenal states. Yet there is a view about the nature of phenomenal experience that precludes those experiences being the direct objects of such awareness, and hence desires. That view is strong representationalism about phenomenal consciousness, which, roughly, claims that the phenomenal character of experience—what-it-is-like to undergo an experience—is identical and reducible to its representational content. Furthermore, strong representationalism is committed to the following: we cannot directly introspectively attend to our own phenomenal episodes since such episodes are *transparent* to the extra-mental features they represent. Yet, it is just this sort of direct attention precluded by strong representationalism which some theorists, e.g., evaluativists, need to explain the normative and motivational features of sensory experience.

So, evaluativists better not be strong representationalists if they want to invoke experience-directed desires in their normative and motivational explanations. But even more problematic, and what is the focus of this paper, is the more general claim that strong representationalists—whether they be evaluativists or not—*cannot* invoke the necessary experience-directed desires to explain the normative and motivational features of affect. Strong representationalism is incompatible with a highly attractive and intuitive naturalistic account of what makes our unpleasant pains bad and why we take painkillers: we *desire* not to be experiencing the unpleasantness we are currently undergoing. But if we are strong representationalists about phenomenal consciousness, then we cannot avail ourselves to those attractive desire-based explanations, for we need to be aware of our experiences in a manner precluded by strong representationalism's commitment to the transparency of experience. Therefore, once we consider the normative and motivational challenges unpleasantness presents us with, and we acknowledge the popularity of and attractiveness in giving a desire-based explanation of those challenges, we ought not be strong representationalists about unpleasantness[75].

[75] Although I couch my argument in terms of 'unpleasantness' I take it that what I say here applies *mutatis mutandis* to pleasant states as well. So, the target of my paper is strong representationalism about affect, more generally.

The paper proceeds as follows. In §2 I present two challenges any account of the nature of unpleasant pain must face: *the normative challenge* and *the motivational challenge* as well a desire-based strategy for meeting these challenges. Here, I motivate the thought that desire-based strategies need the relevant desires to be *de re*. In §3 I make clear the kind of representationalism that is the target of my paper and present the kind of transparency of experience which it entails. In §4 I argue that strong representationalism's commitment to strong transparency entails that strong representationalism is incompatible with any desire-based explanation of the normative and motivational features of unpleasantness. In §5 I consider objections and give my replies. In §6 I sketch the potential problems representationalists face when trying to accommodate the motivational challenge in a non-desire-based way. In §7 I briefly sketch where all this leaves strong representationalism vis-à-vis the motivationality and normativity of affect. I then conclude.

2 Two conditions and the desire-based strategy

In this section, I present the normative and motivational challenges that an account of unpleasantness must meet. I then introduce an attractive and popular desire-based strategy for meeting these constraints: that we direct *de re* desires at our own unpleasant experiences.

Say I place my hand on the hot stove which I've very recently been cooking on. My unpleasant pain no doubt motivates me to quickly remove my hand from the stove. But after a while, my unpleasant pain remains, and wanting to continue with my cooking, I decide to head to the medicine cabinet to take a painkiller. In such a case, my unpleasant pain—independent of anything it might tell me about the state of my hand—gives me a reason to *tend to the unpleasantness of the experience itself*, perhaps by taking painkillers and eliminating the unpleasant experience. Why, the story might go, I have this reason is because my unpleasant pain *itself* is a *bad* state to be in. It is pro tanto non-instrumentally bad-for-me[76] *independent* of any downstream negative consequences it may cause. The explanatory challenge, then, is to explain unpleasant pain's *reason-giving power*. Call this the normative condition.

> **Normative Condition**
> An unpleasant pain is non-instrumentally *bad* for its subject—it provides a subject undergoing that experience with a non-instrumental *justifying reason* to eliminate the unpleasantness itself.

[76] By 'pro tanto' I mean the badness is not extinguishable by some other kind of value but is defeasible.

Note some clarificatory points. First, we're switching gears here from the previous chapter in thinking about *motivating reasons* to now thinking about *normative reasons*. Of course, when one acts for a reason which is their unpleasant pain—e.g., when they take a painkiller—they will have a motivating reason. But we're no longer thinking about how one knows whether or not some thought of theirs is their motivating reason.

Second, we are concerned now with the *normative property* that a quality like unpleasantness provides to its subject. For most of the paper, we will be concerned with how one might go about *explaining* that normativity in light of certain theoretical commitments. But towards the end of the paper, we will see the troubles those theoretical commitments pose for how we *access* that normativity, i.e., how we rationally respond to the normative reason itself.

Third, the normativity under discussion here, and what will concern us throughout the paper, applies to the *unpleasantness itself* and not to the affectively neutral pain experience, nor to the putative normativity of the represented extra-mental bodily damage: we are not concerned about the reasons provided, if any, by *bodily damage* nor by mere *represented* bodily damage. Furthermore, the non-affective pain experience itself is *not* a non-instrumental normative reason for any action, at least I don't think it's plausible that it is one *on its own*. The relationship between the pain experience and the bodily damage, more plausibly, is that of a motivating reason (the pain) that *represents* a normative reason for action (the bodily damage).[77] Although much more needs to be filled in, that picture plausibly, and very roughly, looks something like this: affectively *neutral* pain experiences represent bodily damage, a putative normative (perhaps even *instrumental*) reason for a subject to engage in avoidance behaviour, e.g., by lifting their hand out of scalding hot water. By doing so, not only is one in contact with a worldly normative reason, but importantly, one's behaviour is marked-off as one belonging to the 'space of reasons.' Although the extra-mental bodily damage and its normativity is of *some* concern in this paper, it is only of concern in relation to the normativity of the *unpleasantness*, specifically, as contributing to an explanation of unpleasantness' reason-giving power by being that which is represented.[78] Importantly, our focus here is on the *unpleasantness* of pain.

[77] See Bain (ms) for something like this picture.
[78] I discuss this later when discussing Bain's "perceptualist" strategy for accounting for the normativity of unpleasantness.

Moving on. Note, further, that not only does my unpleasant pain provide me with a reason, but that *I do in fact act* on the reason provided by the unpleasantness itself, for instance, when I reach for the cabinet and take a painkiller. My action here is importantly *not* directed at my body—the anti-damage behaviour I typically display when I *remove* my hand from the hot stove—but *rather* is directed at my *experience*. Although my being provided with a reason is closely related to my acting on it, we should distinguish between these two facts. We can formulate another plausible condition:

> **Motivational Condition**
> Subjects, when undergoing unpleasant pains, typically *rationally respond* to the reasons provided by those unpleasant pains, hence, why they take painkillers[79].

So, the normative condition and the motivational condition are normative and motivational truths, I take it, that *any* account of the nature of unpleasantness must accommodate. One extremely popular way of accounting for the normative and motivational features of our sensory experiences has been to appeal to desire[80]. More recently, desire-based strategies have been proposed for capturing these conditions, one presented by Brian Cutter and Michael Tye (2014) and the other by David Bain (2017). Let's consider them in turn.

2.1 The desire explanation of normativity

Our desire-based strategy accommodates the normative constraint by explaining the badness of unpleasantness in virtue of some feature *independent* of the unpleasantness itself. This is the strategy invoked by Brian Cutter and Michael Tye (2014):

> The mere painfulness of a pain sensation, absent any aversion to it, does not provide the subject with a reason to get rid of the sensation. In other words, the experience of pain does not provide its subject with a reason to get rid of it *simply* in virtue of its painfulness; the subject must also have an aversion to it. (431)

[79] See Bain (2013), Cutter and Tye (2014), Brady (2018b), and Jacobson (2019a) for discussion regarding the normative condition. See Bain (2017) and Aydede and Fulkerson (2018) for discussion of the latter claim as well as for discussion of the motivational condition.
[80] See, e.g., Armstrong (1968), Brady (2018), Hall (1989), Heathwood (2007) and Pitcher (1970b) to only name very few. See Bain (2013) and Aydede (2014) for more references and discussion of such views.

Cutter and Tye explain the reason-giving power of unpleasantness (what they call 'painfulness'), and our responsiveness to that painfulness, by appealing to a subject's aversion to such experiences. On their view, it is the subject's *additional* aversion—what we might call the subject's frustrated *intrinsic* desire not to be undergoing that unpleasant experience—that explains how unpleasant pain is non-instrumentally reason-giving. The normative constraint is met by appealing to the intrinsic desire not to be undergoing that current unpleasant experience, and our responsiveness to our unpleasant pain is explained by that same desire. What is so bad about your unpleasant pain is that you want not to be experiencing it. So, on this strategy, if one can explain the normative condition, they are in a position to also explain the motivational condition.

2.2 The desire explanation of motivation

The second strategy, espoused by David Bain (2017), does not appeal to desire to explain the normative constraint, but rather to explain the motivational constraint. In explaining the normative constraint, Bain appeals to something *intrinsic* to the unpleasantness, namely, to the perceptuality of an interoceptive pain experience. Briefly, the badness of unpleasant pain is explained, in part, by the way in which one perceptually experiences their bodily damage as being bad-for-them, in the sense that one "encounters" the badness of one's bodily states. It's not that one is merely being *told* about some bad state of one's body—as in the case of evaluative belief—but rather that *that* particular normative state impresses itself in an intrusive manner on our perceptual awareness. And appealing to the notions of *encounter* and *impress* seems to capture what might be so bad about pain experiences. Unpleasant pains are bad because they are constituted by our perceptual encounter with our bad bodily sates.[81]

But, importantly, for Bain, our *responsiveness* to our unpleasant pain—our painkiller-taking action—*is* grounded in an *independent* desire to eliminate the unpleasant pain itself, that is, a desire for the unpleasantness (and its badness) not to be occurring (2017: 485). Why you head to the medicine cabinet to take a painkiller is because you desire not to be experiencing the unpleasant pain you are currently experiencing.

[81] More on this in Sect. VII.

Although our desire-based strategies are aimed at explaining different things, they share one crucial feature: *both explanations claim that we direct desires at phenomenal episodes to explain a feature(s) of unpleasant pain*. Both invoke experience-directed desires. Importantly, notice something implicit in this thinking. The sort of desires our strategies *need* are ones *directed at* phenomenal episodes. And what this seems to further require is that we be intimately aware of our experiences such that we can *direct desires at them*. In other words, if we are to invoke desires in our normative and motivational explanations of unpleasantness then they must be *de re* desires, or so I want to suggest.

2.3 De re desire

Intuitively, I take it that what underpins our attraction to the idea that desires can make a normative and motivational difference is the further thought that subjects can have desires—or in our case have anti-desires—about specific things. What it is that you want not to be occurring is *that* experience; it's the specific *instance* of unpleasantness that our anti-unpleasantness desires are directed at that seems to make a normative and motivational difference. What we find so bad about unpleasant pains and what putatively *moves me* to take a painkiller is—most plausibly—a particular instance of unpleasantness my anti-unpleasantness desire is about, a desire whose reference is fixed *directly,* presumably by a demonstrative thought about, or direct awareness of, the unpleasantness itself. In other words, the desires evaluativsts invoke must be *de re*.

To illustrate further, consider the following. Say Max has a general, *de dicto* desire not to be in a phenomenal state of the *stabby-unpleasant-pain-kind*; and further, that Max is now in a phenomenal state which instantiates the property *being a stabby unpleasant pain*. Say, also, that Max is *unaware* that he is in such an unpleasant state. Intuitively, Max's unpleasantness which he happens to want not to be in is *not* bad for him. Arguably, this is because Max is not made aware, in some relevant sense, of his state. But notice that it won't simply be a matter of Max being made aware *doxastically*, that is, coming to *believe* that he is undergoing such an unpleasant experience. For, say Max comes to believe he is in a state of the stabby-unpleasant-pain-kind because someone tells him; Max believing that he is in a stabby unpleasant pain state plus his general desire not to be in such states, arguably, does not suffice to make *the unpleasant pain itself* bad. His desire is only *indirectly* about the pain in virtue of the unpleasantness satisfying some descriptive conditions, namely, *'being a stabby unpleasant pain'*. Is it plausible

that what we find so bad about our unpleasant pains is that they are experiences that satisfy some descriptive conditions we more generally want not to occur? I take it that it is not, or at least it is much less plausible that such general desires—rather than *de re* desires—ground the badness we are interested in when we claim that our unpleasant pains are bad. What is so bad about unpleasant pains is the way they *feel*. And if desires are to play a role in that normative explanation, they must bear a direct—*de re*—relation to unpleasantness.

The same is true, I claim, for our motivation to end our unpleasant pains, especially if we assume that our anti-unpleasantness behaviour is a response to the particular *awfulness* of our unpleasantness: surely, what I want my painkillers to stop is *that* instance of unpleasantness; and what most plausibly explains this behaviour is my having a *de re* desire targeting the specific instance of unpleasantness I want not to be occurring. Contrast this with the idea expressed above that we are merely *told* about—without being made *de re* aware of—our unpleasant pain and our having a general desire not to be in that state. It stretches credulity to think that this adequately captures the sort of anti-unpleasantness behaviour we are after when theorizing about unpleasant pain.

So, if we are attracted to desire-based explanations of the normativity and motivationality of unpleasantness then we must invoke desires in the *de re* sense. We can formulate the following requirement for the explanatorily relevant desires:

> ***De re* requirement**
> For any desire D of some subject S, and for any phenomenal episode P of S, for D to explain the normative or motivational features of P, S must be able to be directly aware of P.

Note one final thing about how I've formulated the *de re* requirement. Underpinning the notion of a *de re* desire seems to be the further thought that when we make *de re* judgements about either the objects in our environment or mental objects such as our own experiences, there is the further requirement that one be *directly aware of the object their desire is about*. In our case, a *de re* desire about one's own experience implies that one is directly aware of one's own experience. For instance, perception is the paradigmatic mode of acquaintance for *de re* thoughts about objects in one's environment. Perceiving a juicy red strawberry, I can form the *de re* desire for *that* strawberry on the basis of my direct perceptual acquaintance with the strawberry. The

obvious analogue for *de re* thoughts about our own *mental states* is direct introspective attention or awareness of our own mental states: introspectively attending to the unpleasantness of my pain experience, I form the *de re* desire for *that* experience to stop[82].

But importantly, there is a wildly popular view about the nature of phenomenal experience that directly rules out the possibility of direct introspective awareness of phenomenal experiences: strong representationalism. If that's right, then strong representationalism is incompatible with a highly attractive and intuitive naturalistic account of the normativity and motivationality of unpleasantness, an account that relies simply on the normative and motivational role of desires directed at phenomenal states.

Before I explicitly spell out my argument against strong representationalism about affect, let me make clear the kind of representationalism that is the main target of my argument.

3 Transparency, strong representationalism, and direct awareness

Strong representationalism (hereafter, 'representationalism'), as I'll understand it, is the view that sensory phenomenal episodes, like visually perceiving a red and round tomato, are identical and reducible to (exhausted by) the representational contents of experience. That is, the phenomenal character of experience—the what-it-is-likeness of experience, e.g., the "red-feelingness" or "reddishness" of perceiving the tomato—are identical and reducible to the representational contents of the experience. Furthermore, phenomenal character is *reducible* to representational content in the sense that phenomenal character is *explained in terms of* representational content. That is, our theory of representational content is more fundamental than phenomenal character[83].

Representationalism, crucially, is supported by appealing to the transparency of experience (Harman 1990; Tye 1995b, 1996a, 2003; Dretske 2003). Roughly, it is the view that when we try to attend to features of our experience, we find nothing but the extra-mental objects and properties represented by our experience. The fact that our attention seems to always glom-on to the putative represented extra-mental objects and properties, and never to something

[82] To be explicit, I am assuming that *de re* desires about phenomenal states entail a degree of awareness of those phenomenal states. Later, I raise some objections to my argument which question this assumption.
[83] There are many distinctions to be made regarding representationalism that need not concern us here. For a good overview see Macpherson (2014).

independent of those represented objects and properties, supports the conclusion that the phenomenal character of our experiences just are representational contents. Consider the following example from Michael Tye (1996a). Undergoing the perceptual experience of seeing a blue and round disk in front of me, I can try to direct my attention to features of my experience. But when I attend to the features of my experience, what I notice, so the argument goes, are not the properties *of* my experience, but rather the properties *of the objects* of my experience, namely, the blueness and the roundness *in* the disk. There are no properties found in my experience that are not implicated in the representational contents of my experience. Our experiences are *transparent* in the sense that we 'see through' them to the extra-mental properties represented. We do not (and cannot) attend to any properties over and above the properties experienced as obtaining in the world. So, when I 'attend to the qualities found in my experience' what I am really attending to are the putative properties instantiated in the objects perceived.

Although the transparency of experience provides strong support for representationalism, it is also seen as a *commitment* of representationalism. Here is Amy Kind nicely explaining this commitment:

> If experience were only weakly transparent [that it is difficult, yet not impossible to attend directly to your experience], then we could (at least in principle), avoid seeing through it—and this is in tension with the claim that awareness of surface qualities provides us with our only means for becoming aware that our visual experience has the phenomenal character that it does. Moreover, this same consideration shows that weak transparency not only undermines the above argument for representationalism but also undermines the theory itself… To avoid seeing through an experience to what is represented is to become introspectively aware of some properties of that experience that go beyond its representational contents. (2007: 419)

Since phenomenal character is identical to representational content, there should not be any phenomenal character of any experience that is introspectively available *independent* of representational content. For if there were, then this would be a straightforward case, it would seem, of phenomenal character separating from representational content, and so, a straightforward case of phenomenal character not being identical to representational content. In other words, it would be to prove representationalism wrong.

Note the implications for a theory of introspection. Representationalists hold that phenomenal properties, e.g., unpleasantness, cannot be *directly introspectively attended to*, at least not independent of what our experiences purport to represent. The transparency of experience puts restrictions on a theory of introspection as well as states a metaphysical thesis about the nature of phenomenal episodes. But this is *not* to deny that we can introspect "features" of our experience. Rather, attributing a feature to my experience in introspection is dependent on my prior perceptual *awareness of* some putative *represented* extra-mental object and/or property. For instance, visually perceiving a red tomato, I can make the introspective judgement *that I am experiencing a red tomato* or *that I am having a red-like experience* or *that I am having a reddish experience*. But this introspective judgement about my experience will be dependent on my first having an *awareness of* the represented red tomato in front of me and *not* on my *directly attending to* the phenomenal property itself.

To see this more clearly, consider the distinction between *awareness of* and *awareness that* (Dretske 1999; Tye 2003, 2014a)[84]. Awareness of doesn't require concepts whereas awareness that does. For instance, I can be *aware that* the cake is done baking by being *aware of* the ringing of the timer. The latter need not involve any concepts, but the former does. Representationalists are committed to saying that introspective attention to phenomenal episodes, is, strictly speaking, *awareness that*. We are *aware of* what is being putatively represented in experience—objects and properties—and thereby *aware that* our experience has certain features—the phenomenal counterparts of the represented objects and properties. If it were possible to directly attend to the phenomenal features of our experience—be *aware of*—without first attending to the representational contents of our experience, then this would undermine representationalism[85]. Importantly, introspective judgements of perceptual states or perceptual state self-ascriptions themselves are systematically dependent upon perceptual experiences[86]. In other words, the epistemic order mirrors the metaphysical order. This view about introspection is called the 'displaced perception model' of introspection (Dretske 1995, 2003; Tye 2000,). Importantly, representationalism precludes the possibility of being *aware of* phenomenal properties: they cannot be the direct objects of introspective attention.

[84] Sometimes called 'thing-awareness' and 'fact-awareness', respectively. See also Giustina and Kriegel (2017) for recent discussion.
[85] See also Aydede (2019b).
[86] See also Aydede (2017).

We can, then, following Amy Kind, summarize transparency as follows:

> **Strong Transparency**
> It is impossible to attend directly to our experience (and features thereof), i.e., we cannot attend to our experience except by attending to (being aware of) the objects represented by that experience.

What does this all mean for an account of the nature of unpleasantness? One natural way to spell out a representationalist account of the nature of unpleasantness is along evaluativist lines. Evaluativism is a natural bedfellow with representationalism since evaluativists explain unpleasantness in terms of representational content. In fact, two outspoken evaluativists—Brian Cutter and Michael Tye—wield a strongly representationalist brand of evaluativism, that is, they reduce the phenomenology of unpleasant pain *entirely* to representational content, in particular, to two specific representations: one constituting the pain which consists in a representation of one's body *as damaged*; and the other constituting the unpleasantness which consists in an additional representation of one's damaged body *as bad*[87].

> **Evaluativism**
> (1) Your being in pain consists in your undergoing an interoceptive (sensory) experience (the pain) that represents bodily damage.
> (2) Your pain's being *unpleasant* consists in its additionally representing that damage as *bad for you* (Bain 2017).

For our purposes, note the following. When we try to attend *directly* to our unpleasant pain experiences, our attention rather "slips through" and finds two things instead: i) the represented bodily damage and ii) the represented *badness* of that bodily damage. When I burn my hand on the hot stove, the subsequent unpleasant pain I experience—the throbbing, stinging feeling of unpleasant pain—is a perceptual experience that represents the area where I burned my hand as damaged, as well as that damage being bad-for-me. And gathering myself for a quick second, when I try to directly attend to the feeling of unpleasant pain, my attention will slip through any

[87] There is nothing in the typical formulations of evaluativism—and here I have in mind those given by Brian Cutter and Michael Tye (2011) and David Bain (2013, 2017)—which entails that it must be a species of strong representationalism (SR), although Michael Tye is a fervent outspoken proponent of SR, something which I briefly discuss below. There is, however, a question to what extent evaluativism is an attractive account of unpleasantness given it is not construed along strong representationalist lines. For what it's worth, I think evaluativism becomes much less attractive once it is divorced from strong representationalism. But such issues are beyond the scope of this paper.

putative mental paint to the extra-mental bodily damage *and* to the badness-for-me of that putative damage[88]. Importantly, representationalism rules out the possibility of direct introspective awareness of one's own unpleasant pain.

4 Against representationalism about unpleasantness

Recall, I ended Sect. II by claiming that if we go in for a desire-based explanation of the normative and motivational truths of unpleasantness, then *de re* desires will be required. I noted further that *de re* desires require a kind of direct awareness of the phenomenal states they are about. But now, notice a serious incompatibility here. Representationalism entails that subjects *cannot* be directly aware of their own phenomenal experiences, at least not in the way required to form *de re* desires about them. Hence, representationalists cannot avail themselves to a highly attractive and intuitive naturalistic desire-based explanation of the normative and motivational features of unpleasantness. That is, representationalists cannot appeal to desires to explain what makes unpleasantness *so bad* as well as *why* we take painkillers. Here's the argument.

> **The problem of *de re* desire**
> P1: Desire-based strategies for explaining the normative and motivational features of unpleasantness require *de re* desires.
>
> P2: *De re* desires require subjects to be directly aware of the unpleasantness of their own experiences.
>
> P3: Strong representationalism rules out the possibility of attending directly to (or being directly aware of) one's own unpleasantness.
>
> C: On representationalism, desire-based strategies for explaining the normative and motivational features of unpleasantness fail.

Note two things here. First, this argument generalizes to *any* phenomenal experience for which a desire-based strategy of its normative and motivational features is attractive. Second, and relatedly, note that so-called 'second-order desire' theorists about unpleasant pain, who claim that *unpleasantness* is constituted by a *de re* desire directed at a sensation, cannot be representationalists about those sensations. Those theorists typically direct *de re* desires at the *pain sensation* to explain unpleasantness (e.g., Heathwood 2007). So, insofar as anyone wants to wield a desire-based account of the normativity and motivationality of phenomenal experience, they better not be representationalists.

---

[88] See Tye (2006a) for explicit endorsement of pain experiences being transparent in the above specified way.

Premise one is the attractive suggestion that what makes unpleasant pains so bad and what explains our unpleasantness-directed behaviour—e.g., to take painkillers—is that we desire not to be experiencing *that* unpleasantness we are currently undergoing, that is, our having *de re* anti-unpleasantness desires. Premise two is the thought that in order for our *de re* desires to get referential grip on our unpleasant experiences—something akin to making a successful *demonstrative* reference to those experiences—subjects need to be directly aware of their phenomenal experiences. In other words, you can want the unpleasantness not to be occurring only if it can be directly attended to, i.e., what you want not to occur is *this* feature of your experience. But representationalism's commitment to transparency—premise three—means it can't accommodate that key idea to which desire-based strategies appeal. Therefore, representationalism is incompatible with such desire-based strategies.

In the remainder of the paper, I raise some objections against my argument. In particular, one might question two things: i) to what extent is it really *de re* desires that are required; and ii) to what extent direct awareness of a phenomenal state is required for *de re* desires. After showing that neither objection works, I then briefly consider to what extent representationalists can account for the motivation condition in terms of non-desire-based attitudes. I sketch why I think attempting to accommodate the motivation condition in such a way spells trouble. Then, I move on to sketch where this leaves representationalism with respect to explaining the normativity and motivationality of affect. And finally, I conclude.

5 Objections and replies

The key idea underlying the argument so far is that *direct awareness* is necessary if desires are to make a normative and motivational difference. But the intuition underlying this thought might be the following: bad things can't happen to us if we aren't aware of them. And surely, that isn't something I can take for granted. For surely, it is highly controversial whether awareness makes a *normative* and *motivational* difference. For instance, intuitively, it is still *bad* for the unassuming spouse that their partner is cheating on them, even though they are completely unaware of the cheating. The same will go for our motivations. I might be motivated towards something without being aware of what's motivating me. Presumably, what *makes it bad* is that the spouse has a general—perhaps *de dicto*—desire against infidelity or against being lied to, and it's the frustration—unbeknownst to her—of that desire which makes it bad for her.

Similarly, what might *move me* to avoid a particular person is that I find them unpleasant to be around, even though, again, I am *unaware* of the unpleasantness they cause me.

To be clear, the reply is a challenge to my argument since the incompatibility I argued for between representationalism and the formation of *de re* desires is really an incompatibility between *representationalism* (with its commitment to strong transparency) and the assumption that *de re* desires entail *direct awareness of a phenomenal episode* or *demonstrative reference to a phenomenal episode*. But once we recognize that we can have normatively and motivationally relevant desires *without* being directly aware of the phenomenal episodes they are about, that incompatibility vanishes. Hence, my argument fails.

But I doubt this move works. That is, I doubt that desires formed without direct awareness of the unpleasantness of the experience can sufficiently explain the normative and motivational features of unpleasantness. There are two ways in which one might elaborate this challenge. First, one might think that dropping the necessity of direct awareness allows for the possibility that *de dicto* desires suffice to accommodate our two constraints. And second, one might think that it is possible to form *de re* desires without being directly aware of the objects they are about. After all, representationalists can still attribute, in introspection, phenomenal concepts to experience, concepts that, as we saw in Sect. III, are systematically dependent on the perceptual concepts implicated in our epistemically prior perceptual awareness. That is, opponents of my argument can challenge premise one and premise two, respectively. Let me take each point in turn.

5.1 Against premise one

Premise one states:

> P1: Desire-based strategies for explaining the normative and motivational features of unpleasantness require *de re* desires.

The reply against premise two is that *de dicto* desires—desires that take the following form: I don't want (to undergo) unpleasant experiences *whatever else they happen to be*—can adequately capture the normative and motivational features of unpleasantness. I think appealing to *de dicto* desires will require *de re* desires anyways, at least when it comes to forming *de dicto* desires about our own experiences. But setting this to one side, I have two worries. Notice, first,

that for an unpleasant pain to be bad for someone (and motivating)[89], one would need to be able to form the relevant *de dicto thoughts* containing the relevant concepts to ensure that the unpleasant pain is bad. In other words, *de dicto* desires necessarily contain conceptual content. This, however, unattractively makes the badness of unpleasant pain dependent on whether one possesses the pertinent *concepts*. Second, *de dicto* thoughts can't capture the degrees to which unpleasantness can be bad for us, a fineness-of-grain that *is* capturable by *de re* desires. I'll take each point in turn[90].

To appreciate the first point, notice that *de dicto* desires necessarily involve concepts[91]. In this case, unpleasant pain won't be bad for non-human animals and young children that lack the appropriate conceptual repertoire, since they will be unable to form the relevant *de dicto* anti-unpleasantness desire. *De dicto* desires are too demanding. If my *de dicto* desire not to be undergoing unpleasant pains is, in some sense, dependent on my being able to conceptualize unpleasant pains, namely, wielding the concepts UNPLEASANTNESS and PAIN, then those without the relevant concepts will be unable to form the relevant *de dicto* desire, and therefore, will fail to have unpleasant pains that are bad-for-them (or reason-providing). But surely, as is familiar from normative ethics, unpleasant pains *are bad independent of whether one can conceptualize their own unpleasantness and pain*. Appealing to *de dicto* desires, therefore, overintellectualizes what is needed for unpleasant pains *to be bad*.

But even jettisoning over intellectualization worries, I doubt still that appealing to *de dicto* desires can adequately account for the normative and motivational conditions of unpleasantness. What is bad about unpleasant pain and what motivates me to end it is the *particular* way it *feels*, or at least, what our desire makes bad and moves us to eliminate is a specific feeling we want not to be undergoing. And we might further wonder how, if at all, *de dicto* desires can ground

[89] In what follows, for ease of exposition, I mostly speak only of the badness of unpleasantness (the normative condition) but assume that what I say in that case applies *mutatis mutandis* to our responsiveness to our unpleasantness (the motivation condition) unless otherwise stated.

[90] The point about *de dicto* desires requiring concepts might not be so forceful a reply in the case of our motivation to end our unpleasantness, e.g., to take painkillers. That does seem like a conceptually demanding action. I do, however, think the point about capturing the fineness-of-grain applies equally to the normative condition and the motivation condition.

[91] One might think that forming *de re* desires also involves the use of concepts, albeit *de re* or *demonstrative* ones. If so, then the over-intellectualization worry I am about to mount against *de dicto* desires, will apply to *de re* desires as well. But in the name of charity, I take it that forming *de re* desires does not entail wielding certain concepts, and that one should understand *de re* here in the less demanding sense. For instance, I take it that young infants and non-human animals can very much desire things without wielding any conceptual schema. See Ventham (2021) for discussion.

those features without being *directed* at such a feeling. To get a better sense of this point, consider the example of surprise badness.

> **Surprise Badness**
> Max the masochist is a perfect masochist. He has no anti-unpleasantness desires, but lots of pro-unpleasantness desires. His unpleasant pain experiences are, according to desire strategists, *good for him*. They give him at least a pro tanto reason to continue their existence, or to seek out their attainment. One day, Max encounters Sam the sadist, a new member to his local sado-masochistic community. Sam begins to inflict pain on Max in a way Max has never experienced; in a way that is more *stabby* than typical. On this occasion, Max finds himself encountering his unpleasant pain differently from how he usually does. In this case, he finds himself *hating* the unpleasant pain, even finding *it* especially *bad.* Presumably, Max has formed a new *anti*-unpleasantness desire. This new desire—to Max's surprise—is why he now hates the experience.

We can ask: what kind of desire could this new anti-unpleasantness desire be? For starters, Max's surprise badness is a product of having experienced a new kind of unpleasantness, a new *stabby shade* of unpleasantness. Can *de dicto* desires account for the badness of this new shade of unpleasantness? I claim they cannot. Max's new shade of unpleasantness entails a fineness-of-grain not capturable through mere concept application. He possesses the concept UNPLEASANT, PAIN, and perhaps even STABBINESS, but notice that this isn't fine-grained enough to capture the unpleasantness in question, the new stabby *shade* of unpleasantness, and to explain why Max finds *that* bad. For, as we've stipulated, Max holds pro-unpleasantness desires of the general, *de dicto* sort which are *good* for him and for which he is positively motivated to continue or seek out. And this new shade of unpleasantness, although it satisfies the description STABBY UNPLEASANT PAIN is nonetheless *bad* for Max due to the *shade, degree, intensity,* or what have you, of the unpleasantness in question. In other words, one might like one kind of unpleasantness and hate another (in the relevant sense of like and hate) and our desire-based strategy better accommodate this possibility, i.e., that some of the relevant anti-X-desires won't just be anti-unpleasantness desires but anti-unpleasantness374 and anti-unpleasantness765 desires. We are capable of experiencing a variety of kinds of unpleasantness that outstrips our capacity to articulate those experiences in terms of non-*de re,* or non-demonstrative, concepts—concepts representationalists are forced to embrace due to strong transparency—and surely those instances of unpleasantness will still be bad for us, and we will still respond to those bad states when we take painkillers. Therefore, *de re* desires are indeed required if desires are to capture the normative and motivational conditions of unpleasantness.

But this reply leaves open the possibility that subjects might form *de re* desires without attending to the phenomenal states they are about. In other words, opponents of my argument will question premise two.

5.2 Against premise two

Premise two states:

> P2: *De re* desires require subjects to be directly aware of the unpleasantness of their own experiences.

Now, there are two ways in which one might form a *de re* desire without being directly aware of the phenomenal feature the desire is about. First, it might be suggested that if subjects can undergo unconscious phenomenal experiences, then it is plausible for them to have desires for which they are not conscious of. In other words, subjects can, and often do, have unconscious desires which are motivational (see Pettit and Smith, 1990). So, why can't they have unconscious desires for states they cannot attend to?[92] But appealing to unconscious, or sub-personal, involuntary desires won't avoid the problems presented by strong transparency. What, we can ask, is the desire targeting in this case? Strong transparency is a metaphysical thesis about phenomenal experience which has the general implication that nothing (conscious or otherwise) can directly target phenomenal experiences. The idea that unconscious desires could do that would count as violating strong transparency, hence, would refute representationalism.

Second, subjects might form *de re* desires in the following way:

> (I) "(Introspectively) I am now having an unpleasant experience (pain). I want *it* to stop."

Here, it seems as though "it" expresses a *de re* desire. And granted we are not directly aware of the features we attribute to our experience in our introspective judgement, this seems to be a case of forming *de re* desires without direct awareness of a phenomenal episode. In this case, our putative *de re* desire to want the experience to stop seems dependent on us categorizing the

[92] This seems to be the view of, e.g., Ventham (2021) and Feldman (2018). But compare Heathwood (2007). Thanks to an anonymous reviewer at *Philosophical Studies* for raising this point.

token experience under the phenomenal concept UNPLEASANT without any direct awareness of the property in question. Note, first, that at least in the case of the normative condition—unpleasantness's badness—this, again, will unattractively make the existence of an *evaluative property* dependent on whether a subject does or does not wield the relevant concepts. Second, and more importantly, this reply faces the same difficulty the *de dicto* strategy faced above. That is, if our *de re* desires—putatively expressed by the 'it' above—are formed based on our introspective judgements which categorize the token experience under a phenomenal concept UNPLEASANT EXPERIENCE, then there will be unpleasant experiences which outstrip our ability to categorize those experiences in terms of the phenomenal descriptions available to us: the phenomenal concepts at hand (those expressed in (I) above) won't properly capture the degrees of unpleasantness one can undergo, and hence, won't generate de re desires that fully capture the normative and motivational conditions of unpleasantness.[93] One way to avoid that issue is for subjects to be able to make introspective judgements like the following:

(I*) "(Introspectively) I want *this* (unpleasant) feature of my experience to stop."

But that judgement expresses a demonstrative thought which will require direct awareness of the property referred to, namely: the unpleasantness. And that is precisely what is impossible given representationalism's commitment to strong transparency[94].

But the representationalist has a reply. Although *de re* desires and demonstrative thoughts are intimately connected to direct awareness, they do not entail it. After all, there seem to be cases

[93] Tye (1996: 52) even speaks of the "general and uninformative way" we *categorize our experiences* in introspection, according to representationalism.

[94] Note here an ad hominem point against Michael Tye. Tye himself, a fervent proponent of strong representationalism, wields a desire-based explanation of the normativity and motivationality of unpleasantness (Cutter and Tye 2014). But Tye (2014a) explicitly precludes the possibility of forming *de re* attitudes about phenomenal episodes when arguing against qualia realism (and for his brand of representationalism), the view, roughly, that there exist intrinsic features of our experience that are not representational. He claims:

> If one is aware (de re) of some entity, one's awareness directly puts one in a position/enables one to form de re cognitive attitudes with respect to that entity… Now forming a de re cognitive attitude with respect to a thing directly on the basis of one's awareness requires attending to that thing… If one cannot attend to a thing…then one is not aware of that thing (45).

So, either Tye gives up his brand of representationalism or his desire-based explanation of the normativity and motivationality of unpleasantness.

of successful demonstrative reference without direct awareness of the item referred to. These cases involve *indirect* demonstrations and plausibly suffice to get *de re* desires to referentially glom on to phenomenal properties, however finely grained they happen to be. Consider two hunter-trackers on the trail of a possibly wounded deer. Dina says to Marge, gesturing toward some animal tracks: "*This* buck *was* limping. You got *him* after all Harry!" Supposing that here there is an *informational link* of the appropriate sort between the tracks and a particular animal that was in fact limping, Bill's thought is plausibly de re with respect to that thing (the animal) that was claimed to be limping—even if it weren't a buck, or a deer, or an animal indeed. But there is no direct awareness of that thing in this case.[95]

Similarly, we can form *de re* desires against our unpleasant experiences via an *indirect* demonstration on the basis of some epistemically prior perceptual demonstration.

Recall the displaced perception model (DPM) I alluded to in Sect. III. There we saw that, according to representationalism and its commitment to strong transparency, for subjects to make introspective judgements about their own phenomenal states they must first make *perceptual* judgements about the putative objects and properties represented in their first-order experience. Importantly, for representationalists, our demonstrations to the phenomenal properties of our experiences—ones expressed in (I*) above—will systematically depend upon an *informational link* between the concepts deployed in *direct* perceptual demonstrations and the concepts deployed in *indirect* phenomenal ones. This allows the representationalist to potentially overcome the issue I raised above regarding the fineness-of-grain of phenomenal properties and the phenomenal concepts available to the representationalist. As long as we can indirectly demonstrate the exact level of grain with respect to the phenomenal properties of our experiences *via* directly demonstrating the exact fineness-of-grain of what our experiences putatively represent, then we will be able to form *de re* desires about our phenomenal episodes on the basis of such direct (perceptual) demonstrations. In other words, so long as our phenomenal demonstrations are anchored in our epistemically prior (direct) perceptual demonstration then we can secure the explanatorily relevant *de re* desires. Fineness-of-grain is no longer an issue for our phenomenal demonstrations can piggyback onto the direct perceptual demonstration.

[95] Thanks to an anonymous reviewer at *Philosophical Studies* for pressing me on this point and for the example.

For instance, in Surprise Badness, Max can indirectly demonstrate the exact level of grain of his new stabby-unpleasant experience on the basis of a prior demonstration to the putative represented bodily-badness of his experience. Again, since representationalists are committed to strong transparency, they are committed to the following introspective story about phenomenal qualities: anything we can come to know phenomenally about our own experiences is dependent on knowing what our experiences represent (Aydede and Fulkerson, 2014). Our attention always, recall, finds the representational properties of our experiences; we then conceptually construct our introspective judgements about our phenomenal states *on the basis* of our perceptual awareness of what our experiences putatively represent. So, something like the following *is* compatible with strong transparency

(I*) "(Introspectively) I want *this* (unpleasant) feature of my experience to stop."

so long as I first perceptually demonstrate bodily-badness. Undergoing an unpleasant pain, my attention first and foremost focuses on what my experience represents, namely, the represented bodily badness constitutive of the affective phenomenology. I then directly (perceptually) demonstrate the exact fineness-of-grain of the badness-for-me experienced as being instantiated on a particular part of my body: "*that* is bad-for-me". I am then able to *indirectly* demonstrate the unpleasantness I am currently experiencing by forming *a phenomenal belief about my experience* on the basis of my prior perceptual demonstration*:* "My experience has *this* (unpleasant) quality". Subsequently, the content of my *de re* anti-unpleasantness desire is the conceptually articulated phenomenal content dependent on the perceptual concepts applied to the objects and properties putatively represented in my experience, concepts which partly constitute my demonstrative, *de re* perceptual belief. That belief—the perceptual one—is the basis for my *de re* anti-unpleasantness desire.

I doubt, however, that DPM does provide a good enough basis to account for the normativity and motivationality of unpleasant pain. I don't have enough space to mount a full attack against that account but note three important things. First, DPM assumes that subjects can plausibly perceptually demonstrate bodily-badness to a relevantly fine-grained level. Second, the phenomenal belief (and hence, *de re* desire) formed on the basis of one's prior perceptual awareness is *inferentially* arrived at. And third, representationalism (and DPM) entails that there is an informational link between the perceptual concepts (p-concepts) invoked in the

epistemically prior perceptual demonstration and the phenomenal concepts (i-concepts) invoked in the indirect introspective demonstration. Hence, we conceptually construct introspective judgements—come to know *what* our experiences are like by first attending to *what* they represent. Each point has its problems. Let me take each in turn.

First, representationalists are committed to the idea that subjects are able to perceptually demonstrate bodily-badness. This seems dubious, especially when we compare it to paradigmatic cases of perceptual demonstration. Deciding on what colour to paint my newly renovated kitchen, I point to the shade of denim blue on the colour wheel and say to my partner: "Let's paint it *that* colour". Do we really have an analogous way of demonstrating bodily-badness? Representationalists will insist that it is nonetheless possible to perceptually gesture to one's bodily-badness in such a manner. But even if it's possible to gesture to one's bodily-badness, it's far from obvious whether we can demonstrate and distinguish between various *degrees* of bodily-badness, like we can shades of blue, required to accommodate cases like Surprise Badness. If I can't appropriately discriminate between degrees of bodily-badness, then I will fail to form phenomenal beliefs, and hence, *de re* desires capturing the fineness-of-grain of my experiential unpleasantness.

Furthermore, representationalists make our introspective knowledge of our phenomenal experiences importantly *indirect*. The indirect nature of DPM has been discussed and criticized elsewhere.[96] Here I note two problems. First, the indirectness of DPM entails that the relevant introspective judgements are *inferentially* formed. Second, for indirectness to yield introspective knowledge—i.e., to introspect the qualitative content of an experience and subsequently form the relevant *de re* desire—representationalists are committed to the claim that there is an informational link between phenomenal concepts like UNPLEASANTNESS and perceptual concepts like BODILY-BADNESS or HARMFULNESS. These points seriously undermine the possibility that we form the explanatorily relevant *de re* desires against unpleasant experiences in a manner consistent with strong transparency.

Regarding the inferential nature of DPM, Aydede (2003) has convincingly argued against the plausibility that introspection of our phenomenal experiences is inferential. His claim is roughly the following: there is no connecting belief which warrants the inference from one's

[96] See, e.g., Aydede (2003).

epistemically prior direct perceptual judgement—e.g., that *this* ball is red—to the indirectly introspected phenomenal belief—e.g., that I am now experiencing/seeing the ball as red. In the affect case, the connecting belief would have to be something like the following:

> (CB): If that bodily-damage is bad, then I am now experiencing that bodily-damage as bad.

Note two things here. First, as Aydede points out, beliefs like this seem to be false and known to be so (2003: 57). There are various ways in which one could come to know the antecedent without thereby perceptually experiencing one's bodily-damage as bad. But to hold a belief like (CB), or any alternative which employs perceptual concepts like BODILY-BADNESS in an inference to what one's experience *is like*, is "epistemically irresponsible". Aydede claims: "*as a matter of fact* most people (perhaps all), being more or less epistemically responsible, just lack a connecting belief like (CB)" (Ibid.). It's highly unlikely that *all* subjects hold such background beliefs. Do we really think that subjects who form anti-unpleasantness desires *must* possess such conceptually articulated beliefs? So, the DPM is empirically and epistemically dubious.

Note, also, what the resulting belief would be:

> (B): "I am now experiencing *that* bodily-damage as bad."

I would then apparently re-appropriate the content of that judgement (e.g., EXPERIENCING-BODILY-BADNESS or its ilk) to finally produce the following:

> (I*) "(Introspectively) I want *this* (unpleasant) feature of my experience to stop."[97]

Again, it is highly questionable whether subjects employ these sorts of concepts and engage in these sorts of processes when forming experience-directed *de re* desires against unpleasantness. If we agree with Aydede's conclusion, then our painkiller-taking actions will seem seriously suspect. In the case where I know that what my experience represents—i.e., represents-bodily-

[97] Of course, subjects could come to simply hold the following conditional connecting belief instead of first holding (CB). They could hold (CB1): If that bodily damage is bad, then I am now having an unpleasant experience. But I take it that the connecting belief (CB1) is even more implausible than (CB) since the conceptual connection between the antecedent and the consequent in (CB1) is highly questionable.

damage-as-bad—is false, then my *de re* desire against the unpleasantness of my experience will be based on a (knowingly) false belief. But notice that this story is seriously incompatible with the motivation condition. Our painkiller-taking actions are no longer paradigms of rational action. The structure of our painkiller-taking actions, according to representationalists, will sometimes entail that we are acting irrationally when—knowing that the perceptual scene before us is falsidical—we nonetheless form the desire for our unpleasantness to stop, and accordingly, *irrationally* take a painkiller[98]. That is a very unwelcome consequence.

Second, even if one thinks that the inference from (CB) to (B) is epistemically and rationally kosher[99], the move from (B) to (I*) seems deeply problematic. Recall that in order to arrive at (I*) I need to conceptually re-appropriate the perceptual concepts employed in my prior perceptual demonstration associated with (B). But that one can reasonably move from the conceptual content of (B) to the conceptual content of (I*) is highly dubious. As Aydede and Fulkerson (2014) point out, there is no proper informational link between experienced bodily-badness and experiential unpleasantness.[100]. It's just not true that in order to possess the concept UNPLEASANTNESS, and therefore, to introspect the phenomenal property of unpleasantness, I must first possess the concept BODILY-BADNESS. For it is clear that one may completely lack the relevant concept BODILY-BADNESS and its ilk yet still possess and apply the concept UNPLEASANTNESS to one's experience. As Aydede and Fulkerson claim:

> As should be obvious, there is no conceptual connection between the concepts PAINFUL [UNPLEASANT] and HARMFUL [BODILY-BADNESS], and for that matter, REPRESENTS-HARM [REPRESENTS-BODILY-BADNESS].

In short, possessing the concept UNPLEASANT does not entail that one possesses the concept BODILY-BADNESS, and hence, cannot be conceptually articulated out of the prior perceptual

---

[98] Rather, one might think that when we take painkillers, we *directly respond* to the unpleasantness (and its badness) of the experience. But, of course, that isn't compatible with strong transparency.

[99] For instance, one might be inclined to make a move similar to Evans (1982) whereby in order to move away from any sort of endorsement of our first-order perceptual experience in the case where we know it is falsidical, we exclude any knowledge we may have that is "of an extraneous kind", and make a judgement about how things *seem to us here and now* with the added prefix 'It seems to me as though…' (227-229). Although this does seem to entail that desires based on how things seem to one here and now avoid the problems of irrationality mentioned above, it is still highly dubious that subjects actually engage in this sort of inference in order to avoid forming irrational desires. Also, in this case, subjects, I take it, would need to *know* that their affective experiences are indeed false, i.e., that the bodily damage putatively represented is not bad-for-them. Again, this is a highly complicated and questionable process which subjects seem not to engage in.

[100] For Aydede and Fulkerson, the corresponding representational property is represents-harm.

contents in order to form the relevant *de re* desire.[101] Contrast this with the idea that possessing the concept REDISHNESS or RED-LIKE or REPRESENTS-RED entails that one possesses the concept RED. One cannot possess the concept REDISHNESS—and make introspective judgements and form desires about their visual experiences of redness—without possessing the concept RED and applying it *first* in one's epistemically prior perceptual judgments. But this is not the case for unpleasantness. That is, subjects can apply a concept like UNPLEASANTNESS without first applying any corresponding concept, or for that matter, "without having the slightest clue what that quality may represent, or even whether it represents anything" (Ibid., 195). And to insist that subjects nonetheless *do* attribute phenomenal concepts like UNPLEASANTNESS without first applying perceptual concepts to their experiences such that those phenomenal concepts really are only picking out intentional properties is just to admit that *it is possible to become directly aware of properties of our experience without first becoming aware of what those experiences putatively represent*. And that is a direct violation of strong transparency.[102] Hence, P2 must be true.

Representationalism's incompatibility with desire-based strategies in value theory, metaethics and moral psychology constitutes a major blow to the theory. Representationalists are barred from wielding any desire-based explanation of the normativity and motivationality of unpleasantness. And so, the explanatory options available to the representationalist interested in the normativity and motivationality of unpleasantness are severely restricted. Not only will representationalism look less attractive to those in the philosophy of mind interested in explaining affective phenomenology, but it will importantly alienate a whole host of philosophers who wield, in the first instance, desire-based strategies in value theory, metaethics, and moral psychology.

I now want to briefly sketch ways in which representationalists might accommodate the motivation condition which *do not* appeal to desires, and the potential problems they would face.

---

[101] Note also that UNPLEASANTNESS is not a perceptual concept.

[102] Note, as should be familiar at this point, that whether an unpleasant pain is bad for someone will depend on whether they can wield the conceptual scheme outlined above and engage in the various processes needed to arrive at *de re* desires targeting their own experiences. This is highly implausible.

6 Non-desire-based explanations of the motivation condition

Let us take for granted the fact that unpleasantness is bad.[103] Can representationalists accommodate the motivation condition without appeal to desire? Recall the motivation condition:

> **Motivation Condition**
> Subjects, when undergoing unpleasant pains typically rationally *respond* to the reasons provided by those unpleasant pains, hence, why they take painkillers.

As a reminder, representationalists fail to accommodate the motivation condition by appealing to desires since, on representationalism, we cannot in the relevant sense want the experience to stop. Here's another way of putting the problem. Wanting my headache to stop, I reach for the painkillers. This is a paradigmatic case of responding to a reason I have, a reason, namely, to end my unpleasant experience. But notice the story representationalists must give. In attempting to respond to one's reason, one's attention only ever finds the reason provided by the *extra-mental bodily damage* and never the reason provided by the unpleasantness itself: we cannot directly attend to the reason provided by unpleasantness, hence cannot directly rationally respond to the unpleasantness. But we clearly *do* directly rationally respond to our unpleasant pains: we take painkillers. Therefore, representationalists fail to accommodate the key fact that subjects rationally respond to unpleasant pain[104].

But underlying this thought seems to be the further idea that in order to rationally respond to something—to act for a good reason, e.g., to take painkillers—one must attend to the features that make it a reason. But this is questionable. To illustrate, consider the following. Say there is a torturer in a windowless room torturing someone. They are causing them a serious amount of pain that is surely bad for the person being tortured. You are outside of the room and, unbeknownst to you are two things: the torturing and the fact that you could end the torturing

[103] See Bain (2017) for a potential strongly representationalist strategy. See Jacobson (2019a) for criticism. I don't think Bain's (2017) strategy will be available to the strong representationalist anyways since his view seems to entail the possibility that two different subjects can have the same representational content yet differ in their phenomenal experiences.

[104] For a similar, although importantly different argument, see Boswell (2016). Boswell considers the tension between, on the one hand, subjects directly responding to the reason provided by unpleasant pain, and on the other, evaluativism's commitment to strong transparency. Boswell, however, fails to appreciate that *de re* awareness is precluded by evaluativism (given that he seems to be interpreting it as a strongly representationalist account), leading him to fail to see that various responses he gives on behalf of the evaluativist (as well as his own preferred response which entails *de re* attitudes) are incompatible with strong transparency, and hence, his interpretation of evaluativism.

by entering the room. Now, say someone nearer to the room can hear the screams of the person being tortured, knows *you* could stop the torturing by entering the room yet themselves are unable to enter the room. They yell to you: "There is someone being tortured in there! You can save them by entering the room!" Believing them, you then enter, saving the person being tortured. Here, you have responded to a reason for action *without attending to anything that makes the situation reason-giving*[105].

Can representationalists avail themselves to a similar explanation of our motivation to end our unpleasant experiences that does not involve attending directly to the reason-giving features of the experience? I think it is doubtful at best. For starters, if there are cases (and surely there are) where we respond to a reason by attending to the features that make it a reason, then our responsiveness to the reason provided by our unpleasantness will surely be such a case. And recall, what my argument above showed was that *whatever else we say about the normative and motivation-constituting attitude* it must be *de re*.[106] So, representationalists are forced to accept that the motivating attitude is not *de re*.

However initially plausible we might now find that idea, in what follows, I assume, for the sake of argument, that the relevant motivating attitude is not *de re.* With all that in mind, here is what is available to the representationalist to accommodate the motivation condition.

Eschewing *de re* attitudes altogether as an explanation of our rational motivation to take painkillers, representationalists have two options: either our rational motivation to take a painkiller to end the badness of our unpleasant pain is explained by i) a further second-order *experience* of the badness or reason-providingness of our unpleasant pain or ii) a second-order *de dicto judgement*[107] that our unpleasant pain is *bad* or reason-providing. Call each strategy the *experiential strategy* and the *judgemental strategy*[108], respectively.

---

[105] We should distinguish here between the *subjective* reason you have—the reason you have in virtue of your epistemic position—provided by the other person's testimony and the *objective* reason—the reason you have independent of your epistemic position—provided by the badness of the torturing and your relation to its cessation. The case of the torturer is supposed to highlight the possibility of acting for an objective reason without attending to the features that make it a reason.

[106] Cases like Surprise Badness show this.

[107] In what follows, I drop the *de dicto* qualification for ease of exposition.

[108] Of course, since we are, for the sake of argument, considering the plausibility of some *de dicto* attitudes, one might be inclined to now appeal to the plausibility of *de dicto* desires. On such a picture, it would be a mental state pair consisting of a *de dicto* desire and a *de dicto* judgement which explains the motivation to end the badness. On that strategy, the point about motivational internalism below won't apply. But what follows after that with respect

I argue trouble lies in both directions. First, both strategies commit representationalism not only to a version of motivational internalism but a highly controversial one. Second, the strategy some evaluativsts (who, again, tend to be representationalists) have found most promising—the experiential strategy—is unavailable to them. And finally, left with the judgmental strategy, representationalism is unable to make sense of *why* subjects would make that relevant motivation-constituting, evaluative judgement in the first place. Let me take each point in turn.

First, notice that each strategy appeals *only* to truth-apt mental states to explain the motivation condition. This entails for its plausibility a version of motivational internalism, the view roughly that, necessarily, evaluative experiences or judgements (mental states with evaluative content) motivate (Smith 1994). This, I take it, is a somewhat costly price to pay. But note, further, the explanation representationalism must give of the relationship between our affective experiences or judgements and the motivation to end the target mental state our experiences or judgements are about. It seems The Humean Theory of Motivation, which states that beliefs are motivationally inert, and therefore, must be accompanied by desire or desire-like motivational states, is unavailable to the representationalist. Absent accompanying desires, the representationalist must adopt a version of motivational internalism that accepts the idea that *purely truth-apt states* with evaluative content can motivate. So, not only do representationalists have to accept a controversial metaethical thesis in motivational internalism, but they have to accept perhaps an even more controversial version of that thesis: non-Humean or purely truth-apt motivational internalism[109].

Second, some have attempted to make sense of the idea of purely truth-apt motivational states by appealing to evaluative *experiences*. Bain (2013) himself attempts to make sense of purely truth-apt motivational states to explain our *bodily-directed* motivation—to tend to our damaged bodily condition—by distinguishing two vehicles of truth-apt motivational content: *experiential* and *judgemental*. In explaining our bodily-directed damage avoidance motivation—e.g., our motivation to quickly remove our hand from a hot stove—Bain claims it is a rather natural idea

---

to how subjects might make a judgement about the badness of their unpleasant pain will apply equally to a judgement-desire pair strategy. But in what follows I drop discussion of desire.

[109] Of course, the matter is more serious (and complicated) than this. For even some anti-Humean versions of motivational internalism will be unavailable to the representationalist since those versions will entail that the relevant motivating attitude be *de re*.

to think of *evaluative content* within an *experiential episode* as being inherently *motivational*. He claims:

> Some will insist that there is nonetheless something objectionably ‘‘queer’’ about the idea of episodes that are both motivational and entirely truth-apt, even if they are experiences. But our version of that idea rather strikes me as utterly natural: when the badness for you of a state of your own body is impressed on you, this— independently of further desires—defeasibly motivates you to do something about that bodily state. (S84).

The plausibility of appealing to experientiality to account for the normativity of unpleasantness—that is, what is so *bad* about your unpleasant pain is that it is a *perceptual encounter* with the badness of your bodily condition—relies on the plausibility of purely truth-apt experiences with evaluative content to be *motivationally* and normatively explanatory. Note, importantly, that the appeal to evaluative (representational) content in the first-order experience is supposed to explain our *world-directed motivation* and *not* our experience-directed motivation to stop the unpleasantness. This might be a plausible way for representationalists to accommodate the normative condition, as well as a separate, *world-directed* motivation condition, namely, to tend to one's bodily damage.[110] But can it account for our *experience-directed* responsiveness to that normativity, namely, to end the *unpleasantness* itself? I think not. With the above in mind, we get something like this:

> When the badness of your unpleasant experience is impressed on you, this—independently of further desires—defeasibly motivates you to do something about that experience.

Can we make sense of the idea that the badness of your *experience impressing itself on you* motivates you to eliminate it? For starters, that idea would be strengthened if we could make sense of the idea that we have further *second-order* experiential episodes in which it is *not* the badness of your *bodily state* impressing itself on you and motivating you to do something about it but rather *the badness of your experiential episode*—the unpleasantness—impressing *itself* on you and motivating you to do something about *it*. That is, you have a representational experience of the badness of your first-order unpleasant experience which then motivates to you to end it. But that would entail a quasi-perceptual account or inner awareness account of our phenomenal episodes, something strong transparency precludes. It seems highly dubious that we could

[110] Although, see footnote 23.

experience the badness of our first-order experiences without invoking something like direct awareness of the first-order experience itself.

So, one of the more recent attempts to make sense of purely truth-apt motivational states (at least according to one evaluativist) is unavailable to the representationalist. That leaves us with the judgemental strategy: we respond to our unpleasant pain purely based on a judgement that our experience is bad or that we ought to cease our experience's existence (again, noting the non-*de re* nature of the judgement)[111]. Now, granting for the sake of argument that this strategy doesn't entail any sort of direct awareness of, or demonstrative reference to, our first-order experiences, problems linger. If subjects can't directly attend to the evaluative features of unpleasant pain (its badness), then why would they make such evaluative judgements about unpleasant pain in the first place? Even granting that such a judgement that one's unpleasant pain is bad *can motivate* (independently, of course, of desires), we might wonder why one would come to make it. How, I ask, is that judgement justified, in the eyes of the subject, *given* it must be *mediated* by our awareness of what our perceptual experiences putatively represent?[112] Representationalism cannot account for the justification of that judgement.

To see this charge more clearly, note an asymmetry between our introspective judgements of pain perception and colour perception. Say I am *directly* perceptually aware

that this tomato is red

and then make the introspective judgement

that my experience is *of a* red tomato

or

that I am having a *reddish* experience

This much seems straightforward. But notice what the analogous judgement is in the pain case (assuming an evaluativist account of unpleasant pain):

that my bodily damage is *bad for me*

or

that I am having an experience *of my body as bad for me*

But the judgement *that we need* is the following

that *this experience* is bad for me

---

[111] For the sake of argument, we can bracket worries about fineness-of-grain.

[112] Some might even balk at the suggestion that our access to the reason provided by our unpleasantness is *indirect*.

and not that *the object of my experience* (my damaged bodily condition) is bad for me or seems bad for me. But there is a deep disanalogy between the colour perception case and what is needed in the pain perception case. In the colour case, my perceptual judgement that the tomato in front of me is coloured red only warrants the introspective judgement that my experience is *of the red tomato*, or that it is *red-like,* or that I am having a *reddish experience*. It does not warrant me, and nor should it, in making the introspective judgement that my *experience is literally* coloured red. For a subject to make the relevant motivating judgement (and ultimately, to respond to the reasons provided by their unpleasant pain), they need to be able ascribe *not* some *phenomenal counterpart of that property* but rather the property perceived—the badness—to the experience itself. But why would we do this in the pain case, when not only do we *not* do this in the colour case, but we *should not* do it in the colour case. In short, our introspective judgements, you might think, especially if you think representationalism is true, rely for their content on the world-directed content of the experiences they are judgements about. But this won't explain how subjects comes to make a judgement about the *badness* of unpleasantness, and hence how subjects are motivated—independently of desires—to eliminate their unpleasant experiences.

To put it another way, one might point out that we *do* ascribe the property 'red' to our experiences of red things. It seems natural enough to say: "I am having a *red* experience." But this is ambiguous between two readings of 'red'. On one reading, 'red' refers to the extra-mental property of *redness* perceived *in the tomato*, presumably a determinate of the property red. On another, 'red' refers to a determinate of something *other than* the property *red*, presumably a phenomenal property best described as "red-like" or "reddish". But importantly, these properties are entirely distinct properties: one is a perceptible extra-mental property and the other is an introspectable phenomenal property. That is, the sense in which *my experience is red* and the sense in which *I experience something being red* are different, is as follows: The redness of my experience—more appropriately, the *red-like quality* of my experience—and the perceived redness are different determinates of *distinct* determinables; again, one is a token instance of an extra-mental property and the other a token instance of a mental property. However, the sense in which *my experience is bad* and the sense in which *I experience my body being in a bad state* are different, is as follows: The badness of my experience and the badness of my body are different determinates of the *same* determinable. They are, arguably, both token instances of

prudential badness: bodily damage is bad-for-you and negative affect is bad-for-you.[113] What is it about the perception of one kind of determinable—namely, an evaluative one—that warrants one in ascribing the same determinable to one's mental states?

There is nothing about the normativity of the extra-mental world that can provide us with information about the normativity of our minds. In short, it is difficult to see what the basis of our *experience-directed judgement* could be, especially given that this judgement is supposed to be a *rational* one to make: it's a response to the badness of our unpleasantness.

To see a potential representationalist reply, consider, again, the inferential nature of the judgement representationalists are committed to. Here's what I think the representationalist needs to say. Call the badness perceived at some location on one's body *B-badness*. And call the badness of one's unpleasant pain *U-badness*. To move from a perception of B-badness to a judgement of U-badness we need a mediating premise not found in the colour perception case. The mediating premise would have to be something like:

> **Mediating Premise***:* If I have an experience of something as bad for me, then the experience is itself bad for me.

We get the following sort of inference:

> **Inference to U-badness**
>
> **P1**: If I have an experience of something as bad for me, then the experience is itself bad for me.
>
> **P2**: My experience makes it seem as though my body is in a condition that is bad for me.
>
> **C**: My experience is itself bad for me[114]

The inference is valid. But it is far from obvious that it is sound. What justifies us in believing P1 (the mediating premise)? It's difficult to justify that premise without appealing to the fact that U-badness often *phenomenologically* accompanies B-badness, or that B-badness guarantees

[113] Note that it won't do to have bodily badness be *instrumentally* prudentially bad—i.e., because it is instrumental to one's ill-being—and affective badness be *non-instrumentally* bad. In fact, that distinction might make it even more baffling how we move from the perception of instrumental bodily badness to non-instrumental affective badness.

[114] If an inference like this is indeed needed to arrive at the judgement that one's unpleasant pain is bad, then the over intellectualization worry raised against the desire strategy in section 4.1 applies as well.

(conceptually) or makes more likely U-badness. And what could further justify that thought if not the fact that I am, or have been, directly aware of U-badness accompanying B-badness, and therefore, directly aware of *two distinct things*, namely, the phenomenal character of unpleasant pain and bodily badness? Any premise mediating the inference from my introspective judgement of seeming B-badness to my introspective judgement of U-badness needs to be justified on the grounds that U-badness itself is reason-providing (and surely, we need not have a piece of philosophical theory—something like our above inference—to make that connection). That is a highly demanding and unrealistic depiction of how subjects' access (and respond to) the badness of their unpleasantness. An awareness of the connection between the seeming badness-for-me that an experience presents me with and the badness of the experience itself would make it rational to judge, *on the basis of that awareness*, that one's unpleasant pain is bad. But the representationalist cannot account for this awareness, not without giving up on the transparency of experience. There is nothing in the perception of an extra-mental evaluative property nor in the *seeming badness* of something extra-mental (even one's own body) that warrants one in believing that one's *own experience* of the thing so perceived is itself bad, and hence must be eliminated. Representationalists cannot avail themselves to that justification since this would directly refute the transparency of experience, and so, representationalism about unpleasant pain.

At this point, representationalists might insist that the piece of knowledge we have that justifies the required judgement—P1 or the mediating premise—is a *conceptual truth*: that is, *being presented with badness* conceptually entails that *the presentation is itself bad*. As long as I know that I am being presented with something bad, I can know that the state of being presented with that badness is itself bad: it will apparently follow that one is having a bad experience if they are presented (in some manner) with badness.

Is this a conceptual truth? This is a difficult question to answer for I take it that any putative counter-example to the idea that the presentation of badness entails that the presentation itself is bad can be countered further by simply insisting that *either* the initial presentation *wasn't really* a proper experience of badness to begin with or that the putative badness of the presentation is really there, it's just been defeated. For example, say I am looking out my living room window, enjoying a cup of coffee, when I notice someone breaking into my car. Here it seems as though I am presented with the badness of an event—it is bad-for-me that someone is

breaking into my car—but surely, the presentation of that badness is itself *not* a bad thing. In fact, it seems like a very *good* thing that I am presented with this badness. One might even say to me: "Good thing you saw that!" where what is "good" here is that we were *visually presented with* something bad. But of course one could push back and simply deny that visual experiences can represent evaluative properties. Or perhaps some will insist that it is both a good and bad thing for you: it's good for you *overall* but still pro tanto bad for you (the experience that is). Or perhaps what is really good-for-me in this situation is *not* the visual experience itself but something more extra-mental, namely, the car-break-in itself.[115] Again, difficult to decide.[116]

A lot will hang on what the representationalist means by "presentation of badness". For I think if that is understood broadly enough as to include the car-break-in case above, then the representationalist is highly vulnerable to counter-examples. Of course, they can make the notion of "presentation" more precise which I think they should.

The representationalist might suggest that it isn't merely *any* presentation of badness that is itself bad, but rather the *interoceptive* presentation that is itself bad (I come back to this point later on). But this is a highly dubious claim, and even if it were true, it's one that most subjects definitely wouldn't have access to. For in order to accommodate the motivation condition, representationalists must now attribute to *all subjects* who rationally respond to their unpleasantness a belief that the presentation of badness entails that the presentation itself is bad. That, to me, seems like a very unreasonable thing to think, but it would seem that it's what the representationalist must say. So, construing our judgement *that* our unpleasant pain states are bad as dependent on a conceptual truth seems unpromising.

---

[115] This last suggestion seems doubtful. For if there really is something good about this case and something bad about it, I think it isn't too plausible to place *both* those evaluative facts in the extra-mental state-of-affairs, namely, the car-break-in itself and perhaps even more implausible to suggest that the goodness-for-me is the car-break-in and the badness-for-me is the visual experience. Rather if we are going to countenance one thing as good-for-me and one thing as bad-for-me, it is more natural and intuitive to have the car-break-in be objectively bad-for-me and the visual experience of that objective badness-for-me be what is good-for-me.

[116] We might, perhaps, employ what is called a "pity" or "sympathy" test (Mitchell, 2019b) whereby, in order to test for whether some mental state is in fact a bad state to be in is whether we, from our third-person perspective, feel pity or sympathy for the person in that state in virtue of them being in that state. One might feel sympathy for my having had my car broken into, but one would not, so it seems, feel sympathy for me in virtue of it striking me (perceptually) as bad that my car was broken into. One would feel bad for me whether they knew that I knew about the car break-in. Compare the state of grieving where we *do* seem to pity the person in virtue of being in that state.

7 Where does this leave representationalism?

The representationalist is in a difficult spot with respect to the normativity and motivationality of phenomenal experience. For, again, as I have argued, if we want to invoke desire-based explanations of the normativity and motivationality of experience, e.g., of affective experiences, then the relevant desires must be *object-directed*. I further argued that they must be object-directed in the sense that for the subject to form the relevant desire they need to be *directly aware* of the object of the desire. In other words, the desires need to be direct-awareness-involving object-directed desires. So, they must abandon desire-based explanations. And we further saw that abandoning desire-based explanations in the motivation case forced the representationalist to embrace a rather unintuitive and demanding picture of how we rationally respond to the badness of unpleasantness: via an inference with a questionable belief about the connection between being presented with badness and presentational badness. Again, my point earlier applies here as well: there is nothing that we could *normatively encounter* in the extra-mental world which would epistemically licence us to arrive at normative knowledge of our minds.

So, where does this leave representationalism vis-à-vis the normativity and motivationality of experience?

At this point, one might naturally wonder whether the representationalist can appeal to non-attitudinal-based explanations. After all, they do explain the phenomenology of unpleasantness in virtue of the *content* of an experience. What might the prospects here be for a content-based explanation of the normativity and motivationality of unpleasantness? Notice that I have already dealt with this to some extent in the case of our *motivation* to end our unpleasant experience. I claimed that one way to explain the motivation condition is to appeal to a second-order experience which itself takes as its content the evaluative property of unpleasantness. And to re-iterate, that is implausible for the representationalist since appealing to a second-order experience would entail that one could directly attend to the first-order phenomenal experience, hence a counterexample to the transparency of experience.

There is one content-based option worth briefly mentioning which is to appeal to the content of the unpleasant experience itself to explain our experience-directed motivation. The idea would be that the content of the unpleasant experience *intrinsically motivates* one to eliminate the

experience itself. The only content that could plausibly do that would be the evaluative content constitutive of the unpleasantness itself. But the evaluative content of the first-order perceptual experience is apparently supposed to explain *not* our experience-directed motivation *but rather* our world-directed motivation, e.g., the motivation we exhibit when we remove our hand from scalding water. How, then, could evaluative content *intrinsically* motivate *both* world-directed and experience-directed behaviour? It arguably cannot.[117]

So, although content-based explanations for meeting the motivation condition have not been fully explored under representationalism, I take it they are quite unpromising for the reasons outlined above. What about the normative condition? We have seen that an attitude-based account will not do. What about a content-based account? Is it possible to meet the normative condition—i.e., explain the badness and reason-giving force of unpleasantness—by invoking representational content *only*? I think that is equally unpromising. Let me say a few things about why, focusing on a recent attempt by Bain (2017) to explain how a representation with evaluative content might make the representation itself intrinsically bad/reason-providing. This will be important with respect to the motivation condition as well, for *if* the representationalist can plausibly accommodate the normative condition—i.e., they can adequately explain what is so bad about having an experience with a particular sort of *content*—then that explanation will lend itself to helping them meet the motivation condition. How so? For if there is some conceptual truth relating the badness of unpleasant experiences with the sort of content those experiences contain, hence capturing the normative condition, then that will presumably be a *conceptual truth* the representationalist can then appeal to to justify the kind of inference

[117] See, e.g., Aydede and Fulkerson (2018) for an argument about the inadequacy of evaluativism about affect to explain the *inherent* motivationality of unpleasantness. And insofar as evaluativism is the most plausible representationalist treatment of affect, it applies equally to representationalism. Bain (2017) is responding to this charge when he invokes desires (as we saw above) to explain our motivation to end the unpleasantness of our pain experiences, in effect denying Aydede and Fulkerson's *inherent* motivationality condition (while still, to an extent, accommodating it). Note, also, a perhaps odd requirement on the sort of content needed here. If we are to explain the *inherent* motivationality of our experience-directed behaviour—e.g., to take a painkiller—then the content the representationalist invokes will need to not only motivate *two* distinct behaviours at the same time—one world-directed and one experience-directed—but it will also have to be a kind of content which inherently motivates its own elimination. It's difficult to imagine what that could be. But some think it's plausible, in particular imperativists about pain content. See Barlassina and Hayward (2019) for a recent discussion and defense of this view. But see Bain (2011) for arguments against the idea that imperative content can play the proper motivational role vis-à-vis pain's world-directed motivation. Bain doesn't have in mind the view of Barlassina and Hayward, rather his targets are Klein (2007) and Hall (2008). It's worth reflecting whether Bain's (2011) arguments apply to Barlassina and Hayward's version of reflexive imperativism—the idea that imperative content intrinsically motivates its own elimination because it commands to eliminate itself. On the surface, the same issues about the explanatory impotency of imperative content seem to apply to them as well, that is, it applies to *experience*-directed motivation as well as world-directed motivation.

subjects make when judging that their unpleasant pains are bad. I will say more about this as we move on.

Recently, David Bain (2017) has tried to accommodate the normative condition of unpleasant pain by appealing to what he calls *perceptual encounter*.[118] This strategy, importantly, denies that experience-directed desires explain the normativity of unpleasantness. Rather Bain's strategy accommodates the normative constraint by invoking a particular kind of content, namely, *evaluative content* that is intrinsic to the feeling of unpleasantness. This much is familiar. And since we invoke evaluative content to explain the phenomenological *feel* of unpleasantness, and it's just this *feeling that is bad*, we can invoke the same content to explain that badness. Explaining the feeling *is,* in part, explaining why it's bad, therefore, accommodating the normative constraint.

Importantly, it's not *just* evaluative content that does the explaining, but the manner in which that content is represented: namely, *perceptually*. Why should this make a difference? To appreciate this point, consider how we might explain the difference between the phenomenology of judging that the tomato is red *versus* perceiving that the tomato is red. Presumably, the phenomenological difference between the two is explained by appealing to the *perceptuality* of some colour content—what makes it the case that two states with the same content have different phenomenal feels is due in part to the fact that one is a *perception* rather than a *judgement*. We might ask 'why is the *normative belief* about one's damaged bodily condition with content [bad-for-the-subject] *not* itself bad for the subject, when a perceptual experience about one's damaged bodily condition with content [bad-for-the-subject] *is* itself bad for the subject? My regular evaluative beliefs, which carry evaluative content are themselves not bad in virtue of the content they have, so why think that my evaluative experiences are any different? Focusing on the *perceptuality* of painful experiences should help draw out a more general question lingering in the background. By focusing on the phenomenological *feel* of unpleasant pains and identifying *it* (in the sense of predication) with badness, we should see that the question about the difference

[118] Note that Bain himself is not an outspoken strong representationalist nor for that matter clearly a strong representationalist at all. But I focus on his account because it is the only explanation on offer of the normative condition which does not appeal to desires or attitudes directed at the unpleasantness itself. As we will see, the plausibility of his account seems to entail a departure from strong representationalism. Note also that Bain here is responding to what is called 'the messenger shooting objection' (Brady 2015; Aydede and Fulkerson 2018; Jacobson 2013, 2018) where it is pressed upon the evaluativist qua representationalist: what is so bad about representing something as bad?

in the *normativity* of *kinds* of representational state (one being neutral and the other not), is akin to a difference in the *phenomenality* of different *kinds* of representational states (as in the phenomenal difference between believing that there is a red tomato in front of you and perceiving that there is a red tomato in front of you). And the question now becomes: what explains the phenomenal difference between a *judgement* that one is seeing a red and round tomato versus a *perception* of seeing a red and round tomato? The former presumably doesn't have a distinct *feel* compared to the latter. But notice now that the worry about how an experience can be bad rather than a judgement (an apparent problem for evaluativism) is now a worry about how perceptual states give rise to *a phenomenal feel that is bad* whereas judgements don't. The question, according to Bain, is no longer specific to evaluativism per se but rather to representationalism overall (Ibid., p. 484). The question now, I take it, is how perceptual representations give rise to a distinct phenomenal *feel*.

Put another way, the question of how a representational state is itself bad becomes the question of how a representational state has the phenomenal content it has. And notice further, that we have no problem in appealing to the perceptuality of colour experiences to explain the difference in phenomenality between judgements of colour versus perceptions of colour. So, then, we shouldn't have any issue in appealing to the same thing when trying to explain the particular *badness* of unpleasant pain, namely, in appealing to its *perceptuality*.

The phenomenal character of unpleasant pain—it's unpleasantness—is explained by appeal to representational content that represents bodily damage as bad-for-the-subject. And just like how we appeal to the *perceptuality of colour experiences* to explain the phenomenology of those experiences, we appeal to the same thing in *not only* explaining the phenomenology of unpleasantness, but the normative properties of that unpleasantness. That is, the badness of unpleasantness is explained by dint of the very same thing that explains it's phenomenology; there is no additional ingredient needed to explain why unpleasantness is bad over and above those features that make unpleasantness what it is.

But why should representing bodily badness *make* the experience so bad? The normative condition is apparently met by appealing to the specific way in which subject's *perceptually encounter the normative world*, or as Bain himself says, to the way in which one is "putatively encountering those [bodily damaged] states' badness-for-you, or your having their badness-for-

you putatively *impressed* on you" (484-485; his italics). Similarly to how one perceptually encounters the *redness* of the tomato thereby constituting the intrinsic feeling of reddishness of one's experience, one perceptually encounters the *badness* of one's bodily damage thereby constituting not only the intrinsic *feeling* of badness but the intrinsic badness itself.

Note something related to an earlier discussion. It might strike someone as at least an interesting disanalogy, and perhaps a rather troubling one, that when we perceptually encounter colour properties like redness, our experience does not *itself* become some kind of red; there is no *third* property over and above the red-like phenomenal property of my experience and the putative redness which we encounter. Of course, there is phenomenal redness, but importantly, or so it would seem to me, that is not a *kind* of red either; rather it is a phenomenal property: encountering the redness of the tomato is precisely what makes my experience feel *red-like*. And this is how it should be if we are representationalists. The transparency of experience fits nicely with this picture of encounter since to try to attend to the phenomenal property of one's experience—the red-like quality—one ends up attending *only* to (perhaps *encountering*) the *redness* of the tomato.

But notice that, according to Bain, when we perceptually encounter bodily badness not only is our experience determined to be a certain way, phenomenologically speaking—it's unpleasant *because* of our encounter with bodily badness—but it's also determined to have this *extra normative* feature: we are supposed to locate a second normative property, this time *in* our experience.[119] But notice a perhaps odd and troubling disanalogy between our perceptual encounters with colour properties and our perceptual encounters with normative properties. We've added another property over and above the phenomenal property of our experience and the putative extra-mental property we encounter.[120] Why, I ask, is there this other, distinct

[119] We might naturally ask: how do we locate it? To suggest that it is something we are aware of *directly* would seem to be a direct violation of strong transparency. Representationalists, of course, will point out that, although there is a distinct kind of badness determined by one's perceptual encounter with bodily badness, it *nonetheless isn't* something we are aware of directly. Rather, the story they tell here is that we are aware of what we perceptually encounter, and it is via such awareness that we come to know that our experience has this other experiential, normative property, albeit *indirectly*.

[120] See Jacobson (2019b: 397) for a similar although slightly different point. She claims: "Hence, had pain consisted in a perceptual encounter with the badness of an extra-mental property, then either the badness of this extra-mental property would have been the only badness in its vicinity [akin to our encounter with perceptible redness], or if pain itself were also intrinsically bad, this fact would have remained inexplicable." I think this is a false dilemma. For, as we have seen, the representationalist need not commit themselves to the idea that the badness of unpleasantness is something we are directly aware of in our experiences. In fact, it's something they must deny. And they have a story to tell about how we might access that badness, however demanding and implausible it is to begin with. So, it's not true that the extra-mental badness would have to be the only badness in one's vicinity; it

property formed with respect to our perceptual encounters in the normative case but not in the non-normative case? I take it this is something which calls out for explanation without sounding ad hoc.

In any case, let's say you find talk of "encounter' and "being impressed upon" plausible ways to spell out what might make a perceptual representation of something bad, itself bad. I want to suggest that this notion isn't plausible within a representationalist framework of phenomenal consciousness. Note, importantly, that if talk of perceptual encounter is a plausible way to capture the idea that being presented with badness is itself bad, then this would provide the representationalist with some needed ammunition. For, now having explained the normative condition by appeal to some conceptual normative truth about encountering normative states, they at least can make sense of the inference subjects would need to make in order to arrive at judgements that their unpleasant pain was bad for them. Recall the problematic premise for representationalists:

> **Mediating Premise***:* If I have an experience of something as bad for me, then the experience is itself bad for me.

We can make this more precise:

> **Mediating Premise (2):** If I perceptually encounter the badness of my bodily damage, then the perceptual encounter is itself bad for me.

MP2 just is the bit of philosophical theory representationalists need to capture the normative condition. And if that's true, then it plausibly allows for subjects to infer that their perceptual encounters with their bodily badness are bad for them. Hence, it will make sense for subjects to inferentially judge that their experience—their unpleasant pain—is bad for them since MP2 is true. Now, of course, as I mentioned above, MP2 is quite a bit of heavy-duty theoretical machinery to employ merely to explain how subjects access the badness of their pains, and hence how they are motivated to take a painkiller. But even in light of that, problems arise for the representationalist. Let me briefly explain.

---

would just have to be true that it is the only badness one is *directly aware* of or has direct access to, but the experiential badness is still there, it's just hiding. In other words, we need not (and should not) conflate a metaphysical point about what *makes* unpleasantness bad (perceptual encounter), and an epistemic point about how we access that badness.

Bain's strategy—call it the *perceptual strategy* (PS)—makes an important distinction between evaluative *experience* and evaluative *judgement.* An important question which, I think, is framing the issue is the following: how can a mere representation of something—e.g., bodily badness—itself be bad? Again, our answer is that an experience of badness is itself bad and not a judgement because the experience is an encounter with badness whereas the judgement apparently is not.[121] Again, we are trying to put meat on the bones of the idea that representing something can itself be bad and one way to do that is to appeal to the *manner of representation*.

But restricting those powers to *perceptual evaluative content* raises, I think, an important question: does that apply across *all* sensory perception? That is, would any sensory perceptual experience with the evaluative content *that bodily damage is bad* count as an encounter with bodily badness? I take it the answer is surely not; or at least there seem to be prima facie plausible cases where we are *perceptually* encountering our bodies as bad but nonetheless fail to have unpleasant phenomenology (e.g., is visually perceiving my broken finger really bad for me?) But then another question arises. What role is *perceptuality* itself playing? Take, for example, the difference between visually perceiving the badness of my damaged foot versus my interoceptively perceiving the badness of my damaged foot. I take it the latter is constitutive of unpleasant pain and the former is not. Why? Well one is a perceptual encounter with bodily badness—the interoceptive experience—and the other isn't—the visual experience. But both are *perceptual*, so the difference must come down to something else rather than mere perceptuality. Either the difference in phenomenology (hence normativity) comes down to a difference in perceptual modality or it doesn't. But, of course, if we are representationalists, perceptual modality cannot make a phenomenological difference for then we have a straightforward case of phenomenology separating from representational content.

But now notice where we are dialectically. The representationalist either admits that we can perceptually encounter bodily badness across sense modalities, or we cannot. If we can, then

[121] Note that if we restrict representationalism to perceptual experiences, one can still be a strong representationalist. It follows from this that one can think that all perceptual phenomenology is reducible to and identical with content. One might, then, think there is cognitive phenomenology with respect to our evaluative judgements but think that it involves different content. Or one might deny cognitive phenomenology all together. What I take it one *can't do* if one is a strong representationalist is say that judgements with one content and perceptions with the same content have differing phenomenology. This would be a straightforward case of phenomenology separating from content, and a counterexample to strong representationalism. So, if a representationalist is going to exploit the experiential/judgemental distinction to capture the normativity of unpleasant pain, then they must either deny cognitive phenomenology or claim there are different contents with respect to evaluative experiences of bodily badness and evaluative judgements of bodily badness.

there seem to be cases (e.g., visually perceiving one's damaged foot) where we have two *phenomenally distinct* experiences with the same representational content. That is precisely the denial of (strong) representationalism. If we cannot perceptually encounter bodily badness across perceptual modalities, then the notion of perceptual encounter no longer plays the explanatory role it initially did. For now representationalists must appeal to the role of a *specific* perceptual modality and the content it tracks. But, again, a specific perceptual modality can't play any explanatory role vis-à-vis phenomenological properties (including normative ones). So, we are left with *pure* content to explain the normativity of unpleasantness, a kind of content which itself is only perceivable within a specific perceptual modality (e.g., interoception).

But now we are back where we began. That is, how can the mere representation of some property itself be bad? Appealing to notions of "perceptual encounter" will not help the representationalist.

Conclusion

In its strong form, representationalism entails that subjects cannot be directly aware of their phenomenal experiences. This is due to representationalism's commitment to strong transparency, the idea that when we try to attend to features of our experience our attention 'slips through' to the extra-mental objects putatively represented by our experience. But this commitment has troubling implications. For a highly attractive way in which one might account for the badness of our mental states—e.g., the badness of unpleasantness—and our rational motivation to end those states—e.g., to take painkillers—appeals to desires for those phenomenal states *to stop*. But as I've argued, the idea that we can direct explanatorily relevant *de re* desires at phenomenal states entails that subjects must be able to directly attend to those phenomenal states. Hence, representationalism is incompatible with any view in value theory, metaethics, or moral psychology that will require *de re* desires to be directed at the unpleasantness or pleasantness of experiences themselves.

Eschewing *de re* desires, representationalists face serious problems. If I'm right, then they seem forced to embrace a *non-attitudinal account* of the badness of unpleasantness. But extant accounts to do just that—e.g., Cutter and Tye (2011)[122] and David Bain (2013) who appeal to

[122] Note here I am referring to their earlier paper regarding unpleasantness and its badness where they *do not* appeal to desires but rather attempted to explain unpleasantness and its badness by dint of evaluative content.

evaluative content rather than attitude—have faced a swath of criticism.[123]In fact, the criticism against the idea that evaluative content *alone* can explain the normativity of unpleasantness is *precisely* why some theorists, who themselves are representationalists, have turned to desire-based accounts, e.g., Cutter and Tye (2014). So, accommodating the normative condition of unpleasantness seems unpromising for representationalists.

Regarding the motivation condition, non-attitudinal accounts compatible with strong transparency have not been much explored. But as I briefly gestured to above, invoking a second-order experience to explain our painkiller-taking behaviour seems to violate representationalism's commitment to strong transparency, and hence is equally unpromising. I also suggested that an inferential picture of our access to the badness of unpleasantness is unpromising. For even granting that such evaluative judgements could motivate us, the judgements themselves are highly implausible, unlikely to be justified and are overly conceptually demanding. In light of all this, it might be time to abandon strong representationalism about affect.

Interlude

In the last two chapters, we saw two difficulties arise. First, it was questionable whether we could give a non-inferential account of our access to our motivating reasons. Motivating reasons, for some, just aren't the sorts of things we have distinctive self-knowledge of, however pre-theoretically plausible it might otherwise seem. So, it follows that if one restricts distinctive self-knowledge to what we can know non-inferentially, then our motivating reasons are no longer candidates for distinctive self-knowledge. Second, we switched gears, and saw the difficulties which arise from restricting how we access phenomenal states like unpleasant pain. One of the difficulties we saw in Chapter 2 was in some sense the flip side of the difficulties raised in Chapter 1: in Chapter 1 we saw that motivating reasons can't plausibly be known *purely non-inferentially* whereas in Chapter 2 we saw that normative reasons provided by mental states can't plausibly be known *purely inferentially*. I want to turn now to think about the theoretical possibilities that open up once we abandon a transparency restriction—and with it an inferential account—of our access to our own mental states and fully embrace the non-inferential nature of introspection. That is, I turn to investigate the theoretical upshot of fully appreciating the *direct non-inferential access* we enjoy to our own mental states. Once we

[123] See, e.g., Jacobson (2013), Brady (2015), Aydede and Fulkerson (2018)

appreciate this fact, I think an interesting, novel epistemological thesis arises with respect to moral knowledge: Introspective Intuitionism. This is what I now turn to.

CHAPTER 3: Introspective Intuitionism

Introduction

Experiences are often qualified to varying degrees as either pleasant or unpleasant, for instance, when experiencing the *pleasantness* of biting into a juicy, ripe strawberry, or experiencing the *unpleasantness* of smelling rotten yogurt. Emotional experiences, also, take on either a positively or negatively valenced dimension. The recalcitrant guilt one feels after a night out of drinking typically includes a negatively valenced quality attaching to, or perhaps constituting, the guilt itself. But notice another feature of our affective experiences, namely, their *evaluative* and *normative* dimension. Not only is tasting a juicy, ripe strawberry a pleasant experience, but it is also an experience that is non-instrumentally *good,* or *good-for-me*, at least pro tanto, in virtue of the fact that it is pleasant. Similarly, the unpleasantness of an intense migraine not only presents me with a distinct phenomenal quality—the unpleasantness—but it is also non-instrumentally *bad*, or *bad-for-me*, at least pro tanto, in virtue of the fact that it is unpleasant. These experiences are not only good or bad in themselves, but they seem to provide us with pro tanto *reasons* to either continue or eliminate their existence. Some of our experiences, then, seem to take on an evaluative and normative dimension.

Similarly, our propositional attitudes can take on an evaluative, normative, and perhaps even a *deontic* dimension. For instance, an intention to break a promise seems pro tanto *bad* or *wrong*. A desire indicative of a racial fetish is, arguably, also *bad* or *wrong*.[124] Both attitudes are, or so it seems, reason-providing in the sense that they give us pro tanto reasons to eliminate their existence. Some of our attitudes, then, seem to take on an evaluative, normative, and deontic dimension.

Furthermore, note that it is a truism that these mental states—affective experiences and propositional attitudes—are *introspectively* accessible. Turning my attention 'inward' I find the particular *unpleasantness* of the experience of smelling rotten yogurt, or the *intention* to break a promise. But what about the evaluative, normative, and deontic properties (hereafter just

[124] See Zheng (2016) for the claim that racial fetishism is morally objectionable.

'normative properties') of our mental states? In other words, can we introspect normative properties?[125] The goal of this paper is to investigate this hitherto underexplored question.

Notice what the question is not asking. The question is not asking, as has traditionally been the case, whether, for instance, experiences with an affective dimension are *experiences of value*, i.e., experiences which put us in touch with an evaluative world. Rather our question is whether the normative properties *of those experiences* are themselves accessible to us. To illustrate further, compare two kinds of value with regard to one and the same experience, namely, watching an autumn sunrise. The experience of watching an autumn sunrise not only presents us with something beautiful—e.g., the beauty of the sunlight reflecting the colour of the leaves—but is itself something that has value, arguably, rooted in the pleasure we receive from undergoing such an experience. My perception is not only a way in which I can access aesthetic values such as beauty, but is itself something that has value, perhaps consisting in a kind of aesthetic pleasure, but a pleasure that is nonetheless *good*. And just like we can have perceptual access to the value of an autumnal sunrise, surely, we can have introspective access to the value of such an experience, albeit a value importantly *different* than the one our perceptual experience gives us access to. Similarly, our pleasant and unpleasant (hereafter 'affective') experiences and propositional attitudes might be *about something worldly*, e.g., one's bodily damage or breaking promises, but they themselves are nevertheless taken to be good or bad, right or wrong, or reason-providing.

So, importantly, our focus here is not about how our mental states present us with a kind of *worldly* or *extra-mental* value such as the beauty of a sunset (or normative properties more generally). Rather our focus is about how our mental states present us with a kind of *mental* (dis)value, or *mental normativity*. We can, then, distinguish between the normativity attached to the mental state—call it *mental normativity*—and the normativity attached to that which is being experienced—call it *worldly normativity*.

The focus of this paper is on mental normativity, how we access it, and the upshot this has for moral epistemology. One reason why we should care about whether we can introspect normative

[125] There are questions about which properties qualify which mental states. For instance, it seems straightforward enough that our unpleasant experiences can be bad, even reason-providing, but can they be wrong? Similarly, we speak of 'bad' intentions but also of 'wrong' intentions. Are these merely synonyms, or is there an important difference? These are important questions which I leave to the side for now.

properties is its potential to provide us with a novel naturalistic moral epistemology. Let me explain. There has been renewed interest in the metaethical view Ethical Intuitionism which claims the following:

> Ethical Intuitionism (EI): normal ethical agents have at least some non-inferentially justified first-order normative beliefs. (Cowan, 2013a)

Put this way, EI is attractively broad to include the possibility of having states of the mind justify first-order ethical beliefs *without* positing a faculty of rational intuition, opening the door for an a posteriori intuitionism. For instance, one highly plausible naturalistic candidate for providing non-inferential justification is introspection. Prima facie, introspection can provide the justificatory grounds needed to support EI, something like the following:

> Introspective Intuitionism (II): normal ethical agents can and do have non-inferentially justified first-order normative beliefs by having introspective states[126].

Now, for (II) to be true, we need the following to be true:

> Normative Introspection (NISP)*:* at least some normative properties are introspectable.

Take, for instance, the badness of unpleasant pain. If, like our unpleasant pains, we can introspect the *badness* of our unpleasant pains, and introspection of that badness is non-inferential, then we have the resources necessary to account for how at least *one* first-order *evaluative* belief—that unpleasantness is *bad*[127]—could be non-inferentially justified. Given that other (normative) properties are similarly introspectable, we have then the resources to account for other first-order normative beliefs. Introspecting the *wrongness* of my intention or desire, I have a pro tanto justification for the belief that *my intention to such-and-such is wrong*, and so on for any introspectively available normative properties of my mental states. From this, and with a little bit of thinking, we can use our introspectively justified first-order normative

---

[126] I have put things this way as to not commit myself to whether the introspective states themselves need to have *ethical/normative content* to justify first-order ethical beliefs or whether introspective states *without* ethical/normative content can nevertheless justify first-order ethical beliefs. For a similar view of the latter claim in the case of ethical perception, see McGrath (2018). I also want to remain neutral on what view of introspection is best.

[127] I have been intentionally ambiguous here about whether the evaluative belief is about badness *simpliciter* or not. Later on, I consider a potential worry about the limited scope with which introspection can deliver *ethical* knowledge. In any case, whether it is evaluative prudential badness or evaluative badness simpliciter we can introspect, I take the general claim that we can introspect badness *of some kind* to be a substantial claim.

beliefs to arrive at new justified first-order normative beliefs beyond the contents of our minds. In other words, we take what is fundamentally a feature of our experience and use it as the bedrock for our judgements about the world beyond our experience.

Here's the plan. In section II, I set out the preliminaries. In section III, I motivate the claim that mental states can have normative properties. In section IV, I argue that we should take seriously the idea of normative introspection (NISP) on phenomenological grounds. Importantly, I do not think that to be able to engage in normative introspection our mental states *by themselves must* take on normative properties, although I do motivate NISP under the assumption that mental states *do* have normative properties. In section V, I outline how something like NISP could be true without thinking that mental states can have normative properties, e.g., be wrong. In particular, I borrow from the *affordance* literature and argue that insofar as it makes sense to think that mental states can afford actions, and hence, give us knowledge of these affordances, we can gain normative knowledge that is importantly directed *at action* via introspection without entailing that the mental states we introspect are themselves right or wrong. In section VI, I sketch how Introspective Intuitionism (II) could provide the requisite *epistemically independent* non-inferential justification for our first-order normative beliefs by considering some objections to extant accounts of Intuitionism as well as objections that are proprietary to (II). In section VII, I conclude.

2 Preliminaries

Above, I spoke loosely of 'ethical', 'deontic', 'evaluative', and 'normative' properties. Let me clarify things a little.

In what follows, by 'normative' I mean to refer to both evaluative properties and deontic properties.[128] Also, I will typically drop use of the term 'ethical' or 'moral', even though the intuitionism I am interested in has historically gone by the name 'ethical intuitionism' or 'moral intuitionism'. But it's not always made clear from within the intuitionism literature whether 'ethical' or 'moral' are supposed to encompass both the evaluative and the deontic domain rather than just one of them.[129] The same goes for 'ethical perception' or 'moral perception'. These

[128] See Tappolet (2013, 2014) and Werner (2019a).
[129] In fact, sometimes philosophers often interchangeably refer to the same thing by 'moral wrongness' and 'moral badness' without clearly delineating any boundary between the two.

terms seem to be most often associated with deontic properties[130], although, again, things are not always so clear.[131] Also, the 'moral' is often times used in contradistinction to the 'prudential' which itself is typically related to the evaluative, i.e., value concerning oneself or one's well-being. This might naturally incline one to think of the content of a moral reason as essentially *other-regarding* but that would be hasty, for some (e.g., Kant) think that we have duties towards ourselves, i.e., self-regarding duties. And even then, our self-regarding duties, say, to cultivate our natural skills, might conflict with our general well-being, i.e., our prudential reasons.[132] Theorists in moral epistemology as well as in the ethical perception literature are typically not sensitive to these distinctions.[133] So, in the light of all this, I've decided to speak as broadly as possible by using 'normative' to cover everything from the evaluative (prudential and otherwise) to the deontic. That is, for ease of exposition, I will be using the term 'normative' to cover everything from the evaluative—e.g., *goodness, badness, courageousness, cruelty, estimability, admirability, etc.*—to the deontic—e.g., *wrongness, permissibility, oughts, obligations, etc..*[134] I also speak from here on out of 'normative intuitionism' rather than 'ethical intuitionism'. Of course, where appropriate, I will distinguish between the various categories of normative properties, especially in motivating the claim that normative properties do indeed supervene our mental states; and sometimes, I will speak very simply of basic normative reasons for or against something without having anything particularly evaluative or deontic in mind[135]. The reader is, of course, welcome to have in mind the properties they find most plausible (if any) that qualify our mental lives, and which of those (if any) are most plausibly introspectable. In Sect. IV, I focus primarily on the putative wrongness of an intention to lie to help motivate normative introspection.

---

[130] For example, Harman's (1977) canonical example of 'seeing' wrongness in the torturing of a cat is frequently cited as the main example of moral perception. Cullison (2010) is strictly concerned with perceiving the 'wrongness' of an action (see his fn. 2).

[131] See Berqvist and Cowan (2018) for a very broad understanding of 'evaluative'.

[132] See Crisp (2018) for a good discussion regarding the distinction between moral reasons and prudential reasons.

[133] Although Werner (2019) is a start, and more recently Müller (2020) presents a careful discussion of the possibility to perceive concern-dependent practical reasons, i.e., reasons whose existence depend on the concerns or ends of the agent whose reasons they are.

[134] At the moment, I am not endorsing any claim about whether such properties just mentioned actually do qualify our mental states.

[135] Although, NORMATIVE REASON is typically taken to be a deontic concept. Things become unclear once philosophers begin attempting to analyze deontic properties in terms of reasons, or as it is sometimes called, once we "pass the deontic buck". See Stratton-Lake (2002) and Bedke (2011) for discussion of reducing deontic properties to reasons. See also Dancy (2000) for general discussion about "passing the buck".

Aside from ease of exposition, I also speak broadly of the 'normative' in order to leave room for the possibility of holding one kind of intuitionism and not another. In order for what I call 'introspective intuitionism' to seem plausible, there needs to be, among other things, *some* introspectable *normative* properties; which ones *exactly* is a question for another day. For instance, perhaps some will find it more convincing that *pro tanto* properties[136] are introspectable rather than *all-things-considered* properties[137]; and perhaps some will find it more convincing that evaluative properties rather than deontic properties qualify mental states, and hence, will find more plausible a version of intuitionism that justifies first-order pro tanto evaluative beliefs rather than pro tanto deontic ones (or vice versa). And, of course, it would be a very good thing to get to the bottom of which kinds of properties exactly are or are not introspectable or even perceivable. For there is a general question as to what extent normative metaphysics restricts normative epistemology. But those are questions for another day.

In any case, to drive the point home: the main purpose of this paper is to sketch a hitherto unexplored novel a posteriori intuitionism, one grounded in the introspection of our own mental states and the possible normative properties (whatever they happen to be) supervenient upon them, and not to nail down any of the finer details. I will, of course, along the way gesture towards various ways of filling in the details.

Finally, the plausibility of what I call 'normative introspection', where that refers only to deontic properties, shouldn't depend on whether one accepts a restricted view about the scope of morality. That is, that morality's scope (where 'morality' here is best understood as encompassing what is morally *permissible* and *impermissible*, and hence, as encompassing the deontic) does not extend to the purely private such as our thoughts, fantasies, feelings, imaginings, desires, etc., and it is only when such mental states are appropriately conjoined with outward behaviour that they take on any normative cum deontic dimension. On one picture of normative introspection, merely introspecting the non-normative mental base is enough to

[136] By 'pro tanto' I mean properties that are not *extinguishable* but are nonetheless *defeasible*. For instance, it might be pro tanto wrong of me to push you off the bridge, but nevertheless defeated by the stronger pro tanto (or perhaps all-things-considered) reasons to save thousands of lives which pushing you off the bridge would cause.
[137] See Werner (2019a) for the claim that in moral perception, it is the *pro tanto* evaluative and deontic properties that are most plausibly perceivable. I don't want to model what I have to say here about normative introspection on any account of the plausible candidates of moral perception. Although, if one is inclined to think of introspection along broadly perceptualist lines à la David Armstrong (1968) then I suppose a good deal of what is said about moral *perception* could carry over to normative *introspection*. But I think there are important differences anyways between perception and perceptual accounts of introspection which should make anyone skeptical of any sort of symmetry between the two.

provide the non-inferential justificatory grounds for first-order normative beliefs. But for now, I merely want to point out that those who are sympathetic to the claim that morality's scope does *not* extend to the purely private, need not think that introspection of normative properties is a non-starter. In other words, theorists inclined to restrict morality's scope to the extra-mental need not deny the possibility of normative introspection. I will return to this below.

But even though *that* idea is perhaps the minority view, let me anyways motivate its rival claim that our mental states *do* in fact have normative properties of various kinds.

3 Normative properties and mental states

In the opening paragraphs, I briefly motivated the idea that a certain sub-set of our mental states have what I broadly called 'normative properties', that is: evaluative and deontic properties. The further idea there was that *if* these mental states have normative properties, then prima facie those normative properties are themselves introspectable. I take it that the claim that we can introspect the normative properties of our mental states has an initial pre-theoretic plausibility; at the very least, it seems to me pre-theoretically plausible that our mental states can *have* normative properties. But aside from that initial plausibility, I also pointed out that there are good reasons to consider whether something like normative introspection is possible, especially if one is interested in giving a naturalistic account of ethical intuitionism (hereafter 'normative intuitionism'). The idea here is, roughly, given that the core claim of normative intuitionism (NI) centres around the *non-inferential justification* of first-order normative beliefs (see Vayrynen 2008; Cowan 2013a, 2013b), and introspection is an uncontroversial *natural* source of non-inferential justification, it follows that introspection is a highly attractive naturalistic candidate for the non-inferential justification of our first-order normative beliefs. The question, then, is: *can we introspect normative properties?*

One way to move forward is to break down our investigation into two parts. First, we can focus on the plausibility of the introspection of normative properties—what I will call 'normative introspection' (NISP). We can further divide this part of our investigation into two parts: A) whether normative properties supervene mental states (and which ones), and B) whether those normative properties are themselves introspectable. And second, we can focus on the theoretical advantages of normative introspection and its plausible support for what I above called 'introspective intuitionism' (II), the view, again, which claims that: normal ethical agents can

and do have non-inferentially justified first-order normative beliefs by having introspective states.

Let us then proceed to think first about whether normative properties supervene our mental states.

3.1 Mental normativity

Before moving on to consider our main question, it will be useful to map out which normative properties qualify our mental lives. Above, I noted that there seem to be two classes of mental states that can take on normative properties: affective states and propositional attitudes. I also noted two different classes of normative properties: evaluative properties and deontic properties. Let us first consider evaluative properties and the possible mental states they qualify.

3.1.1 Evaluative properties

The most plausible normative properties that predicate our mental states seem to be *evaluative properties* such as goodness and badness[138]. Take, for instance, the feeling of unpleasantness that constitutes a throbbing migraine. Unpleasant pain is a paradigmatic case of something that is *bad*. And pleasure, which is often assumed to be the opposite of pain, is a paradigmatic case of something that is *good*. Again, take, for instance, the pleasure one might feel upon watching an autumn sunrise. The pleasure is often synonymously described as *good*. Emotions like guilt also take on an evaluative dimension. The recalcitrant guilt one feels after a night out of heavy drinking seems to be a *bad* state to be in. So, I take the following examples to be highly plausible (perhaps even paradigmatic) cases of mental states which take on an evaluative dimension:

> *Recalcitrant guilt***:** the *badness* of the feeling of recalcitrant guilt after a night out of drinking.
>
> *Visual pleasure:* the *good* feeling of pleasure received from watching an autumn sunrise.
>
> *Unpleasant pain:* the *badness* intrinsic to the unpleasant pain felt when undergoing an intense migraine.

[138] The history of philosophy is replete with examples. The most notable being Jeremy Bentham (1789 [1970]) and John Stuart Mill (1979 [1861])). See, e.g., also Armstrong (1962), Aydede and Fulkerson (2018), Bain (2013, 2017), Bradford (2020), Brandt (1979), Cutter and Tye (2011, 2014), Fletcher (2018), Heathwood (2007), Helm (2002), Parfit (1984), Pitcher (1970a) for discussions centring around the badness of pain.

The above examples seem to be paradigmatic instances of states of affairs instantiating evaluative properties. A possible issue here, however, is that the above examples only show that affective states instantiate *prudentially* good and bad properties: the badness of unpleasantness is only something that is *bad-for-me* and not bad *simpliciter*. I have my reservations about whether the unpleasantness I experience is only bad-for-me rather than also being bad simpliciter.[139] For it is highly plausible that if anything is bad simpliciter it is unpleasantness. And the fact that it is *me* who experiences unpleasantness shouldn't detract from the fact that it is badness simpliciter I am aware of when I undergo it. But in any case, let me turn to some other examples to help show that mental states can take on (moral) goodness and badness.

Evaluative properties also qualify our *motivational states*. Consider the following cases:

> *Motivated Mal:* Mal is extremely motivated to support her partner, Pam. Mal tries to make sure Pam has the proper resources to succeed in her career and her personal life. Mal also tries to support Pam with words of encouragement. But unfortunately, for Mal (and for Pam), every time Mal tries to be supportive, she never seems to successfully provide the goods necessary for Pam's success. In fact, she ends up making things *worse.* She never seems to do and say the right things. Mal always fails, even though she genuinely *wants* what is best for Pam[140].
>
> *Indifferent Ian:* Ian is the opposite of Mal. He does not support his partner, Patrick. In fact, Ian is positively motivated to undermine Patrick's success in his career and personal life. But, every time Ian tries to undermine Patrick's well-being, he ends up *supporting* Patrick's success in his career and personal life: he ends up making things *better*. He always seems to do and say the right things. Ian always succeeds, even though he genuinely wants to *undermine* Patrick's success.

Mal's motivation to support her partner, and Ian's motivation to undermine his are, respectively, morally good and morally bad motivations to have. Arguably, they are also the *right* and the *wrong* motivations to have, at least in the context of their relationships. I will return to the issue of deontic properties later, but for now note that, intuitively, we might reasonably hold certain evaluative attitudes like *esteem* toward Mal's deep motivation to support Pam. Conversely, we

[139] In any case, it would still seem to me that establishing that we can introspect prudential badness would be a substantive claim.

[140] This seems to be as good a case as any, where the following Nina Simone quote applies: "I'm just a soul whose intentions are good. Oh Lord! Please don't let me be misunderstood."

might be inclined to reasonably hold a negative, perhaps *disparaging* moral evaluation toward Ian's indifference, and positive motivation to undermine Patrick's well-being and flourishing. Note, further, that we seem to hold these attitudes towards Mal's and Ian's motivations *independent* of the actions they happen to produce.[141] Our evaluations are restricted to the *motivations themselves*. As Phillipa Foot (2004) claims: "dispositions, motives, and other 'internal' elements are the primary subjects and determinants of moral goodness and badness."

To see this more clearly, think about what happens when we switch the behaviours that Mal's and Ian's motivations produce; that is, when we hold fixed the motivations and swap the behaviours they produce such that they *match*. Intuitively, our evaluative attitudes don't change, at least not in the sense that one is generally *positive*—the attitude directed at Mal—and one is generally *negative*—the attitude directed at Ian. If, in the matching case, our attitudes *do* change when we align motivations with behaviour by swapping only the behaviour produced, it is most plausibly a change in *degree* of evaluation. Our evaluation of Mal might be of a higher degree now that her motivation matches the behaviour produced, and vice versa for Ian.[142]

But notice what happens when we swap Mal's and Ian's *motivations* rather than their behaviour. Intuitively, our evaluative attitudes *swap* as well. That is, our attitudes change in the *kind* of evaluation we give rather than merely in the *degree.* Mal's *new* motivation to *undermine* her partner's success *matches* the behaviours which we typically associate with such a motivation; Ian's *new* motivation to *promote* his partner's success *matches* the behaviours which we typically associate with such a motivation. Our evaluative attitudes about Mal and Ian will surely switch as well: we develop a *negative evaluation toward Mal* and a *positive evaluation toward Ian*. Why? One powerful explanation is that our evaluative attitudes track motivating states. And although our attitudes in this case might be directed at the *actions* rather than merely the *motivating states* themselves (as in the original example), this fact can be explained by the fact that the actions *match* (and implicate) particular motivating states which direct our evaluations to the whole motivation-action composite. Importantly, once we divorce the motivation-action

[141] But even if we didn't, that is, even if some version of act consequentialism is correct and our mental states are good/bad, right/wrong only insofar as they tend to produce overall good/bad actions, this doesn't undermine the claim that our mental states still have the relevant normative properties in question. In such a case, there will be a question of how we could come to know the normative properties of such mental states only by 'seeing' the mental states themselves that stand in some wider relation to a set of good/bad consequences. Below, I address this worry.
[142] Perhaps the actions now become praiseworthy and blameworthy, respectively. It might even be the case that Mal's motivation on its own is worthy of praise since she seems to be subject to some sort of bad moral luck.

composite such that there is a *mismatch* between the motivation and the resultant behaviour, our evaluative attitudes only change in kind when we change the motivational state itself rather than merely the behaviour. Holding fixed behaviour and changing the motivational state will alter the kind of evaluation we give. But holding fixed the motivational state and changing the behaviour won't. So, our evaluative attitudes track motivating states. Hence, our motivational states can be good or bad.

One might object that what is really going on in such cases is that our intuitions are tracking what our subjects are *trying* to do. Mal is trying to help her partner; Ian trying to undermine his. And it is the trying which is commendable and condemnable. Furthermore, trying is not merely a motivational state. There is something above merely having a motivation that amounts to my trying. So, my argument hasn't quite shown that motivational states can be the bearers of evaluative properties.

Even so, I take it that what *makes* trying in Motivated Mal good and what makes trying in Indifferent Ian bad as well as what makes the various matching cases good and bad is the putative *endorsement* of the motivation in question. We can run the same line of reasoning as we ran above. Holding fixed the overt behaviour (the behaviour added to the endorsement that partly constitutes the trying) and changing the endorsement will result in a change in kind of evaluation. That is, Mal endorses the motivation *to support her partner*; Ian endorses the motivation *to undermine his*. If we swap these endorsements, surely our evaluative attitudes swap as well. But if we hold fixed the original endorsements and subtract the bit of overt behaviour that partly constitutes the trying (either by swapping them or just merely subtracting them from the equation) our evaluative attitudes of Mal's and Ian's situations do not change, at least not in kind. Why? Our evaluations track whether one is or is not endorsing some putatively good or bad motivational state *independent* of whether they actually *try* to fulfill that motivation.[143]

[143] Note that one could develop this thought further by suggesting that some mental states, e.g., motivations to murder someone, are morally bad states *independent* of whether one endorses the motivation.

3.1.2 Deontic properties

Above, I motivated the idea that a variety of our mental states can be good and bad: affective states and motivational states All these states, I claim, can be evaluated as good or bad, either prudentially or morally, and perhaps even take on a *thicker* evaluative dimension by being estimable or disparaging. And, assuming a stance-independent evaluative realism, evaluative properties, then, can supervene our mental states. But when it comes to deontic properties like moral *rightness* and *wrongness*, *impermissibility*, *requirements, obligations,* and so forth, things are more complicated. For instance, in the case of some of our affective states like pain, the locution 'My pain is wrong' or 'My pleasure is right' is infelicitous. But in the case of our desires, motivations, fantasies, and more generally, thoughts, things are not so straightforward. There is no infelicity in claiming that my motivation, e.g., to lie, is morally wrong; or that one ought not, *morally speaking,* fantasize about raping young children. And even though some (e.g., Sher, 2019) have gone to great lengths to argue against the thought that purely mental states can be morally wrong or impermissible[144], the status quo is that they can be[145].

Note that it won't matter for our purposes *why* we think we have moral reasons for or against certain kinds of mental states. That is, it won't matter whether mental states themselves are the *primary* subjects of moral evaluation alluded to in the Foot quote above, or whether mental states are subject to moral evaluation only in relation to something else, e.g., to the consequences of the actions they tend to produce. Perhaps rape fantasies increase the likelihood that people who entertain them will act in such a way and produce bad consequences; perhaps there is something inconsistent about willing such fantasies into universal law; or perhaps such fantasies are indicative of a vicious character. Whatever the case may be, the point stands that those are morally wrong mental states to occupy. Of course, *if* mental states such as affective responses, dispositions, motivations, beliefs, desires, and intentions *are* the primary subjects of moral evaluation—i.e., independent of what they tend to produce or effect— then all the more plausible it is that they are directly qualified by normative properties. But, again, even if they

[144] Note that there seems to be at least *one* moral property that would apply to mental states on Sher's view: the moral property of *permissibility.* If purely private mental states are never morally impermissible, then surely, they are morally permissible. See Director (2022) and Coates (2023) for arguments against Sher. Note, also, as I mentioned in the preliminaries of the paper, that my positive arguments for normative introspection need not rely on the claim that morality's scope extends to the purely private realm of the mental, although I take it that if it does extend that this bolsters the plausibility of normative introspection.

[145] See, e.g., Smith, (2011: 243), Brewer, (2000: 38), Hazlett (2009: 245), Cox and Levine (2004: 225) D'Arms and Jacobson (2000: 80, 82), Schroeder (2018: 115-116), Basu (2019b: 2500; 2018; 2019a), Zheng (2016: 407) The first five references are from Sher (2019).

aren't the primary *subjects* of normative evaluation (see, e.g., Driver, 2001), our mental states are still subject to normative evaluation: they will be right or wrong, good or bad, insofar as they, e.g., tend to produce certain consequences. In either case, we still morally assess our mental states. And at this point in the paper, this is all I need to establish: mental states can have the properties of *being morally right, being morally wrong*, *being morally impermissible, being required to give up, etc*. How we choose to cash out those properties shouldn't concern us.

But let me say more in favour of the thought that mental states can have normative properties, in particular, *deontic* properties. There are two considerations I want to consider in favour of this claim. First, for some normative theories such as Kantianism mental states seem to be the primary subjects of moral assessment. And second, some of our experience-directed behaviour such as therapy is best explained by appeal to the fact that mental states can provide us with *reasons* to eliminate their existence.

First, according to Kant, our intentions or "subjective principles of action" are the primary subjects of moral evaluation. Kant (1997) is perhaps the most obvious case in the history of philosophy of someone who explicitly claims that our internal mental states are subject to moral evaluation. Importantly, for Kant, our actions will be right and wrong only insofar as they are constructed from motives that pass the categorical imperative (CI) test; what we test, in the first instance, are our motives for action; and we may act from a motive if *it* passes the CI test; but importantly, our actions are permissible or impermissible insofar as they arise from permissible or impermissible motives, motives which either do or not indicate a Good Will. Motives are not made right or wrong in relation to the actions they tend to be conjoined with or, for that matter, with *any* action they are conjoined with but rather with whether they spring from a Good Will.

Second, I think there are cases where one is motivated to act in ways to eliminate their own mental states which help show that some mental states can be at least *reason-providing* and perhaps even morally right and wrong. To begin, let me start off with the uncontroversial claim that many of our instances of unpleasantness are followed by trips to the medicine cabinet to try to eliminate the unpleasantness as much as possible; in other words, we are motivated to take painkillers. Now, one plausible story about our painkiller-taking action is that we respond to a reason by taking a painkiller, in particular, we respond to the putative *badness* in the unpleasant state itself. What we want to eliminate is the particular *bad feeling* impressed upon our own

consciousness. Importantly, what this picture takes for granted here is that the mental state itself—the unpleasantness—is a bad state to be in; and that when we take painkillers, we are responding appropriately to the features that give us a reason: to the badness. What I want to suggest is that many of our other mental states have a similar structure to our painkiller-taking actions. I think cases of people who seek out therapy help show this.

Take, for instance, Mark, who forms, at first, a relatively *weak* desire to cheat on his partner;[146] Mark doesn't endorse the desire nor does he pay much attention to it; but as time goes on, the desire slowly strengthens and his attention to it increases; when, in moments of weakness, he finds himself attending to the desire and imagining himself fulfilling it, he subsequently feels a small amount of pleasure from it. After a while, our subject decides to see a therapist to help him eliminate the desire. After a few sessions, he and the therapist end up successfully eliminating the desire. Now, I take it that this isn't an obscure or even uncommon case. It seems that what is happening in the therapy case is the same thing that is happening in the painkiller case: there is a response to a (good) reason(s) for action that, in the case of the unfaithful desire, is putatively partly grounded in the normative properties of the desire; Mark has at least a pro tanto normative reason to *eliminate* his desire, and he acts *well* when he seeks therapy and eliminates it. Furthermore, if we think Mark's action is *morally praiseworthy*, then one attractive explanation of that fact is that Mark acted for the *right reasons*; and what that putatively entails is that Mark responded to the *reason-making features*: his mental state.

So, I take it to be a reasonably plausible claim that mental states have normative properties.

4 Normative Introspection

Recall in the opening paragraphs the view of normative introspection I put forth:

> Normative Introspection (NISP)*:* at least some normative properties are introspectable.

There seems to be strong intuitive pull to the idea that evaluative properties like the *badness* of a painful pain or the badness of an unpleasant emotion like grief are introspectable. For instance, when we undergo particular unpleasant experiences—like migraines—we typically are motivated to take painkillers. This experience-directed action seems intuitively to be the result

[146] Or take your favourite unsavoury desire, thought, fantasy, etc.

of some sort of *introspective* knowledge. That's why we want to eliminate it. We experience something that is putatively *bad-for-us* and we respond to the reasons provided by the badness. Importantly, that rational response to the badness of our unpleasant pain is underpinned by our *introspective access* to the badness: given that our unpleasant pains provide us with reasons to eliminate their existence, our painkiller-taking actions seem to be direct *introspective* responses to those *normative properties*[147]. So, I think it is quite reasonable to suggest that in the case of the evaluative and normative properties of our affective states—i.e., the particular *badness* of unpleasant pains or painful emotions—we can and *do* introspect such properties. And in cases such as affect, the preceding kind of normative introspection seems highly plausible. In short, it seems highly plausible that introspection provides the non-inferential justification for our first-order evaluative belief that pain is bad.

Let us reflect for a moment on other alternatives. Epistemic concepts like intuition and emotion are simply misplaced to account for the justificatory source of our evaluative beliefs, at least in the pain case. Typically, intuition provides justification through subjects applying their grasp of general truths to particular cases. But that seems unnecessary in the case of pain. Emotion is also misplaced since it would be odd to posit *another* affective state—e.g., an emotional experience *of* one's first-order unpleasant pain—in order to explain the source of justification for first-order evaluative beliefs about unpleasant pain. Notice that when we experience unpleasant pain, we already are introspectively aware of our unpleasantness. So, it would be superfluous to posit these other epistemic states to explain such evaluative justification.

But one might reasonably raise eyebrows at the prospects of introspection operating with respect to the *deontic properties* of our mental states, e.g. the putative *wrongness* in an intention to lie. For there seems to be an important disanalogy between the evaluative properties of our affective states and the deontic properties of our propositional attitudes. Phenomenologically speaking, there are important differences. Our cognition with respect to our affective states is rooted in an awareness of a *phenomenological* datum making other epistemic options potentially otiose. But in the case of the wrongness of an intention to lie, introspection doesn't lend itself so obviously to explanation for we seem to lack any phenomenological analogue in such a case. So, this at

[147] Things become slightly more complicated once we consider representationalist accounts of pain, but I take it that those who endorse such accounts of pain would still want to construe our painkiller-taking actions as primarily an introspectively based action. Of course, we saw reason to reject such an account in Chapter two precisely because of its difficulties in accommodating such introspectively-directed action.

least invites some suspicion with respect to the introspection of normative properties in other, non-evaluative and non-phenomenological cases.

Now, I can't go through all the potential properties to see whether they plausibly are or are not introspectable. Rather what I want to do in the remainder of this section is motivate the idea of normative introspection with regard to one normative property, namely, wrongness. I take what I say about the introspectability of wrongness to help support the introspectability of other normative properties since it seems to be the most potentially problematic. I think a case can be made for the plausibility of normative introspection with respect to something like the wrongness of an intention to lie, a case which itself involves a distinct phenomenological state.

Here is a fairly close instance a philosopher has gotten to thinking about the introspection of normative (deontic) properties:

> The core idea behind some of the main deontological theories is that the intentions of the agent are what matter for determining the rightness or wrongness of an action. One example of a deontological theory holds that for an action to be morally right, it must flow from a good will or proper intentions—where proper intentions are thoughts of duty. The bottom line is that you must perform the action for the right reasons in order for it to be morally right. Given this kind of theory, right action will involve mental states that we could probably have perceptual knowledge of. (Cullison, 2010: 165).

Since Cullison is strictly concerned with *moral perception,* he naturally thinks of the rightful/wrongful-constituting mental states as mental states of *other people*. But wouldn't a seemingly simpler and more intuitive claim to make be that we *introspect* the normatively relevant mental states?[148] That is, right actions will involve *our own* mental states that we *definitely* have introspective knowledge of.

Here is an example to help motivate normative introspection:

> **Susan:** Susan is a typical daughter and occasionally lies to her mother. When pressed about whether she's cleaned her room, Susan feels the motivational pull to lie, and in a rather non-deliberative, unreflective manner, forms the intention to lie, and lies to her

[148] If one is sympathetic to normative perception in the case of the deontic properties of other's mental states, then one should be at least, and perhaps slightly more, sympathetic to normative introspection in such cases since introspecting our own mental states is much less controversial than perceiving the mental states of others.

> mother, claiming she did. But as Susan continues her education, she takes an ethics class. The next time Susan is pressed about whether she's cleaned her room (and she hasn't) she feels that same motivational pull to lie, and in a rather non-deliberative, unreflective manner, forms the same intention to lie, yet this time she is *struck* by the wrongness of her intention; and accordingly, quickly changes her intention and tells the truth.

There are a few things we can say about this case. It seems, at first glance, as though Susan's being *struck* by the wrongness of her intention is equivalent to her 'just seeing', introspectively that is, the wrongness of her intention; she plausibly comes to learn something new when in contact with her habitually formed intention to lie. Susan also undergoes a *phenomenological change* with respect to her mental state. That is, there is a change in, or generation of, a *second-order* mental state that is about her *first-order* mental state in which Susan has a new *moral experience*, one best characterized as involving access to a new *phenomenological datum*. The focus moving forward will not only centre around this phenomenological point, but also Susan's seemingly newly formed normative belief *that her intention to lie is wrong*. To be clear, we are asking what Susan's "being struck" consists of. One way to be somewhat neutral with respect to that question is to describe Susan as developing a *second-order state*, one that importantly is accompanied by a phenomenological feeling of wrongness[149].

So, plausibly, Susan occupies a novel, second-order state with the following sort of content:

> Second-Order State (SOS): "*That* is wrong"[150]

---

[149] There is a question here about how best to characterize Susan's experience, e.g., as either a felt-reflexive demand or a feeling of wrongness. But I don't think too much hangs on that, so in what follows I sometimes speak interchangeably of the two. One might think that construing Susan's experience as a felt-reflexive demand is too relational to be introspectable. I have my doubts about that (for arguably many mental states which are relational are still introspectable), but if one is sympathetic to that idea, then the account of normative introspection I give in Sect V should alleviate those worries.

[150] Or: '*My* intention to lie is wrong'. Importantly, the state is about something having to do with *her mental states* and need not involve the demonstrative 'that' nor, as I go on to argue below, need it contain any particularly *normative content* to constitute a genuine instance of normative introspection. Note, also, that we may substitute the predicating property for any plausible normative property which might qualify our mental states. For example, in the case of unpleasantness, we might occupy a state such as: '*That* (the unpleasantness) is bad'; in the case of my motivation to support my partner, the state might be: '*That* (my motivation) is good/estimable'; in the case of my unfaithful desire, the state might be: 'I ought to get rid of *that* desire', and so on and so forth for any putative mental state which can be qualified by a normative property.

Also note that the state in which Susan is *struck* by the wrongness of her intention could be *first-order*. That is, the first-order intentional state could possibly present *itself* (and any accompanying normative property) to itself such that there is no numerically distinct higher- or second-order state representing or accessing the lower- or first-order state. Such a view has recently been developed by Chalmers (2003), Gertler (2001, 2012), Kriegel (2007, 2009), Guistina and Kriegel (2017), and discussed in Guistina (2022). One thing to note, however, is that such a view is developed in terms of accounting for phenomenal consciousness and not propositional attitudes so

where the demonstrative 'that' refers to her intention, hence, is second-order.[151] Before I go on to motivate normative introspection with respect to SOS, let me clarify what the desiderata are in this case. I take there to be at least three things in need of explanation. I have so far spoke loosely of "being struck" by something. I think this notion of "being struck" can be parsed into two parts: one about the *immediacy* of the new state and one about the particular *felt-quality* of the new state. The first two desiderata are the following: Susan undergoes a novel moral experience that has two key features. First, her moral experience is *immediate*, hence, the sense in which Susan is *struck* by something. And second, her being struck is accompanied by a particular phenomenological feeling, something-it-is-like to undergo that particular moral experience. Given that Susan decides to tell the truth partly based on this feeling, I take it that her moral experience can be plausibly characterized as a *felt-demand* not to lie: what *immediately* strikes her is the demandingness of her situation with respect to her intention, namely, not to act on *it*.[152] So, we can formulate the first two desiderata:

> **(D1) Immediacy**: An account of SOS ought to accommodate the felt *directness* or *immediacy* of Susan's SOS.
>
> **(D2) Felt-demandingness**: An account of SOS ought to accommodate the felt *demandingness* or *wrongness* of Susan's SOS.

There is a third and final desideratum that I think is crucial to properly capturing the nature of Susan's SOS. It that has to do with what I call the *asymmetry* of felt-demandingness. To get a sense of what I have in mind here, consider SUSAN, except this time rather than have Susan be asked whether *she* has cleaned her room, imagine that Susan's sister Bethany is asked whether she has cleaned her room. Imagine further that Bethany lies to her mom about having done so, and that Susan is the lone spectator with respect to the unfolding of this lie. Let's stipulate that Susan herself is aware that Bethany has not cleaned her room and directly observes her sister tell a lie. In other words, Susan observes the wrongness of her sister's lie. It's not too far-fetched

---

might only be plausible if extended to the normative properties of phenomenal concsiousness. I discuss this in more detail below.

[151] Note, also, that had Susan not formed the intention to lie, but rather only felt the motivational pull to lie, we could still formulate a similar claim about some normative judgement Susan makes about her motivational state. In this case, the normative property might change from being wrongful to being bad. Or we could say that Susan judges that it's wrong to lie/intend to lie on the basis of her introspective access to her motivational pull to lie, a line of thought which will be pursued below.

[152] Or, if one prefers, simply the wrongness of her intention.

to assume that the phenomenological feeling Susan undergoes when she observes her sister's lie (and its putative wrongness) and is thereby struck by the wrongness of the act is importantly *different* from the feeling she undergoes when it concerns her own intention to lie. Say, Susan even forms the following judgement: "*That* is wrong" where 'that' refers to Bethany's putative lie. Intuitively, the felt-demandingness or wrongness of Susan's own SOS is quite a bit *stronger* than any new SOS Susan might occupy with respect to her sister Bethany's wrongful act. I take this to be an important desideratum in accounting for SOS for there seems to be an important difference between the sorts of states (or judgements) we make about *ourselves* versus the states (or judgements) we make about those of *other people*. An account of the nature of our moral experiences ought to accommodate the intuitive felt-difference in phenomenology between first-personal states (or judgements) and third-personal states (or judgement) of wrongness; that is, between the phenomenology of *reflexive demands* and the phenomenology of what we might call *removed moral judgements of rightness and wrongness*[153]. So, we have our third and final desideratum:

> **(D3) Asymmetry**: An account of SOS ought to accommodate the phenomenological difference between first-personal states/judgements of wrongness and third-personal states/judgements of wrongness.

Now, what I want to claim is that Susan arguably develops a *recognitional sensitivity* to the wrongness in her intention to lie which is indicated by the difference in second-order states she has before and after the recognitional sensitivity develops (Siegel, 2006). Furthermore, I claim that this recognitional sensitivity is best captured in terms of her *introspective abilities* to recognize the normative properties of her mental states[154]. And the accompanying phenomenal feel of her mental state's wrongness—the sense in which it *strikes* her as wrong and she feels a demand to do something about it—is explained by her direct introspective access/recognitional

---

[153] See Mandelbaum (1955) for the claim that our *direct* moral judgements (first-personal judgements) are importantly phenomenologically distinct from our *removed* moral judgements (third-personal judgements). In particular, he claims: "Thus, the stirredupness and pressures which are present in direct moral judgements have no counterpart in removed moral judgements" (127). Note that there isn't quite the overlap here between how I have put things in terms of first and third personal and how Mandelbaum puts things in terms of direct and removed judgements. For Mandelbaum includes first-personal memory judgements in the category of removed moral judgements.

[154] For the time being, I speak in terms of one's direct introspective access to the normative properties themselves, and do not want to commit myself to any claims about what 'access' here might entail. For instance, I do not want to commit myself to the claim that when we 'access' these properties we, e.g., *represent* the normative properties in introspection, or that we are *directly acquainted* with them, or that we access them by first *looking outward* onto the world (although we saw reason to reject a view like this latter one in Chapter two).

sensitivity to the wrongness present. In other words, she develops a recognitional sensitivity that is expressed in a state of *normative introspection*. SOS becomes:

Second-Order ***Introspection*** (SO**I**): "(Introspectively) *That* is wrong."

Note two important things here. Although I couch the relevant introspective state in terms of possessing normative content—e.g., *wrongness*—I by no means think this is necessary for normative introspection to be plausible. That is, although SOI is an instance of normative introspection, it isn't the only one. There are two broad ways to develop normative introspection. First, Sarah McGrath, for instance, endorses a view about moral *perception* whereby for a subject to have a genuine moral perception it is enough for them to perceive basic, low-level, non-normative properties which provide immediate, non-inferential justification for first-order normative beliefs.[155] No specifically *moral* content is included in one's perceptual *experience*. In motivating normative introspection, we shouldn't restrict it to the possession of normative *content*. Second, nor for that matter should we restrict the plausibility of normative introspection to a *mediating second-order non-doxastic introspective experience*, whether it contains normative content or not, something akin to a non-doxastic perceptual experience: there need not be an introspective experience-like state that stands between one's introspective beliefs and the mental items those beliefs are about. In this case, SOS can be read as an immediate, non-inferential *judgement* (as opposed to a mediating non-doxastic second-order experiential-like state) *based directly* on either i) an introspective experiential-like state with or without normative content or ii) simply on the first-order state itself; either is compatible with normative introspection. Below, I outline the various ways we might go about filling in SOI. But my language in what follows will more closely resemble talk *as if* Susan's being struck by her wrongness entails that she occupies a non-doxastic second-order experiential-like introspective state (something akin to a first-order perceptual experience) *with* normative content. But I use the term 'state' in an admittedly loose sense.

SOS seems to meet at least three plausible markers of introspection. First, SOS seems to be *direct, immediate,* and importantly, *non-inferential*: it's not inferentially based on any prior doxastic state. Second, if we construe SOS as a non-doxastic state, SOS seems poised to further provide the *justification* for Susan to form the belief *that her intention is wrong*: it seems well-

[155] See McGrath (2004, 2011, 2018).

suited to play the basing role (non-inferentially construed) for the belief that her intention is wrong. Third, SOS seems to be importantly *inwardly directed*, that is, it is about Susan's own mental state rather than something extra-mental. Now, it's not all together uncontroversial whether introspective states can have phenomenal properties. But *if* we grant that Susan has *some* sort of phenomenological experience, then it is plausible to think that that feeling is constituted in part by her direct introspective access to the wrongness of her intention. Of course, that is controversial, but the reader is welcome to bracket that phenomenological datum to the side and focus merely on a phenomenologically-neutral counterpart to SOS above and ask what best explains that (e.g., a non-phenomenological judgement). In such a case, I think normative introspection is just as plausible as alternative explanations without the added phenomenological datum.

Normative introspection accounts for (D1), (D2), and (D3) nicely. As stated above, introspection is paradigmatically a non-inferential process that can result in *immediate* second-order mental states targeting first-order mental states. Although there are some exceptions[156], introspective awareness is typically direct and immediate. Similar to how the unpleasantness of a sensory episode might become immediately and directly present to my introspective awareness, it is reasonable to think that if we can normatively introspect properties like wrongness, then they too will immediately produce second-order introspective states.

And if normative introspection can easily account for (D1), I think it can easily account for (D2). If we are, in our introspective attention, some of the time immediately struck by our own mental states, then it is plausible to think that there might also be an accompanying phenomenology with respect to our direct access to our own mental states. If I can be immediately struck by my mental states, then I see no reason why that idea cannot carry over to the properties of those mental states: plausibly, I can be struck by the properties of my mental states. And if we grant that there is, in some sense, deontic properties which we can access via introspection, it isn't too far-fetched to assume that there can be a particular phenomenal feel to that sort of access: it feels demanding—i.e., my intention demands me to do something about *it*—because I stand in some sort of direct relation to the wrongness of my intention.

[156] See, e.g., Dretske (1995, 1999).

(D3) is also easily accounted for. Normative introspection explains the difference between first-personal deontic judgements (reflexive demands) and third-personal deontic judgements (removed moral judgements) in virtue of the fact that we are more intimately and directly acquainted[157] with our own mental states than we are with the mental states of other people. That difference plausibly tracks the difference in felt-demandingness between what is presently demanded of me in my situation—and my access to it—and what is demanded of another person in a similar situation—and my access to their subsequent success or failure to respond appropriately. In other words, assuming a phenomenological difference in first-personal and third-personal instances of SOS, normative introspection nicely explains that difference: we have sufficiently direct access to the deontic property in the first-personal case whereas we do not have sufficiently direct access to the deontic property in the third-personal case.[158] Note that we may even think of the third-personal reading of SOS as an instance of normative *perception*: Susan perceives her sister's wrongful act, and is struck by the wrongness (or the sense in which her sister ought-not-to-have-done-that; or some similar normative notion) of the act. Even in this case, where we seem to stand in a suitably *direct relation* to the wrongness of the act, I think it is reasonable to assume that the first-personal case about the wrongness of our own intentions/actions is still phenomenologically distinct from the third-personal case in the sense that it feels like something is putting *pressure* on us to do something in a way that just isn't felt with respect to our third-personal stance. That pressure, I claim, is best captured by appeal to introspection.

Now, to help bolster the plausibility of normative introspection as the best explanation of SOS, let me consider other possible explanations. I consider three: inference, a priori intuition, and emotion.

One possible alternative explanation to SOS which does not identify SOS with Susan's introspective abilities, is that Susan engages in a sort of inference from background beliefs and general principles to the conclusion that her intention to lie is wrong. Susan's newly developed SOS is not a product of a newly developed recognitional sensitivity, but rather through a bit of ethical reasoning. SOS gets filled in with an *inferential judgement*, something like the following:

---

[157] In what follows, I sometimes use the term 'acquaintance' for ease of exposition and use it to resemble talk of 'access'. I will flag to the reader if and when I use 'acquaintance' in its more technical sense.

[158] A notable exception to the intuition that there is a felt phenomenological difference between first-personal and third-personal singular deontic judgements is Broad (1944). I consider his discussion below.

(1) (First-order intention): I have an intention to lie (non-normative, introspective knowledge)
(2) (First-order judgement): If I have an intention to lie, then my intention to lie is wrong (normative knowledge)
(3) (Second-order judgement): *That* (my intention) is wrong (inferential normative knowledge)

SOS becomes:

Second-Order ***Judgement*** (SO**J**): "(Inferentially) *That* is wrong"

In other words, no need to posit a normative introspective state to explain SUSAN. Rather, SOS is explained by a seeming cognitive state about her intention arrived at inferentially. The first response to this line of reply is that SOJ might be *psychologically* unrealistic—it fails to account for (D1)— and second, *if* we grant that there is a what-it-is-likeness to Susan's *being struck* by her intention's wrongness—a *felt demand* not to lie or a felt demand to get rid of her intention—then it might also be *phenomenologically* unrealistic; in other words, it fails to account for (D2). And furthermore, it's not clear it has the resources to account for (D3) either.

First, SOJ cannot account for the seeming *immediacy* of Susan's SOS, that is, the immediate and direct sense by which she is struck by the wrongness since it is dependent on antecedent mental transitions from one content to the next, resulting in SOJ. All that doesn't seem immediate enough.

Of course, proponents of the inferentialist story will point out that the inferential transitions will be fairly automatic and unconscious, resulting in a seemingly *immediate* inferential judgement. One reason to think this is that in SUSAN, she gains what seems to be new background beliefs about the wrongness of lying and is exposed to general moral principles. In other words, Susan's *being struck* by her wrongness is dependent on various background beliefs which give us reason to believe that her SOS is a new second-order *judgement* inferentially arrived at rather than a direct introspective state.

The claim that Susan's SOS is dependent on background beliefs is enough to make it plausible that we read SOS as an instance of SOJ. But it isn't necessarily a *better* explanation than the claim that Susan introspectively accesses the wrongness of her intention. That is, it doesn't necessarily have an advantage over SOS read as an instance of SOI. For starters, pointing out that SOS is dependent on background beliefs does not entail that it is inferentially arrived at.

Vayrynen (2008), following Pryor (2000)[159], points out in the context of ethical perception and its potential to non-inferentially justify first-order ethical beliefs, that just because something like observation is dependent on background beliefs—in other words, is theory-laden—need not entail that it is therefore based on an implicit inference (497). We can say the same thing for introspection. We can say that one's background beliefs *causally affect* the introspective states one has in a way that does not result an inference.

There are at least two ways to fill this in, and although the details need not concern us here, let me briefly mention what they are. The first is to claim that background cognitive states like beliefs, concepts, intuitions and emotions can help subjects gather attentional resources which bear on the direction of one's introspective attention with respect to the properties of one's mental states. In this case, certain background states help make us more introspectively attuned to the normative properties of our mental states.[160] The second way to understand the causal efficacy of background states on introspective states is via the process of cognitive penetration (Cowan, 2015). The rough idea there is that what a subject cognitively thinks can influence the content of what the subject phenomenally experiences in a way that goes beyond merely directing their attention to features of what their experience is about. Rather it must involve the alteration of some content within the subject's experience.[161] Now, the phenomenon of cognitive penetration has been exclusively restricted to perceptual experience and its representational nature, and as far as I can tell nothing has been said about its possible application in the case of introspection. But in any case, the resources are there to account for how a subject's background states might influence the production of a new state like SOS without resulting in the production of a full-blown inferential judgement. We can resist the inferential judgement reading of SOS and instead utilise the evidence in favour of that view for the normative introspection reading of SOS, namely, SOI. The plausibility that cognitive penetration can operate in the case of introspection is especially bolstered if we construe introspection along broadly perceptual, quasi-representationalist lines[162], but again, that might not be necessary.

---

[159] Vayrynen is responding to Sturgeon (2002).

[160] Note that what I above called *first-order*, self-representational accounts of introspection seem reliant on this picture of the influence of background states on introspective access, namely, that they contribute to the distribution of one's attentional resources.

[161] See Cowan (2013a, 2013b, and 2015) for discussion of the possibility that cognitive penetration is required to explain the possibility of ethical perception.

[162] I take it that cognitive penetration paradigmatically operates on perceptual experiences. So, if we construe introspection along quasi-perceptualist lines, then this will prima facie make cognitive penetration in the context of introspection more plausible compared to other, non-perceptualist accounts of introspection.

If we grant that SOS is a distinct *phenomenal state*—thinking about the *felt-demandingness* of Susan's state—then the inferential reply seems lacking. How might an unconsciously produced inferential judgement account for (D2)? I think it's highly dubious that unconscious inferences can account for robust phenomenology. It's implausible to suggest that something inferential *and* unconscious nonetheless shows up as something entirely different, namely, *conscious* and *non-inferential*. One plausible move they could make is to appeal to the normative content of SOJ. That is, one way to account for the felt-demandingness of SUSAN is to have the phenomenology be *cognitive* in nature rather than *experiential*/*sensory*. That's plausible, but it's not clear it has an advantage over positing a direct introspective state constitutive of the felt-demandingness of SOS. But more importantly, the idea that felt-demandingness emanates from a *cognitive judgement* with normative content is consistent with normative introspection, for the *directness* of the judgement might itself be *directly* produced in automatic response to an underlying introspective state, one which *either* does or does not have as its content the normative property of wrongness.[163] We get either one of the following pictures:

(4) (Second-order non-doxastic state): I intend to lie.

(5) (**Direct, non-inferential introspective judgement**) *That* (my intention) is wrong.

or

(6) (Second-order non-doxastic state): My intention to lie is wrong.

(7) (**Direct, non-inferential introspective judgement**): *That* (my intention) is wrong.

Importantly, a direct cognitive judgement which constitutes the phenomenology of SOS is consistent with no mediating background normative belief helping inferentially produce the cognitive judgement.[164] That is, if we appeal to the directness of an inferentially (unconsciously)

---

[163] This might be similar to the McGrathian idea discussed above.

[164] Note another compatible picture between implicit inferential processes and 'seeing', introspectively or perceptually, normative properties, that is, having a particular normative phenomenology. Pekka Vayrynen (2018) claims that rather than account for perceptual moral phenomenology in terms of moral content figuring in directly to the perceptual content, the non-moral content can figure in the perceptual experience which is part of a broader mental state which also includes a representation (albeit non-perceptual) of the moral properties in question. He states: "When it comes to experiences like [seeing a cat burned alive], it is one thing to say that an overall mental state that has a perceptual experience as a component can also involve a representation of a moral property as another component, quite another to say that the moral property figures in the content of that perceptual experience." Rather, our moral phenomenology of 'seeing' the wrongness of the cat being burned alive can consist in something like the following: "The alternative I'll adopt in order to facilitate concrete comparisons is that when Norma sees what the hoodlums are doing in Cat and represents it as bad, this representation results from an implicit habitual inference or some other type of transition in thought which can be reliably prompted by the non-moral

produced cognitive judgement with normative content to account for (D2), then we can equally appeal to such a thing in the case of introspection. The difference is whether the cognitive judgement is mediated by an inference or not. SOS might be a cognitive judgement but one that is in direct response to an underlying introspective state. But note that if we think the prior introspective state contains normative content, as I've been assuming it does—e.g., it represents wrongness or is a state directly acquainted with it as expressed in (6) above—then it is more likely that the phenomenology—the felt-demandingness—is determined by the prior introspective state (6) rather than by the cognitive phenomenology of the direct cognitive judgement produced. And this explanation is at least less controversial since it doesn't appeal to cognitive phenomenology[165].

Furthermore, if the best way to account for the felt-demandingness of SOS is to posit cognitive phenomenology, it's unclear then that inferentialists can properly account for (D3). Why would we expect to have distinct phenomenology between first-order states and third-order states about wrongness when both seem to entail the same normative, cognitive content? In the first-personal and third-personal cases, both refer to a singular proposition expressed by: '*that* is wrong'. Now, the only thing that could account for a difference in phenomenology between the two cases is the change in content underlying 'that' above. But note what that change amounts to. It's a change between my intention and another's action. It's unclear how those two distinct *cognitive* elements of one's singular deontic judgements is supposed to explain the difference in the felt-demandingness experienced in the first-personal case and the lack of that feeling in the third-personal case. More simply, in the first-personal case, one directly introspectively accesses the wrongness; in the third-personal case, one stands in an unsuitably indirect relation to the wrongness of the observed act.

Normative introspection seems to fair *at least* as good as an inferential account of SOS[166]. But if the inferentialist reply is committed to the notion of cognitive phenomenology to explain the

---

perceptual inputs jointly with the relevant background moral beliefs." A similar story, I take it, could be appealed to in the case of introspection. On this account of moral introspection, (4) and (5) would be unified as one kind of mental state and would appropriately be classified as an introspective state/experience. But note this will have negative consequences for the thesis that normative introspection provides an epistemically independent way to arrive at non-inferentially justified first-order beliefs.

[165] Normative introspection can eschew appeal to cognitive phenomenology in the sense that it need not locate the phenomenology-constituting fact in an act of cognition.

[166] I should note one other way in which the inferentialist story might be rejected. According to McGrath (2018), an inferentialist story about justification of immediate judgements of the wrongness of singular actions of other people is epistemically implausible. McGrath argues that it is implausible to prefer an inferentialist story over a

felt-demandingness of SOS, and normative introspection is *both* compatible with it and can even eschew appeal to it, then normative introspection has the upper hand. For normative introspection could appeal to the phenomenology of a second-order *non-doxastic* introspective state, which, to my mind, need not entail anything about cognitive phenomenology.

Now, let me turn to the possibility of accounting for SOS in terms of *a priori intuition*. On one attractive characterization of intuition, for S to have an intuition that *p* is for S to be in a mental state where it *intellectually seems* to S that *p*.[167] Here we interpret SOS as:

Second-Order ***Seeming State*** (SOSS): '(Intellectually seems) *That* is wrong.'

Now, SOSS is plausibly produced by Susan's adequate reflection on some proposition *p* where *p* refers to a proposition with the following content [I intend to lie] or [my intention is to lie] or some relevant variation such that adequate reflection on the proposition leads to an intellectual seeming state, hence, is a priori. However we fill in the exact content of *p*, for a priori intuition to constitute a plausible explanation of SUSAN, it must tell us a story about the immediacy of Susan being struck by the wrongness of her intention and the phenomenology of the felt-demandingness of that normative property. Now, intellectual seemings are typically understood to have presentational phenomenology of the truth-makers of the propositional content (Huemer 2005; Chudnoff 2011), and hence, seem well-suited to explain the phenomenological challenge in SUSAN: SOS feels demanding—i.e., Susan is *struck* by the wrongness—because SOS consists in an intellectual seeming state with respect to the moral concept WRONGNESS and INTENDING TO LIE. Furthermore, intellectual seeming states can either be causally produced or constituted by an adequate understanding of a self-evident proposition *p* or they can be produced without any prior reasoning, reflection, or inference but on some general notion of concept apprehension (Huemer, 2005). For instance, adequately reflecting on the concept

---

perceptual one in cases where we make immediate judgements about the wrongness of an individual's singular action, for the inferentialist will need to appeal to a premise which entails knowledge of what the action is. McGrath thinks this already involves a kind of perception which will either i) entail that one can perceive properties that are extremely rich which then bolsters the idea that moral properties *are* perceivable, or ii) entail that one only perceives *thin* properties which then makes the premise that the inferentialist needs to account for moral knowledge epistemically implausible. This line of response, however, doesn't quite carry over in the case of introspection. For the non-moral (non-normative) premise which the inferentialist must appeal to—premise (2) above—in order to account for moral knowledge that one's intention is wrong is epistemically innocuous, at least in the sense relative to reject on similar grounds as McGrath rejects it in the perception case. Premise (2) above might be epistemically suspect in that it just isn't *required* for one to have justified beliefs about their particular mental states.

[167] See Chudnoff (2011) and Huemer (2005) for these views. See Bedke (2010) and Cowan (2017) for discussion.

WRONGNESS or MORAL REASON AGAINST, one will be able to conceptually pull out the concept INTENDING TO LIE. Or, conversely, adequately reflecting on the concept INTENDING TO LIE allows one to recognize more general conceptual truths about the relation between the concepts of LYING and WRONGNESS, ultimately resulting in an appreciation that one's singular intention to lie is wrong. Importantly, SOS is accounted for via the phenomenology of seeming states.

Intellectual seemings can plausibly capture (D1) and (D2). But I think there are three problematic things with this story. First, intellectual seemings cannot account for (D3). If SOS is best explained by appeal to an intellectual seeming state which is best characterized as some general apprehension of the general concepts INTENDING TO LIE and WRONGNESS, then it's unclear why such an account would yield different phenomenological verdicts in first-personal and third-personal readings of SOS. For what first and foremost underpins SOS—in either its first-personal and third-personal form—is an intellectual seeming state produced by some apprehension of the general concepts of INTENDING TO LIE, WRONGNESS, and ACTS OF LYING. But it's unclear how we then move from the phenomenology of the intellectual seeming state associated with those *general* concepts to the different phenomenology between the application of that seeming state to one's *own* mental states and the *actions* of other people. The impression of a felt-demand directed at *myself* seems to be livelier in nature than any impression of a felt-demand I might feel directed at *someone else*. Furthermore, not only does Susan's SOS *seem* to present her with something associated with wrongness, but rather something is *impressed* upon her consciousness in a way that isn't in matters concerning the acts of other people. In short, a felt intellectual seeming state does not have the resources to accommodate the robustness of the felt-demandingness of Susan's second-order, first-personal state: there is something *impressing* itself on her consciousness which the mere apprehension of concepts cannot accommodate. Rather Susan's being struck is explained, simply, by her normatively introspecting the wrongness of her intention.

Second, notice how, on some views of intuitions as intellectual seeming states (Chudnoff, 2013), intellectual seeming states have an abstract subject matter (Cowan, 2016). But this is problematic for the present case. For SOS is about a particular, *concrete* case. So, read non-doxastically, SOS can't be an intellectual seeming state for it is about a particular case. One might respond to this and claim that although SOS read non-doxastically can't be an intellectual

seeming state, it can be a second-order judgement arrived at via an inference with two supporting premises, one of which is an intellectual seeming state connecting types to tokens—e.g., if it intellectually seems to me that intentions to lie are wrong, then particular intentions to lie are prima facie wrong (Ibid., 71). But, as Cowan (2016) points out in the context of deontic judgements about singular cases—e.g., that the political budget is wrong—it is highly epistemically suspect that subjects *first* have justification about the *type* of case in order to have justification about the *singular token* case. That is, Susan doesn't seem to have a justified belief about the type of situation she is in with respect to her occurrent mental state in order to arrive at a justified belief that her intention to lie is wrong. Moreover, nor does Susan *need* a justified belief of that kind. Therefore, we have reason to reject the intuition model with respect to SOS.

Note one final thing here. Michael Huemer, a leading contemporary intuitionist, argues for the plausibility that intuitions are (initial) intellectual seeming states on the grounds of the plausibility of the epistemological category of a *seeming state*. According to Huemer, this category already includes perceptual and *introspective* seeming states. So, I see no reason why one couldn't appeal to a plausibly less controversial type of seeming state to account for cases like SUSAN, namely, an introspective seeming state. Importantly, introspective seeming states nicely accommodate the worry raised above about the generality of intellectual seemings: they don't take particulars as their content. But introspective seemings plausibly do.[168]

Finally, consider the following example:

> **Uneasy Ed:** Ed's friend June invites him to a party. June tells Ed that their mutual friend Rebecca isn't invited. Ed knows that if he tells Rebecca that he's going to June's party where she's not invited, Rebecca will become seriously irate. Ed does a bit of thinking and concludes that the best thing to do is to lie. Ed forms the intention to lie to Rebecca. Although feeling *uneasy* about the lie, he makes the tough decision and lies to Rebecca, saving her a potentially serious bout of anger.

[168] For example, Sosa (2012) appeals to introspective seemings in his general account of introspection. Although there are no introspective *experiences* about our own mental states which provide justification for introspective beliefs there are nonetheless introspective *seemings* based on the mental states they are about which provide justification for introspective beliefs. Note that for Sosa these introspective seemings are intellectual attractions to assent to a proposition. So, the extent to which an account like Sosa's is a plausible account of SOS, and hence, normative introspection, depends on the nature of the propositions which one can be intellectually attracted to assent to; in particular, whether we can be attracted to assent to singular propositions.

Grant that Ed's uneasiness about his lie has a particular phenomenal feel to it; that is, there is something-it-is-like to undergo that feeling; we might even call Ed's feeling *uncomfortable* or *unpleasant*; and furthermore, we can reasonably imagine that he'd rather not feel the feeling if he had the choice. But, unfortunately, as is often the case, we can't help feeling how we feel. In the case of Uneasy Ed, *if* we accept the plausibility that we can have substantial phenomenal feelings about our own mental states, then we can ask what most plausibly explains Ed's uneasiness? That is, in this case, we now ask what best explains Ed's *persistent feeling* of uneasiness? So far, I've argued that both an inferential and an intuition strategy lack the explanatory power to accommodate some of the phenomenological data in SUSAN. But let's grant for the sake of argument they can. What could possibly explain *persistent negative feelings toward one's own mental states*? For starters, we might balk at the suggestion that inferential judgements can have a persistent phenomenology. Second, and more importantly, in the case of an inferential judgement about the wrongness of Ed's own intention, he will have made *another* inference about what the all-things-considered reason to do in the situation is. Why, we might ask, has Ed's uneasiness persisted *even though* he's made a different inference defeating the one about the wrongness of his intention to lie. We might think that any phenomenological feeling associated with Ed's inferential judgement about his intention would vanish in the light of his other, *defeating* judgement about what he has all-things-considered reason to do.

A priori intuition might more readily accommodate the idea of persistent phenomenology, for it seems plausible to imagine holding an intuition in mind for an extended period of time: intuitions can persist. And insofar as they can persist, so can their phenomenology. We also might more easily accept the possibility that the uneasiness constituted by Ed's intuition that lying is wrong can persist in the light of his all-things-considered-judgement that he should lie. But here's a potentially simpler explanation: Ed is introspectively aware of the wrongness of his intention to lie, and the persistent feeling of uneasiness he feels—his discomfort with what he intends to do—is constituted by that very same introspective awareness. Ed's uneasiness is the result of his *continual* direct introspective access to the wrongness of his intention to lie. He is still introspectively aware of the pro tanto wrongness of his intention—it hasn't vanished. Ed will be continually aware of the intention he has to lie since he is in the process of bringing it about that he lies to Rebecca.

But notice another, equally plausible explanation for the data found in SUSAN: emotion. Emotions seem to be a natural reading of SOS as well as Ed's uneasiness. Susan occupies a second-order emotional state with respect to her first-order intention; it plausibly can account for the felt-demandingness: the emotion is constituted by one's direct access to the normative property of one's own mental state; the emotion is elicited as a response to the presence of wrongness. Emotions can persist even well-after their objects have vanished. Emotions might also nicely capture (D3). Our emotional responses are plausibly going to be quite distinct when they concern properties relating to ourselves and properties relating to other people: emotions can directly take as their content the particular mental state itself and its accompanying wrongness. So, emotion might be the way to go with respect to SOS.

Emotions are indeed plausible candidates. But note one thing. It's unclear what emotion we might identify the felt-demandingness with. For it does not seem to fall under any of the more traditional moral emotions like guilt, anger, indignation, etc. In that case, those attracted to an emotion account of SOS might have to posit a *sui generis* emotion with respect to the wrongness of one's intention (or the felt-reflexive demand). Also, positing a sui generis emotion to explain SOS might entail something like normative introspection in the first place: we have a distinctive emotional response of felt-demandingness with respect to the wrongness of our intentions to lie because our emotional state is constituted by a normative introspective state. Normative introspection is at least compatible with an emotion account of SOS and might even be needed to illuminate the structure of such introspective emotions.

Of course, I don't pretend that any of this is conclusive reason to think that SOS can't be read along either inferentialist, intuitionist, or emotion lines. What I do think the preceding discussion does is bolster the plausibility of normative introspection and gives us reason to begin to take seriously the idea that we can and do introspect normative properties. If that's true, then normative introspection might constitute a novel a posteriori way of gaining normative knowledge. That is, some of my normative knowledge, e.g., that intending to lie is wrong, is dependent on what I know about the contents of my own mind; sometimes, all I have to do to gain normative knowledge is look within. But before I turn to those issues, note also that although I've framed things in terms of a second-order introspective state which contains normative content, that need not be the case. As will become clearer below, it might be possible

that we gain normative knowledge via introspection *without* having introspective states which contain normative content. Let me know turn to this.

5 Normative introspection without normative properties

In the previous section, I motivated normative introspection with a phenomenal contrast argument and argued that relative to an inferentialist, a priori intuitionist, or emotion account of Susan's newly formed second-order state, normative introspection is *at least* as good of an explanation as those three, and in some cases, might do better than them. Recall:

> Normative Introspection (NISP)*:* at least some normative properties are introspectable.

In SUSAN, I formulated a version of NISP by appealing to an introspective state that contains *normative content*. On this reading of NISP, we get something like the following:

> Contentful Normative Introspection (C-NISP): at least some normative properties figure in the contents of introspective states

So far, I have been neutral with respect to what is supposed be meant by 'state'. On one interpretation (and what I was assuming above), 'state' means something akin to a perceptual experience. Similar to how, on some views of perception, perceptual experiences stand between perceptual beliefs and the extra-mental objects and properties those beliefs are about, introspective states qua experiences stand between introspective beliefs and the mental objects and properties those beliefs are about. On this line of normative introspection, when Susan introspects the wrongness of her intention to lie, she occupies a *non-doxastic* introspective state such that the state itself contains normative content by, e.g., representing normative properties. Recall (6) and (7):

> (6) (Second-order, non-doxastic introspective state): My intention to lie is wrong.
> (7) (**Direct introspective judgement**): *That* (my intention) is wrong.

C-NISP rises and falls with the plausibility that some of our introspective states can stand in suitably direct relations to other mental items in a way that involves the possession of some sort of content. There is debate in the philosophy of mind, and the history of philosophy, more generally, whether introspection ever takes such a form. It certainly seems like Descartes,

Locke, Hume, Berkley, and perhaps even Kant[169] had something like this in mind when they spoke about directly accessing the objects of the mind, "inner sense", or having "perceptions of the mind"[170]. So, the company is good. But nonetheless, philosophers have taken issue with the idea that introspection involves a distinct non-doxastic state separate from the putative object of one's introspective beliefs.[171] That is, there is no non-doxastic second-order state like (6) which can provide the epistemic grounds for introspective beliefs about one's first-order mental states like (7). So, the plausibility of C-NISP will depend on how plausible one thinks it is that we can occupy non-doxastic second-order introspective states which contain conceptual or non-conceptual content. This would lead to something like the following:

> *Non-doxastic (contentful) normative introspection (ND-NISP):* at least some normative properties figure in the contents of non-doxastic introspective states.[172]

But this need not be the case. There are two ways we might eschew appeal to a second-order non-doxastic state. For some (e.g., Gertler, 2001, 2012; Loar, 1990; Chalmers, 2003; Papineau, 2007; and Levine, 2007), introspective judgements about our mental states need not entail any mediating non-doxastic second-order state on which they are based. Rather the first-order state which our introspective judgements are about *directly supply* the content of the introspective

---

[169] Kant spoke of an 'inner sense', but it is controversial whether he thought that knowledge of our propositional attitudes could ever be grounded in a kind of inner sense. Boyle (2009) discusses the two kinds of self-knowledge found it Kant: passive and active. Passive self-knowledge is self-knowledge associated with our own sensations whereas active self-knowledge is self-knowledge associated with our rational capacities.

[170] "Perceptions of the mind" occurs in Hume (1748). See Descartes (1641), Locke (1689), Berkley (1713), and Kant (1781/1797).

[171] Perhaps the most extensive attack on this view, sometimes called the 'inner sense model' or the 'object perception model', is given by Sydney Shoemaker (1996). The very rough idea is that introspection just isn't anything like perception to be modelled on it. But see Gertler (2012), Horgan and Kriegel (2007), and Horgan (2012) for defenses of something resembling these models but note that they would not take on the description that their views were 'inner sense' views, but rather would label them 'direct acquaintance accounts'. Note that these thinkers defend it with respect to the introspection of phenomenal states, so the extent to which it extends to normative properties might be restricted to the normative properties of phenomenal states. See also Smithies and Stojlar (2012) for discussion of the various forms of introspection from which the Gertler and Horgan articles also appear. Another interesting thing to keep in mind here. There are various accounts of introspection not all of which are mutually exclusive. There may indeed be different ways we can introspect our mental lives. An interesting thing to consider which I do not have the space here to explore is whether we might be able to introspect the non-normative mental base in one way and introspect the normative property supervenient upon it in another.

[172] Note that there is also a version of this account whereby normative concepts/properties are not contained in the non-doxastic introspective state (i.e., is *non-contentful*), but rather are produced downstream from that state and appear first in the second-order introspective *belief*. The most plausible way I think to elaborate that view is to have subjects undergo introspective seemings without normative content that then non-inferentially produce justified second-order normative beliefs about first-order mental states. This I take it would be similar to the view briefly discussed here given by McGrath (2018). This version is similar to what I call 'reliable normative introspection' below, except the account below does not include a non-doxastic introspective state. Rather, the first-order mental state—e.g., the intention to lie—directly produces a justified second-order introspective normative *belief*. See below for more on this.

judgement where this notion of 'directness' is supposed to be metaphysical and not merely causal (Gertler, 2012: 96). (6) does not mediate between the first-order intention to lie and the judgement that that intention is wrong; on this account, the first-order intention to lie (and its wrongness) directly contribute to the introspective judgement that it (the mental state itself) is wrong. We get something like the following:

(8) (First-order state) ≈ I intend to lie [is wrong] ≈ [*c*]

(9) I judge that [*c*]

The mental state itself in (8)—the intention to lie and its wrongness—directly *embeds* itself into the content of the second-order introspective judgement in (9): *c* is the mental state itself. Now, how *c* can directly embed itself into the judgement in (9) is a complex story, but the point I want to make is that we need not appeal to second-order mediating states to account for the justification of our introspective beliefs; we can skip that state altogether and embed the first-order mental state directly into the introspective judgement. This would lead to something like the following:

> *Direct acquaintance normative introspection (DA-NISP):* normative properties *directly embed* themselves into the contents of second-order introspective (demonstrative) beliefs. [173]

One appealing thing about this view is that it might make sense of the seeming self-evidence of some of our normative beliefs, especially a belief in the proposition that intending to lie is wrong. For some, such normative beliefs just seem self-evident. And on the above sort of view, "one's epistemic grasp of a bit of reality … can be partly constituted by that reality itself." (Gertler, 2012: 101). Why it seems self-evident that my intention to lie is wrong is because my understanding of that fact is in part constituted by part of the fact itself: the wrongness directly embeds itself into the content of my introspective judgement.

[173] Or, if one feels queasy about normative properties directly embedding themselves into demonstrative judgements, then one may stick simply to non-normative properties and have the second-order judgement somehow become imbued with normative content *after* the direct embedding of the non-normative content. So, some non-normative mental property directly embeds itself into a second-order judgement and the normative content somehow gets plugged in during or after this embedding process. This account will have prima facie difficulties accounting for the phenomenological data in SUSAN.

For others (e.g., Armstrong, 1968; Lycan, 1996; Nichols and Stich, 2003; Goldman, 2006), we eschew second-order non-doxastic states by appealing to the perception-like causal nature of introspective judgements. First-order states do not directly embed themselves into second-order demonstrative judgements but rather reliably causally produce second-order beliefs about first-order mental states. Importantly, the normative content of the second-order belief that '*that* intention is wrong' is going to be produced via some background reliable belief-forming mechanism.[174] Insofar as there are second-order beliefs that result from background reliable *introspective* processes, then we can label those second-order normative beliefs introspective. This would lead to something like the following:

> *Reliable normative introspection (R-NISP):* normative concepts/properties do not figure in the contents of non-doxastic introspective states nor are they directly embedded in the contents of second-order (demonstrative) introspective beliefs, but rather are causally produced in the contents of second-order introspective beliefs via some reliable background process.[175]

In any case, there are many ways to elaborate NISP. In what follows, I want to briefly elaborate on a model of how normative introspection might occur in light of the presumed fact that mental states are never *by themselves* (i.e., purely privately) wrong. This will involve drawing on claims made in the affordance literature.

Let's say you aren't convinced that our purely private mental states can ever take on substantive deontic properties like wrongness.[176] That is, it is never morally impermissible to *only* occupy a given mental state, and that in order for our mental states to properly be qualified by something like wrongness or impermissibility they necessarily have to be conjoined with an action: they have to be a composite of a whole action-thought pair for which a deontic property like wrongness supervenes. For instance, an intention to lie is never *on its own* wrong if the particular intention never manifests in a full-blown action. Or, if one finds that intention a little strange, think of some unsavoury motivation, say, to undermine the successes of your wife or husband. *If* mental states, on their own, never have normative properties, then, the thought goes, we can never normatively introspect such properties: merely attending (introspectively) to our mental

[174] Note there's a question about the viability of this process being relevantly *non-inferential*.
[175] Note, however, that this account will have difficulties accounting for the phenomenological data set out in SUSAN. R-NISP also subsumes the alternative account to ND-NISP I gave above in footnote 51.
[176] See, e.g., Sher (2019).

states cannot give us introspective knowledge that some mental state is wrong. Hence, normative introspection (at least as it concerns deontic properties) is implausible.

I think this line of thought can, in part, be resisted. For starters, if we want to defend normative introspection in this case, then it won't be appropriate to call the knowledge it produces *introspective knowledge*, at least not as it concerns knowledge of things that are right and wrong. But I still think that sense can be made of the idea that the introspection of mental states can provide the grounds for justification of *first-order* normative beliefs. We can resist rejecting normative introspection on the grounds that normative properties are never instantiated by (never supervene upon) purely private mental states by pointing out that we can gain knowledge of things without being aware of *all* the properties which instantiate those things. That is, we can come into contact with some properties which are *reliable indicators* of the instantiation of more complex, normative properties. What we need to be possible is that we can come into contact with properties which by themselves never take on the properties they reliably indicate. I think there are lots of cases like this.

To take a non-normative case, consider the property of *being-almost-out-of-petrol*. I'm driving through the Scottish Highlands and notice my fuel light turn on. Presumably, I *see* that I am almost out of petrol *via* seeing that my fuel light is on. Dretske (1995) calls this *displaced perception*. But notice that I haven't perceived all of the properties which help constitute *being-almost-out-of-fuel*; in fact, I plausibly haven't perceived *any* of the more basic properties which partly make up (help instantiate) the complex property *being-almost-out-of-fuel* since the properties associated with my flashing fuel light—the only properties I am presented with—are not what the property *being-almost-out-of-fuel* instantiates. Nonetheless, I take it, I can *see* that I am almost out of fuel. Here we have a case where we come to see some property F in virtue of seeing some property P which itself has nothing to do with the instantiation of F. Similarly, we can introspect the wrongness of an act—namely, the wrongness of the act of lying—by first introspecting the intention which reliably indicates the property of wrongness.

Notice that we seem to be in a *better position* vis-à-vis awareness of the wrongness of an act than we are vis-à-vis awareness of the (almost) empty fuel tank. For in the normative case, we are aware of a *constitutive part* of the property of wrongness whereas in the fuel case we are not. Why is that better? Well, we can imagine being aware of one of the parts which constitute

the property *being-almost-out-of-petrol*—say, you look down and see a portion of the metal bottom of your fuel tank—which would surely bolster your confidence that you *see* you are almost out of petrol: the visible portion of the bottom of the metal fuel tank surely will reliably indicate that you are almost out of fuel, more so than the flashing fuel light *because* it is a constitutive part of the property *being-almost-out-of-petrol*. What is going to be a more reliable indicator of some property is awareness of a constituent part of the property in question rather than something that is merely causally correlated with it.

One might object that the petrol case is importantly different since as a matter of fact the property *being-almost-out-of-petrol* is indeed instantiated. The point of the fuel light is to reliably indicate when a property is in fact instantiated. But in the normative cases we are concerned with, the idea is that *there is no instantiated complex normative property* simply when one occupies a given mental state. So, introspecting a mental state cannot reliably indicate the presence of a normative property for there never are normative properties present *merely* when one occupies a mental state constitutive of a deontic property.

Two replies can be given here. First, normative introspection is still plausible in cases where one's mental state (e.g., intention to lie) does as a matter of fact partly constitute the thought-behaviour composite—i.e., the action—which wrongness supervenes upon. Merely accessing the mental state constitutive of the deontic property will reliably indicate (produce justified beliefs about) the wrongness. In such cases, I have knowledge of the singular proposition *'that* is wrong' where 'that' refers either to my action or the thought-behaviour composite[177] via my introspective knowledge of my intention: my direct introspective access to my intention reliably indicates the wrongness present. But what about cases where my mental states do not manifest in outward behaviour? That is, cases like Susan where she merely recognizes her intention to lie without putting that intention into motion (or, for that matter, *any* mental state which is not properly linked up to a token action)?

I think that question can be more clearly presented as follows:

[177] It's unclear whether those who claim that mental states are wrong only insofar as they are part of a thought-behaviour composite believe further that the isolated mental state is itself wrong or whether it is still only the composite thought-behaviour which is wrong.

(Q1): can we make sense of the idea that some property F can reliably indicate another property G when only F and not G is present?

My hunch is to say yes to (Q1). I think we can make sense of a positive answer to (Q1) only if the connection between F and G is such that F indicates the *possibility* of G. The literature on *affordances* is a promising place to start (Noe, 2006; Siegel, 2014). For instance, Siegel (2014) characterizes an affordance as involving a possibility of action for some creature. For example, seeing the tree in the back garden, one can perceptually experience the tree *as climbable*; or the rhubarb pie on the table *as edible*. Furthermore, affordances can apparently be experienced *as soliciting* a particular action. In the tree case, perceiving the *climbability* of the tree can solicit—i.e., prompt or invite—one to climb the tree; in the pie case, it can solicit one to *eat the pie*. And if one is *moved* to climb the tree or to eat the pie, then one experiences what Siegel calls "an experienced mandate" which is a sub-class of the category of soliciting affordances which involve a relatively high degree of felt solicitation, i.e., motivation (2014). Affordance properties like the tree's *climbability* or the pie's *edibility* can—if experienced as a mandate—'call out' or 'demand' that certain actions be carried out (Mitchell (2021) calls these "action-properties"). Siegel develops her account of experiences of soliciting affordances by focusing on cases where subjects are motivated in the experience of a soliciting affordance. That is, she focuses on experienced mandates. Let us focus on this complex experience in what follows.[178]

Now, Siegel is concerned with defending the claim that all perceptual experiences have representational contents against the phenomenon of experienced mandates. Since experienced mandates, e.g., one's motivational pull to eat the freshly baked rhubarb pie, have two key distinct components, namely, a soliciting component and a motivating component, Siegel attempts to find contents associated with those components. For our purposes, it will suffice to merely point out what those contents *could be*. For Siegel, the soliciting component of an experienced mandate is most plausibly characterized as involving either content of the form 'X is-to-be-ϕ-ed' or 'A-is-to-be-done' where 'X' is some object of perception, e.g., a piece of pie and 'A' is some possible action, e.g., eating the pie[179]. The motivating component, importantly, must

[178] Some of what follows is not needed to argue for the claim that experiencing affordance properties can non-inferentially justify first-order normative beliefs. But for ease of exposition I speak about the entirety of an experienced mandate even though aspects of what those experiences entail are not crucial for our purposes.

[179] Siegel does note a difference between these two contents insofar as experienced mandates seem to be issued from the objects of one's perception and not the possible actions. So, on this assumption, Siegel thinks it best to characterize the content as 'X is-to-be-ϕ-ed' and not 'A-is-to-be-done'. It's not out of the question that different

account for what Siegel calls the 'phenomenology of answerability' the sense in which an experienced mandate is an *answer to* a soliciting affordance. According to Siegel, the motivating component of experienced mandates have the following *answerability content*: Experience: [It is answered that: X is-to-be-ϕ-ed] or equally plausibly: Experience: [It is answered that: A is to-be-done]. Putting the soliciting (non-motivational) content together with the answerability (motivational) content, we get something like the following. I undergo an experienced mandate to eat the piece of rhubarb pie which includes a *feeling of answerability*. The content of that experience can be characterized as follows: Experience [It is answered that: the piece of rhubarb is to-be-eaten] or Experience: [It is answered that: eating the rhubarb pie is-to-be-done].[180]

Notice what the corresponding beliefs might be for each respective content. It is natural to think that a subject who undergoes an experiential mandate like the above could form one of the following non-inferentially justified beliefs: Belief *that* [the piece of rhubarb pie is-to-be-eaten] or Belief *that* [eating the rhubarb pie is-to-be-done].[181] Can we extend this picture of experienced mandates to cover *mental* affordance properties? I see no reason why we can't.

First, notice that this picture of experienced mandates—experiences of soliciting affordance properties as answerable—fits quite naturally with the phenomenological data set out in SUSAN above. We can characterize Susan's phenomenal contrast as one involving the newly developed recognitional sensitivity to an affordance property of her mental state. The "felt-demandingness" of SUSAN just is an experienced mandate to do something with respect to her mental state grounded in her introspective abilities. Importantly, and what is our main focus, is that we must characterize the possible admissible contents of the relevant experienced mandate

---

experienced mandates can have different contents with respect to objects and actions especially if we, as Siegel does, understand mandates as issuing not only from extra-mental objects but entire environmental situations. If it's one's entire environmental situation which issues the mandate, then I take it that it is more plausible that the content of the experienced mandate involves reference to some possible (immediately) future action.

[180] Siegel herself considers the possibility that the content 'X-is-to-be-ϕ-ed' might induce in one the desire or intention to ϕ. She seems to quickly move past this option on the grounds that "one might worry that this fails to respect the way in which perceptual experience is directed outward, characterizing things external to the subject's mind". But in the context in which we are concerned, any sort of worry about the experience not being directed outward enough is alleviated since we are concerned with experiences directed inward. Note, also, that for our purposes we need not include the *answerability content* for we are solely concerned with the epistemic properties of the experienced soliciting affordance and not necessarily with the motivational component. But I take it that the motivational component of an experienced mandate *could* provide the epistemic grounds for access to the situationally relevant possible action.

[181] Things get interesting when Siegel considers whether *rationalizing properties* in the normative sense might also be part of the contents of experienced mandates. For instance, we would add to the above content after the 'to-be-done' or 'to-be-ϕ-ed' bit the following: 'because it looks tasty' or 'because it has such-and-such aesthetic/gustatory properties.'

as involving *reference to our mental states*. Of course, each mental state might have slightly different contents, but here's a gloss on what some of them might be. In the case of an intention to lie, one might reasonably engender an experiential state with the following content: Experience: [It is answered that: the intention to lie is-not-to-be-acted-upon] or Experience: [It is answered that: the action of lying is-not-to-be-done].[182] Although I've characterized the relevant contents in terms of a seemingly *negative* affordance, I take it this is no problem for a general theory of affordances. For a general theory of affordance properties ought to include affordances *not to do something* since such action-properties seem to be regularly salient. For instance, perceptually experiencing a sharp, barbed-wire fence near my side might afford the negative action 'not-to-be-touched' or simply 'don't touch!'.[183] Similarly, I might experience a mandate *not to eat* the left-over rotten pizza. I naturally might form the belief *that the fence should not be touched* or *that the pizza should not be eaten*.

To see this more clearly in the case of mental states, consider an unpleasant pain. As I've mentioned previously, a typical rational response to undergoing an unpleasant pain like a severe migraine is to pop a painkiller. In the terminology of affordance, we can characterize unpleasantness as soliciting the affordance of eliminating the unpleasantness itself. What unpleasantness invites (and quite strongly *motivates*) you to do is eliminate *it*. Plausibly, we have an experiential mandate with respect to our unpleasant pain which contains something resembling the following content: Experience: [It is answered that: my unpleasantness is-to-be-eliminated]. We can then form the non-inferentially justified belief *that my unpleasantness is to be eliminated* or *that I should eliminate my unpleasantness* or *that I have reason to eliminate it*.

But what about cases like one's intention to lie or desire to cheat on one's partner? How do we go from an experience of a soliciting affordance to the non-inferential justification of a first-order normative belief of wrongness? Recall what the putative content would be in the case of undergoing an experiential mandate with respect to one's intention to lie: Experience: [It is answered that: the intention to lie is-not-to-be-acted-upon/performed] or Experience: [It is

[182] If we wanted to include rationalizing properties in the contents of such experiences, plausible candidates for such contents might be something like the following: 'because it is constitutive of/correlated with acts that are wrong' or 'because it is wrong'.

[183] There is no obvious restriction on what the contents of experienced mandates might be, so long as they plausibly capture the sense in which such experiences solicit affordances (putative non-motivational content) and are motivational. Imperative contents might just be able to play such a dual role. See Klein (2007) and Martinez (2011) for claims like this. But see Bain (2011) for criticism about imperative content's motivational plausibility.

answered that: the action of lying is-not-to-be-done]. The respective beliefs would be: Belief that [the intention to lie is-not-to-be-acted-upon] and Belief that [the action of lying is-not-to-be-done]. Can these beliefs yield full-blown deontic beliefs? I don't see why not. Admittedly, the latter belief's content more easily gets us to the belief we want which is *that lying is wrong* or *that to lie in this situation is/would be wrong*. For the content [the action of lying is-not-to-be-done] is closely conceptually connected to other deontic concepts like SHOULD and OUGHT. If some action is not to be done, then this entails that it shouldn't be done (at least pro tanto), and if it shouldn't be done, then plausibly that action is wrong (again, at least pro tanto). And granted we have relevant background concepts it isn't unreasonable to think that subjects do form beliefs about the wrongness of lying on the basis of concepts having to do with whether an action is or is not to (should or shouldn't) be done.

In short, even if we are skeptical about whether deontic properties like wrongness supervene our mental states, introspecting our mental states can nonetheless provide the epistemic grounds for the non-inferential justification of first-order normative beliefs with respect to actions typically associated with those mental states. How so? We experience the soliciting affordance properties of our mental states which solicit actions either to be or not to be done. On the basis of the content of these kinds of soliciting experiences, we can form non-inferentially justified normative beliefs about the actions afforded or mandated (positively or negatively) by our experiences.

So far, I have done three things. First, I motivated the idea that a certain sub-set of our mental states have normative properties (broadly speaking). Second, I motivated the idea that those properties can be introspected. I did this through a phenomenal contrast argument showing that normative introspection is at least as plausible as alternative explanations and, in some cases, provides better explanations than the alternatives given the phenomenological data. So, if one is already on board with the idea that normative properties can supervene our mental states, then they should be sympathetic to the possibility of normative introspection taking place. But I also motivated normative introspection in the case where one isn't on board with the claim that mental states have normative properties, in particular, *deontic* properties. In that case, it's still plausible that introspecting our mental states and the *affordance properties* they have can provide the non-inferential justification we need for first-order normative beliefs. Importantly, those beliefs are justified without entailing that the mental states themselves are morally

right/wrong, permissible/impermissible, ought to/ought not to be occupied. In other words, we can hold the view that there is nothing normatively (read: deontically) problematic with merely occupying a mental state whilst also holding the view that introspecting mental states nonetheless can provide the necessary epistemic grounds for normative beliefs.

So, having made plausible the idea of normative introspection, we now have good reason to think something like an introspective version of Ethical Intuitionism is also plausible. Recall:

> Normative Intuitionism (NI): normal ethical agents have at least some non-inferentially justified first-order normative beliefs.

We can now put forth a novel a posteriori intuitionism:

> Introspective Intuitionism (II): normal ethical agents can and do have non-inferentially justified first-order normative beliefs by having introspective states.

Before concluding, there is one last thing to consider: whether (II) can plausibly provide an *epistemically independent* ground for the non-inferential justification of *substantive* first-order normative beliefs. In the next section I canvass this worry and some related ones.

6 Epistemic independency and substantive ethical thought

Recall that in the introduction I mentioned that the prospects for an a posteriori normative intuitionism hang on the plausibility of whether we can give a naturalistic account of non-inferential justification. Recall Normative Intuitionism:

> Normative Intuitionism (NI): normal ethical agents have at least some non-inferentially justified first-order normative beliefs.

What I hope to have done so far is make plausible the following claim:

> Normative Introspection (NISP)*:* at least some normative properties are introspectable.

If that's true, and under the assumption that introspection is non-inferential, then we can fill in (NI) as follows:

> Introspective Intuitionism (II): normal ethical agents can and do have non-inferentially justified first-order normative beliefs by having introspective states.

The typical naturalistic candidates for non-inferential justification, and therefore, of cashing out (NI), are perception, emotion, memory, and introspection. A good deal of work has centred around the first two, next to nothing on memory, and as far as I can tell, nothing whatsoever on introspection. What I want to do in this final section is briefly present some of the problems that plague other forms of intuitionism and show how they don't plague (II) either by showing that the initial problems can be given adequate responses (in which case I argue those responses can be extended to (II)) or that the problems don't extend to (II).[184] There is, however, one problem which I think is proprietary to (II) which is that it does not justify *first*-order normative beliefs that are directed at the *extra-mental* world. I end by giving a response to this worry. Again, it is worth noting that I by no means intend this to be a full or even partial defense of the views I've put forth here, and that one of the main aims is primarily exploratory.

In the case of perception, it's argued that perception becomes most plausible as a form of intuitionism—Perceptual Intuitionism—just when the contents of ordinary perceptual experiences represent ethical properties.

> Perceptual Intuitionism (PI): normal ethical agents can and do have non-inferential justification for first-order ethical beliefs by having ethical perceptual experiences.

But it's questionable whether perception (Cowan, 2013a, 2013b) [185] can provide the appropriate epistemic grounds for an *epistemically independent* source of non-inferential justification. That is, if perceptual experiences can non-inferentially justify first-order normative beliefs—i.e., if

---

[184] I skip over a common objection to emotional or affectual intuitionism sometimes called the 'reason-responsiveness' objection which, in one of its forms, claims that emotions can't form the justificatory ground for non-inferential (immediate) belief since justified emotions themselves rely on whether their cognitive base is itself justified, introducing a kind of epistemic dependency thought to be inimical to intuitionist epistemologies. For obvious reasons, I don't consider this objection since introspection isn't a reason-responsive state like emotion nor does it involve a separate cognitive base like emotion does.

[185] Cowan, through personal communication, has claimed that he no longer believes that in order for Perceptual Intuitionism to constitute an epistemically independent intuitionism that the contents of perceptual experiences must include ethical properties (i.e., that ethical perception must be possible). He seems open to something like the view of McGrath's mentioned below.

ethical perception is possible—then it is most plausibly because the justification-conferring content of perceptual experience is epistemically dependent in the following sense:

> Epistemic Dependency (ED): a state, d, epistemically depends on another state, e, with respect to content c iff e must be justified or justification-conferring in order for d to be justified or justification-conferring with respect to content c. (Cowan, 2013a)

Is (II) vulnerable to a similar epistemic dependency objection (EDO)? It's unclear whether it is, for it's unclear whether in order for NISP to be true it also needs to be the case that a particular version of NISP needs to be true, namely, one which posits a non-doxastic state with normative content. Recall:

> *Non-doxastic (contentful) normative introspection (ND-NISP):* at least some normative properties figure in the contents of non-doxastic introspective states.

I take it that EDO could be a forceful objection to NISP if we construed NISP along the lines of ND-NISP. But, as should be clear from Sect. V, we need not do that. But there's another complication. According to Cowan (2013a, 2013b, 2015), there is good reason to believe that *if* we can *perceptually* experience the instantiation of normative properties, then it is possible only in virtue of a process of cognitive penetration. And the trouble for Perceptual Intuitionism results from the fact that the most plausible story of how cognitive penetration enables the perceptual experience of the instantiation of normative properties is only if subjects already hold background justified normative beliefs, hence making Perceptual Intuitionism epistemically *dependent*. But it isn't even remotely clear whether something like ND-NISP will have to rely for its plausibility on justified background normative beliefs cognitively penetrating non-doxastic introspective states. This is mainly because it's not at all clear whether our introspective states can be cognitively penetrated.

Now, one might argue that if introspection isn't a cognitively penetrable process, then there can't be non-doxastic introspective states with normative content, hence ND-NISP isn't plausible. This might be a good argument against ND-NISP but would need much more support. For starters, it's just not clear that in order for *any* non-doxastic state to contain normative content it must be cognitively penetrated by antecedent normative beliefs. There are no a priori constraints about what potential process is involved in making normative properties part of the admissible contents of non-doxastic states, at least none that entail that it must be cognitive

penetration. But in any case, we need not rely on the plausibility that non-doxastic introspective states contain normative content in order to arrive at non-inferentially justified normative beliefs.

One way to avoid the potential problem of epistemic dependency is to *deny* a necessary connection between subjects having perceptual knowledge that x is F and the contents of one's perceptual experience representing x as F (McGrath, 2018: 178). We can do the same in the case of introspective knowledge and deny that normative properties must be admissible contents of non-doxastic introspective states. It's possible, according to McGrath, that one could come to have perceptual knowledge that *that action is wrong* without one's perceptual experience representing the action *as wrong*. For example, in the case of one's perceptual knowledge that some object presently before one is a lemon, McGrath claims the following:

> [F]or example, no visual experience ever literally represents something as *a lemon*, although a visual experience might represent something that is in fact a lemon as having a certain shape and being a certain color. Should we conclude from this that no one ever *sees* that there is a lemon on the table? No, we shouldn't. For it might be like this: the features of the scene that you do take in in your visual experience *trigger* or *prompt* you to take up the immediate, non-inferential belief that there is a lemon on the table. The features of the scene that you are responding to in taking up the immediate, non-inferential belief that there is a lemon present need not (and typically will not) amount to anything like a sufficient condition for the presence of a lemon (178).[186]

There are no a priori reasons why we can't extend this reply to the case of (II). Hence, why I have formulated (II) in a broad manner so as not to include the requirement that subjects undergo particularly *normative* introspective states in order to arrive at non-inferentially justified normative beliefs. So, nothing about how I've presented (II) here entails that it is in tension with the idea that (II) must constitute an epistemically independent source of non-inferential justification. But there are two other worries I want to end with. The first I think is rather innocuous and stems from a reason why we might reject something like Memory Intuitionism: it doesn't hook us up to a substantive *mind-independent* normative world. The second is perhaps more troubling. Although (II) is plausible in the case of beliefs about our mental states, it lacks the resources to account for *first-order* normative beliefs that importantly are directed at the

[186] See also Lyons (2018) for a similar claim.

extra-mental world. In other words, how can a *second*-order belief targeting my mental state yield *first*-order knowledge targeting the external world? I'll take each point in turn.

The first worry we might have about (II) is that it doesn't give us access to a mind-independent normative reality. Cowan (2013b), in rejecting the plausibility of Memory Intuitionism, makes a similar point. Even if memory can generate non-inferential justification it runs into a major problem. The main problem is that memory states don't connect us to a mind-independent normative reality. As Cowan claims:

> In this case it seems that merely positing the justification-conferring powers of memory would be insufficient as an account or explanation of *how we have knowledge of the external world*. Memory is the wrong sort of state to posit as hooking us up to a mind-independent external reality in a way that is plausibly required for knowledge. A similar point can be made about the ethical case; merely positing memory as the source of non-inferentially justified belief seems inadequate because it is not a plausible candidate for the sort of thing that would, by itself, connect us to a mind-independent ethical reality. (1107; Cowan's italics).

Given that introspection is similarly *not* connected to the external world, (II) is vulnerable to a similar charge as Memory Intuitionism: we can't go from a memory state to knowledge of the external world. Similarly: we can't go from an introspective state to knowledge of the external world. But why exactly should that non-normative claim support the normative one? I think there are two ways to cash out this connection neither of which is problematic for (II). The first is to claim that memory states *just aren't* hooked up to the mind-independent ethical world just like they aren't hooked up to the mind-independent external world. Pointing to the mind-independent external world is supposed to illuminate the sense in which the same holds true for the mind-independent normative world. But if so, it's unclear why we should believe the normative claim without some reason for thinking that not being appropriately connected to the mind-independent external world is somehow similar/analogous to not being connected to the mind-independent normative world. If we're supposed to see an analogy here, it's unclear what it is.

The second way to cash out the argument is to make an explicit connection between the two sorts of mind-independent realities. But what might that connection be? It's difficult to see what it might be. For starters, we should distinguish between two things:

(A) A mind-independent *external* reality
(B) A mind-independent *normative* reality

Importantly, (B) need not entail (A) as should be clear from the above. That is, if something does not connect us to a mind-independent external reality, it does not follow that it also doesn't connect us to a mind-independent *normative* reality. For (B) can presumably exist in the non-external (mental) world. So, *if* Cowan's argument is something like the following:

P1: Memory cannot connect us to a mind-independent external reality.
P2: If something cannot connect us to a mind-independent external reality, then it cannot connect us to a mind-independent normative reality.
C: Memory cannot connect us to a mind-independent normative reality.

then his argument is unsound and can't be wielded against introspection. Although it seems reasonable to assume that memory cannot connect us to a mind-independent normative reality, it is questionable whether it has anything to do with its connection (or lack thereof) to the mind-independent external world. In any case, the objection that (II) does not constitute a plausible version of (NI) on the grounds that it doesn't connect us to a mind-independent normative reality should be rejected.

Another way of cashing out the objection that (II) does not connect us to a mind-independent normative reality is to claim that the normative truths it connects us to are entirely contingent on mental objects. In other words, the normative reality it connects us to is importantly *mind-dependent*. In order to deal with this potential objection, it's best to introduce more precise terminology. It's best to think of the normative truths with which normative epistemology is concerned as being truths that are importantly *stance-independent* rather than mind-independent, for it may very well be the case that creatures with minds need to exist in order for there to be normative truths like *causing gratuitous pain is morally wrong*. In this case, of course, if there were no creatures with the capacity to feel pain then the above normative claim would be false. But that's not the dependence at issue in metaethics. Importantly, whether causing someone gratuitous pain is morally wrong does not depend on one's attitudes about the matter. Another way of putting it: (II) does not entail that the normative reality it connects us to

is dependent on the *stances* we take about normative matters. And that's the sense of mind-independence that is at issue in normative epistemology and metaphysics. Just because the normative beliefs that (II) most plausibly non-inferentially justifies are directed at our own mental states, does not entail that the truth of those beliefs is dependent on our own normative stances towards the propositions of those beliefs.

This brings me to my final worry regarding (II): whether (II) can manage to provide justification for beliefs that are importantly *first-order* and *external* to the mind. For notice that introspection is importantly a process that produces knowledge about our mental states. So, how might it give us non-inferential justification for extra-mental normative beliefs? Let's stick with our example of a putatively non-inferentially justified normative belief about my own mental state:

(I-judge): "*That* is wrong.

where 'that' refers to my mental state. How could me move from I-judge to the following:

(P-judge): "*That* is wrong.

where 'that' refers to someone else's action and/or intention?

First, I want to point out that whatever we say about how we move from I-judge to P-judge, I-judge constitutes substantial normative knowledge. The importance of (II), if true, should not be understated simply because the knowledge it generates is directed *inward* rather than *outward*. Having normative knowledge about one's own mental states is just as significant as normative knowledge about the singular actions of other people. And although the normative beliefs that (II) justifies are in a sense *second-order* the normative content is importantly *first-order*. But can I-judge provide the epistemic grounds for perceptual or outward-directed normative judgements? Let me sketch three possible answers to that question. The first is to claim that introspective judgements like I-judge above can constitute the *basic* epistemic *normative* grounds for extra-mental first-order normative judgements about other people's actions. Here's C.D. Broad (1944) nicely outlining this view:

> When a deontic judgment is passed by a person on one of his own acts the above criticism [that we do not perceive the intentions of other people] does not hold. In performing an act a person is or may be directly aware of his own intentions. He knows it directly as an act of intended bribery or forgery or debt-paying or whatever it may be, and not merely as a bit of overt behaviour of a certain kind […] We might suppose that he derives his notions of rightness and wrongness from [introspectively] perceiving those characteristics in certain of his own acts by means of moral sensations. Once he has acquired the notions in this way he can proceed to apply them to the acts of other persons although he cannot perceive these and therefore cannot perceive their rightness or wrongness, but can have only conceptual cognition about them. (144).

The idea here is that we introspectively access the deontic properties of our own mental states, e.g., the wrongness of our intention to lie, and then gain substantive normative knowledge that we then apply to other acts we judge to entail the same sorts of mental states. Importantly, we must make a kind of *inference* from the deontic properties of our own mental states to the mental states/actions of other people which will plausibly involve storing the beliefs obtained in my singular deontic judgements about my own mental states. But once we have the relevant beliefs at hand, it shouldn't be a mystery how we then make first-order extra-mental normative judgements about other people's mental states/actions. And notice, that such an inference need not involve any other normative knowledge, at least not in the case of making first-order singular deontic judgements. In other words, (II) can provide the normative bedrock for extra-mental first-order normative judgements.

We can also give a plausible story about how (II) can provide the grounds for justification in more general propositions like *torture is morally wrong/bad*. We introspect some instance of pain (and its badness) and generalize that introspected badness to the case of torture based on what we conceptually know about torture (e.g., that it involves the causing of immense pain). Admittedly, this won't be the only normative knowledge needed in this case, for there will need to be a piece of normative knowledge referring to the badness/wrongness of causing immense pain. But surely, the knowledge that pain states are bad to be in is indispensable to knowledge of the general proposition *that torture is wrong*.

Recall that the one thing that plagued Perceptual Intuitionism (PI) was that it relied on background normative beliefs penetrating perceptual experiences in order to properly represent normative properties and provide the requisite non-inferential justification. In other words,

according to some, the plausibility of (PI) is dependent on the plausibility of ethical perception; and ethical perception is dependent on subjects wielding prior (justified) epistemic states. Well, here's a plausible epistemic prior: introspective states! Normative introspection, and the subsequent beliefs produced from such a process, are good candidates for the background cognitively penetrating states required for the plausibility of ethical perception. And if that's true, and ethical perception can non-inferentially justify first-order normative beliefs, then the knowledge produced by (II) can function as the basic, fundamental normative belief required for our perceptual experiences to represent normative properties. Here's a fanciful way of putting it: we in essence *normatively project* onto the world the normative properties accessed via our introspective experiences.[187] How might such a process work? Again, this is all rough and ready, but here's a sketch. Take the case of unpleasant pain. Stipulate that unpleasant pain is bad-for-you. If (II) is true, then the badness-for-you of your unpleasant pain is as good a candidate as any for introspectively grounded normative knowledge. Say you now hold the belief that *my unpleasant pain is bad*. Say you experience enough unpleasant pains and come to believe that *unpleasant pain is bad*. Now, your belief *that unpleasant pain is bad* cognitively penetrates your perceptual experiences of cases of people seeming to experience pain. Your perceptual experience now represents the state some person is in (e.g., they've just bashed their knee against the pavement) *as bad* because of the background normative belief *that unpleasant pain is bad*. You've come to perceptually represent that person's condition accurately *as bad*, and then come to non-inferentially believe *that is bad*. Note, for what it's worth, that the idea that our experiences of our own pain (and its badness) cognitively penetrate our perceptual experiences is pre-theoretically plausible. But the explanation going in the other direction is odd: that our perceptions of other people's pain cognitively penetrate our own experiences of our own pain. So, here's one reason to think that introspective states are epistemically prior to perceptual ones.

The third and final way in which introspection might provide the epistemic grounds for first-order extra-mental beliefs is to appeal to the view outlined above regarding *affordances*. If affordances are indeed introspectable, then they can plausibly provide the grounds for the non-inferential justification for first-order extra mental normative beliefs, in particular, about which actions are or are not to-be-done or to-be-ϕ-ed.

[187] This way of talking shouldn't entail that the view is somehow non-cognitivist since projectivism can take many forms, even cognitivist and realist forms. See Joyce (2009) for discussion.

Conclusion

In closing, there are many ways to make plausible the idea that (II) constitutes an epistemically independent source of non-inferential justification for normative beliefs. As well, I've outlined some further attractions of the view, in particular, that it can constitute a basic epistemic prior with respect to other states, e.g., perceptual experience. I've also gestured at a way in which we might construct our normative knowledge of the world out of the basic normative knowledge we have of our own minds. But the extent to which that story is plausible demands more attention. I hope to have motivated taking seriously the idea that normative introspection and with it (II) ought to be taken seriously.

Conclusion

Let me end the thesis by briefly restating the conclusions of each chapter and giving some concluding thoughts.

We've seen that, in Chapter one, if we restrict our picture of distinctive self-knowledge to non-inferential self-knowledge, then something taken to be fairly intuitive as a prime candidate of distinctive self-knowledge is excluded: motivating reasons. The main reason for this is because they are causal. Keeling attempted to give an argument against what she called 'the orthodox position'. Her argument rested on the idea that were we to engage in an inference when pressed for our motivating reasons—i.e., *why* we believe some proposition or act a particular way—we would be failing to respect the dual-role of the question 'why?'; that is, we would fail to provide a justification for the lower-order attitude or the action. But we saw a number of things wrong with that argument. First, it was not obvious how exactly inference failed to provide such a justification. We saw that any apparent connection between inference and purely causal answers (which would necessarily exclude providing justification) needed further argument. It also became clear that if inference is epistemically problematic in the psychological sense—i.e., in the sense that one explicitly engages in an inference—then the belief one forms with respect to their motivating reason must not itself be *epistemically mediate*. The key thought there was that were something to be epistemically mediate in the sense that it relies on the justification of background beliefs for its own justification—e.g., think of one's belief that there is a proton—then there should not be anything epistemically wrong with making that justificatory support explicit. In other words, there shouldn't be anything wrong with making that inferential support explicit. We also saw that giving a "non-inferential" answer might nonetheless be epistemically dependent. So, it's unclear how linguistic conventions with respect to answering why-questions has any rendering on the self-knowledge of motivating reasons.

We then saw a different interpretation of Keeling's argument. There we saw that when subjects engage in inference, they cannot *endorse* the content of a normative judgement; that is, they cannot take some proposition to count in favour of believing some other proposition. This seemed to be the best way to interpret Keeling's argument. However, there were problems with this argument as well. It was unclear why subjects could not come to non-inferentially know

their normative judgement and then infer from this knowledge that it was their motivating reason on the basis of a further background belief about their normative judgements becoming their motivating reasons. I ended that section by claiming that there is nothing problematic—or even *alienating*—about making the sort of normative endorsement required for something to be one's motivating reason while wondering whether such an endorsement plays the proper causal role to actually be one's motivating reason.

I then moved on to consider Keeling's non-inferential account of our access to our motivating reasons: RTM. Roughly, RTM gives us direct non-inferential self-knowledge of our motivating reasons because when we make a normative judgement—e.g., that *p* is a good reason for believing *q*—subjects partly make it the case that *p* is their motivating reason. In short, they become agent aware of *p* as their motivating reason—have an agentive experience of *p* as their motivating reason—and therefore have non-inferential justification for the belief that *p* is their motivating reason. But there were problems with this too. First, it was unclear whether the required background beliefs which penetrate the subject's experience do not confer justification *to* the experience itself making the subsequent motivating reason belief epistemically dependent in a problematic way. Second, we then say that it wasn't at all clear that subjects have non-inferential justification for believing that a causal relation obtains. We concluded that the prospects for giving a non-inferential account of our distinctive access to our motivating reasons needs further support. And anyways, an inferentialist account of our knowledge of our motivating reasons remains highly plausible.

Diagnosing why our intuitions are perhaps misled here is easy. Although we enjoy distinctive self-knowledge of—i.e., direct non-inferential access to—our *judgements* about what we take to be good reasons, we nonetheless do not enjoy any sort of direct non-inferential access to whether those judgements *caused* a relevant belief or action. We tend to simply take our normative judgements at face value: they just are our motivating reasons. But if we want knowledge of that further causal fact, then, it would seem, we will need some further justification about the causal efficacy of our normative judgements. And this, importantly, will involve *inference*, what we were assuming here to be *too alike* our knowledge of other people's mental states to count as *distinctive* self-knowledge. Recall:

1. Jack knows that he weighs 77 kilograms.

2. Sally knows that she has a headache.

So, self-knowledge of our motivating reasons will more closely resemble Jack's knowledge of his weight than it will Sally's knowledge of her headache. Why it might *seem* like we have self-knowledge of our motivating reasons more akin to (2) than to (1) is because like Jack we stand in some privileged relation to a set of information or evidence which we exploit for our epistemic purposes. In Jack's case, it's his proximity to weight-scales, and in the motivating reasons case it's our proximity to *our normative judgements*. And there is one crucial difference between self-knowledge of motivating reasons and Jack's knowledge of his weight: self-knowledge of motivating reasons does seem to entail *some* direct non-inferential access to our mental states, in particular, our judgements about what is a good reason for something. And this fact, I think, puts self-knowledge of motivating reasons somewhere in between (1) and (2) on the scale of distinctive self-knowledge. But importantly, so long as we are restricting distinctive self-knowledge to direct non-inferential access, then motivating reasons fail to fall under the scope and reach of distinctive self-knowledge.

In chapter two, I moved on to consider the various ways an indirect, inferential account of our access to our phenomenal states interacted with some of the motivational and normative constraints typically put on theories of phenomenal consciousness. There we saw that strong representationalists—who are committed to the transparency of experience, and hence an indirect, inferential account of introspection—fail to accommodate the motivational and normative features of particular phenomenal episodes like unpleasant pain. In particular, I argued that *if* strong representationalists want to accommodate those motivational and normative features by appealing to desires, then they will need to invoke what I called *de re* desires: desires which are importantly *object-directed*. Now, I further claimed that for those desires to be relevantly object-directed subjects would need to directly attend to their phenomenal episodes. But that is precisely what is precluded on strong representationalism. So, strong representationalists are barred from wielding a highly attractive naturalistic explanation of motivation and normativity. I then further claimed that strong representationalists are still in trouble with respect to other, non-attitudinal explanations of the motivationality and normativity of unpleasantness. First, they cannot account for how subjects could *access* the badness of their unpleasant pain, and hence rationally respond to it without either having the relevant motivation-constituting desire or without having the right second-order motivational experience. Second, I

argued that one recent attempt to accommodate the normativity of unpleasantness by appeal to perceptuality is unavailable to the strong representationalist. They must rely, implausibly, on content alone to do such explaining. I suggested we abandon strong representationalism.

In Chapter three, I considered the prospects for giving a novel a posteriori ethical intuitionism: Introspection Intuitionism. I did three things there. First, I motivated the idea purely private mental states do have normative properties. Second, I motivated the idea that we can introspect such normative properties. I did this by explaining some intuitive phenomenological data by appeal to introspection and showing how that explanation does at least as good as rival explanations. That gives us good reason to take seriously the idea of normative introspection. Next, I showed how we might even appeal to normative introspection to provide the non-inferential justification of first-order beliefs given that purely private mental states never have normative properties. Here I appealed to the affordance literature and claimed that mental states can afford certain actions, and it is our introspective access to such affordance properties which provides the non-inferential justification for normative beliefs about actions. I then went on to connect normative introspection to introspection intuitionism (II). I defended (II) against the idea that it might not be an epistemically independent source of non-inferential justification. I also sketched a picture with respect to how (II) might further provide the epistemic grounds for non-inferential justification in first-order normative beliefs about the extra-mental world.

References

Alvarez, M. (2017). Reasons for action: Justification, motivation, explanation. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Winter 2017 ed.). Retrieved from https://plato.stanford.edu/archives/win2017/entries/reasons-just-vs-expl/

Alvarez, M. (2018). Reasons for action, acting for reasons, and rationality. *Synthese 195*, 3293–3310.

Alvarez, M. (2008). Reasons and the ambiguity of 'belief'. *Philosophical Explorations: An International Journal for the Philosophy of Mind and Action*, *11:1*, 53-65.

Anscombe, G. E. M. (2000). *Intention* (2nd. ed.). Cambridge, MA; London: Harvard University Press

Armstrong, D. M. (1962). Bodily sensations. London: Routledge and Kegan Paul.

Armstrong, D. M. (1968). *A Materialist Theory of the Mind*. New York, Humanities Press.

Aydede, M. (2014). 'How to Unify Theories of Sensory Pleasure: An Adverbialist Proposal'. *Review of Philosophy and Psychology*, *5*(1), 119–33.

Aydede, M. (2019b). 'Is the experience of pain transparent? Introspecting phenomenal qualities'. *Synthese*, (196) 677-708

Aydede, M. and Fulkerson, M. (2018). 'Reasons and Theories of Sensory Affect'. In David Bain, Michael Brady and Jennifer Corns (Eds.), *The Nature of Pain*. Oxford, UK: Oxford University Press.

Bain, D. (2013). 'What Makes Pains Unpleasant?' *Philosophical Studies*, *166*(1 Supp), 69– 89.

Bain, D. (2017). 'Why Take Pain Killers?' *NOUS,* 53:2, 462–490.

Bain, D. Pain and Action (manuscript).

Basu, R. (2018). Can beliefs wrong? Philosophical Topics, 46(1), 1–18.

Basu, R. (2019a). What we epistemically owe to each other. Philosophical Studies, 176(4), 915–931.

Basu, R. (2019b). The wrongs of racist beliefs. Philosophical Studies, 176(9), 2497–2515.

Basu, R., & Schroeder, M. (2019). Doxastic wronging. In B. Kim & M. McGrath (Eds.), Pragmatic encroachment in epistemology (pp. 181–205). New York: Routledge.

Bedke, M.S. (2010), 'Intuitional Epistemology in Ethics', Philosophy Compass 5 (12): 1069-1083

Bedke, M. (2011). Passing the deontic buck. *Oxford Studies in Metaethics*, **6**, 128–153.

Bergqvist, A., & Cowan, R. (2018). Evaluative perception. Oxford University Press.

Brandt, R. (1979). A theory of the good and the right. Ithaca: Cornell University Press.

Boswell, P. (2016). 'Making Sense of Unpleasantness: Evaluationism and Shooting the Messenger,' Philosophical Studies 173(11), pp. 2969–2992

Boyle, M. (2009). Two kinds of self-knowledge. *Philosophy and Phenomenological Research*, *78*, 133–164. Retrieved from https://doi.org/10.1111/phpr.2008.78.issue-1.

Boyle, M. (2011a). 'Making up your mind' and the activity of reason. *Philosophers' Imprint*, *11*, 1–24.

Boyle, M. (2011b). Transparent self-knowledge. *Proceedings of the Aristotelian Society Supplementary Volume*, *85*, 223–241.

Boyle, M. (2019). Transparency and reflection. *Canadian Journal of Philosophy*, 1–28.

Brady, M. (2015). 'Feeling Bad and Seeing Bad'. *Dialectica* 69(3): 403–416.

Brady, M. (2018b). *Suffering and Virtue*. Oxford: Oxford University Press.

Bradford, G. (2020). The badness of pain. *Utilitas, 32*(2), 236-252. doi:10.1017/S0953820819000475

Broad, C. D. (1944). Some Reflections on Moral-Sense Theories in Ethics. In *New Series* (Vol. 45). https://www.jstor.org/stable/4544400

Brewer, Talbot. (2000). The Bounds of Choice: Unchosen Virtues, Unchosen Commitments, New York: Garland Publishing.

Cassam, Q. (2014). *Self-knowledge for humans*. Oxford: Oxford University Press.

Chalmers, David J. (2003). "The Content and Epistemology of Phenomenal Belief." In *Consciousness: New Philosophical Perspectives*, edited by Q. Smith and A. Jokic. Oxford: Oxford University Press.

Chudnoff, E. 2011. What intuitions are like. *Philosophy and Phenomenological Research* 82(3): 625–654.

Coates, D.J. (2023). A Defense of Weak Moralism: A Reply to Sher. *J Ethics* 27, 131–140. https://doi.org/10.1007/s10892-023-09422-z

Cowan, R. (2013a). Perceptual intuitionism. *Philosophy and Phenomenological Research*. doi:10.1111/phpr.12023

Cowan, R. (2013b). Clarifying ethical intuitionism. European Journal of Philosophy. doi:10.1111/ejop.12031

Cowan, R. (2015). Cognitive Penetrability and Ethical Perception. *Rev.Phil.Psych.* 6, 665–682. https://doi.org/10.1007/s13164-014-0185-4

Cowan, R. (2016). "Epistemic perceptualism and neo-sentimentalist objections". *Canadian Journal of Philosophy*, 46 (1), 59-81.

Cowan, R. (2017). "Rossian Conceptual Intuitionism". *Ethics*, 127 (4), 821-851.

Chudnoff, E. (201)3. *Intuition*. Oxford: Oxford University Press.

Cox, Damien and Michael Levine. (2004). Believing Badly. *Philosophical Papers* 33/3: 309–28.

Crisp, R. (2018). Prudential and Moral Reasons. In: Daniel Star (ed), *The Oxford Handbook of Reasons and Normativity*, (pp. 801-820). Oxford University Press.

Cullison, A. 2010, 'Moral Perception', European Journal of Philosophy 18 (2): 159–75.

Cutter, B. and M. Tye. 2011. 'Tracking Representationalism and the Painfulness of Pain'. *Philosophical Issues* 21: 90–109.

Cutter, B. and M. Tye. (2014). 'Pains and Reasons: Why it Is Rational to Kill the Messenger'. *Philosophical Quarterly* 64(256): 423–433.

Dancy, J. (2000). Should we pass the buck? *Philosophy: The Journal of the Royal Institute of Philosophy,* 47 (Supplement), 159–73.

D'Arms, Justin and Daniel Jacobson, 2000. The Moralistic Fallacy: On the 'Appropriateness' of Emotions, Philosophy and Phenomenological Research 61/1: 65–90.

Davidson, D. (1963). Actions, reasons, and causes. *The Journal of Philosophy*, *60*, 685–700.

Director, S. (2022). The sheriff in our minds: on the morality of the mental. *Journal of Ethics and Social Philosophy*, Vol. 22 (3), 423-437.

Dretske, F. (1995). *Naturalizing the Mind*. Cambridge, Massachusetts: MIT Press.

Dretske, F. (1999). The Mind's Awareness of Itself. *Philosophical Studies*, *95*(1–2), 103–124.

Dretske, F. (2003). 'How Do You Know You Are Not a Zombie?'. In: B. Gertler (ed.), *Privileged Access: Philosophical Accounts of Self-knowledge*, Aldershot: Ashgate Publishing Limited.

Driver, J. (2001). Uneasy Virtue. Cambridge: Cambridge Studies in Philosophy.

Feldman, Fred. 2018. Unconscious Pleasures and Pains: A Problem for Attitudinal Theories? Utilitas 30.4: 472–82.

Fletcher, G. (2018). Pain for the Moral Error Theory? A New Companions-in-Guilt Argument. *Australasian Journal of Philosophy*, 96(3), 474-482.

Foot, P. (2004). Rationality and Goodness" in *Modern Moral Philosophy*, (*Royal Institute of Philosophy Supplement 54*), Anthony O'Hear (ed.), Cambridge: Cambridge University Press, 2004: 1–14. doi:10.1017/CBO9780511550836.002

Gertler, B. (2011). *Self-knowledge*. London: Routledge.

Gertler, Brie. (2001). "Introspecting Phenomenal States." *Philosophy and Phenomenological Research* 63: 305–328.

Gertler, Brie. (2012). “Renewed Acquaintance”. In Declan Smithies and Daniel Stoljiar (eds), *Introspection and Consciousness*, Oxford University Press.

Giustina, A. (2022). Introspective knowledge by acquaintance. *Synthese* 200, 128 https://doi.org/10.1007/s11229-022-03578-1

Giustina, A., & Kriegel, U. (2017). ‘Fact-introspection, thing-introspection, and inner awareness’. *Review of Philosophy and Psychology*, 8(1), 143–164.

Goldman, Alvin. (2006). *Simulating Minds*. New York: Oxford University Press.

Grahek, N. (2007). *Feeling pain and being in pain* (2nd ed.). Cambridge, MA: MIT Press.

Hall, R. J. (1989). Are pains necessarily unpleasant? *Philosophy and Phenomenological Research,* XLIX(4), 643–659.

Harman, G. 1977. The nature of morality: an introduction to ethics. New York: Oxford University Press.

Harman, G. (1990). The Intrinsic Quality of Experience. *Philosophical Perspectives*, 4: 31–52.

Hazlett, Alan. (2009). How to Defend Response Moralism, The British Journal of Aesthetics 49/3: 241–55.

Heathwood, C. (2007). The reduction of sensory pleasure to desire. *Philosophical Studies* 133: 23–44.

Helm, B. (2002). Felt evaluations: A theory of pleasure and pain. American Philosophical Quarterly, 39(1), 13–30.

Horgan, T. (2012). “From Agentive Phenomenology to Cognitive Phenomenology: A Guide for the Perplexed.” In *Cognitive Phenomenology*, edited by T. Bayne and M. Montague. Oxford: Oxford University Press.

Horgan, T., and U. Kriegel. (2007). “Phenomenal Epistemology: What Is Phenomenal Consciousness That We May Know It So Well?” *Philosophical Issues* 17: 123–144.

Huemer, M. (2005), *Ethical Intuitionism*, Basingstoke: Palgrave Macmillan

Jacobson, H. (2019a). ‘Not Only a Messenger: Towards an Attitudinal-Representational Theory of Pain’. *Philosophy and Phenomenological Research.* Vol. XCIX No. 2,: 382-408.

Joyce, R. (2009). Is Moral Projectivism Empirically Tractable? *Ethical Theory and Moral Practice*, *12*(1), 53–75. http://www.jstor.org/stable/40284272

Loar, B. (1997). Phenomenal States. In *The Nature of Consciousness*, edited by N. Block, O. Flanagan, and G. Güzeldere. Cambridge, Mass.: MIT Press.

Lycan, William G. (1996). *Consciousness and Experience*. Cambridge, Mass.: MIT Press.

Lyons, J. (2018). Perception and intuition of evaluative properties. In R. Cowan, & A. Bergqvist (Eds.), *Evaluative perception*. Oxford University Press.

Kant, Immanuel. [1781, 1787] 1997. *Critique of Pure Reason*. Edited and translated by P. Guyer and A.W.Wood. Cambridge: Cambridge University Press.

Kant, I. (1997). Groundwork of the Metaphysics of Morals. (trans.) (ed.) Gregor, M. United Kingdom: Cambridge University Press.

Keeling, S. (2019a). The transparency method and knowing our reasons. *Analysis*. Retrieved from https://doi.org/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*, 00, 1–24.

Kind, A. (2007). 'Restrictions on Representationalism'. *Philosophical Studies* 134 (3):405-427.

Kriegel, U. (2007). "Intentional Inexistence and Phenomenal Intentionality." Philosophical Perspectives 21 (1): 307–40.

Klein, C. (2007). An imperative theory of pains. Journal of Philosophy, 104(10), 517–532.

Kriegel, U. (2009). Subjective Consciousness: A Self-Representational Theory. Oxford: Oxford University Press

Macpherson, F. (2014). Is the sense-data theory a representationalist theory? *Ratio*, *27*(4), 369–392.

Mandelbaum, M. (1955). The Phenomenology of Moral Experience. Glencoe, IL: The Free Press.

Martinez, M. (2011). Imperative content and the painfulness of pain. Phenomenology and the Cognitive Sciences, 10, 67–90

McBrayer, J.P.(2010a). 'A Limited Defense of Moral Perception', Philosophical Studies 149 (3): 305–20.

McBrayer, J.P. 2010b, 'Moral Perception and the Causal Objection', Ratio 23 (3): 291–307.

McGrath, S. (2004). Moral knowledge by perception. Philosophical Perspectives, 18(1), 209–228.

McGrath, S. (2011). Moral knowledge and experience. In R. Shafer-Landau (Ed.), Oxford studies in metaethics, volume 6. Oxford University Press.

McGrath, S. (2018). Moral perception and its rivals. In R. Cowan, & A. Bergqvist (Eds.), Evaluative perception. Oxford University Press.

Mill, J. S. (1979 [1861]). *Utilitarianism*. Indianapolis: Hackett Pub. Co.

Mitchell, J. (2023). Experiencing Mandates: Towards A Hybrid Account, *Australasian Journal of Philosophy*, 101:2, 267-281, DOI: 10.1080/00048402.2021.1995013

Moran, R. (2001). *Authority and estrangement: An essay on self-knowledge*. Princeton; Oxford: Princeton University Press.

Müller, J.M. (2021). Perceptualism and the epistemology of normative reasons. *Synthese* 199, 3557–3586. https://doi.org/10.1007/s11229-020-02947-y

Nichols, S and Stephen Stich. (2003). *Mindreading*. Oxford: Oxford University Press.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231–259. Retrieved from https://doi.org/10.1037/0033-295X.84.3.231.

Noë, A. (2006) *Action in Perception*. Cambridge: MIT Press.

Parfit, D. (1984). Reasons and persons. Oxford: Oxford University Press.

Papineau, David. (2007). "Phenomenal and Perceptual Concepts." In *Phenomenal Concepts and Phenomenal Knowledge*, edited by Torin Alter and Sven Walter. Oxford: Oxford University Press.

Pitcher, G. (1970a). The awfulness of pain. Journal of Philosophy, 68, 481–492.

Pitcher, G. (1970b). Pain perception. *Philosophical Review*, 79, 368–393.

Pryor, J. (2000). The Sceptic and the Dogmatist'. *Nous*, 34, pp. 517-49.

Rey, G (2008). (Even Higher-Order) Intentionality Without Consciousness. *Revue Internationale de Philosophie* 1:51-78.

Rapstine, M. Regrettable beliefs. *Philos Stud* **178**, 2169–2190 (2021). https://doi.org/10.1007/s11098-020-01535-7

Setiya, K. (2013). Epistemic agency: Some doubts. *Philosophical Issues*, *23*, 179–198.

Schroeder, M. (2018). When beliefs wrong. Philosophical Topics, 46(1), 115–127.

Siegel, S. (2006), 'What Properties Are Represented In Perception?' in Gendler, T.S. & Hawthorn J., (eds.), Perceptual Experience, Oxford University Press.

Siegel, S. (2014). Affordances and the Contents of Perception, in Does Perception have Content?, ed. Berit Brogaard, New York: Oxford University Press: 51–76.

Sher, G. (2019). A Wild West of the Mind, Australasian Journal of Philosophy, 97:3, 483-496, DOI: 10.1080/00048402.2018.1490326

Shoemaker, Sydney. 1996. The First-Person Perspective and Other Essays. Cambridge: Cambridge University Press.

Smith, Angela. (2011). Guilty Thoughts, in Morality and the Emotions, ed. Carla Bagnoli, New York: Oxford University Press: 235–56.

Smith, M. (1994). The moral problem. Oxford: Basis Blackwell.

Smithies, D and Daniel Stoljiar. *Introspection and Consciousness*, Oxford University Press.

Sosa, E. (2012). "The epistemology of introspection". In Declan Smithies and Daniel Stoljiar (eds), *Introspection and Consciousness*, Oxford University Press.

Sturgeon, N. (2002). 'Ethical Intuitionism and Ethical Naturalism', in P. Stratton-Lake (ed.), *Ethical Intuitionism: Re-evaluations* (Oxford: Clarendon Press), pp. 184-21 1.

Stratton-Lake, P. (2002). *Ethical intuitionism: Re-evaluations*. Oxford; New York: Clarendon Press; Oxford University Press.

Sylvan, K. (2016a). Epistemic reasons I: Normativity. *Philosophy Compass*, *11*, 364–376.

Sylvan, K. (2016b). Epistemic reasons II: Basing. *Philosophy Compass*, *11*, 377–389.

Tappolet, Christine. (2013). "Evaluative vs. Deontic Concepts," in *International Encyclopedia of Ethics.*

Tappolet, Christine. (2014). "The Normativity of Evaluative Concepts," in *Mind, Values, and Metaphysics: Philosophical Essays in Honor of Kevin Mulligan, Volume 2.*

Tye, M. (1995a). A representational theory of pains and their phenomenal character. *Philosophical Perspectives* 9: AI, Connectionism, and Philosophical Psychology, 223–239.

Tye, M. (1995b). Ten problems of consciousness. Cambridge, MA: MIT Press.

Tye, M. (1996a). The function of consciousness. *Noûs*, *1*(3), 287–305.

Tye, M. (2000). *Consciousness, Color, and Content*. Cambridge, Massachusetts: MIT Press.

Tye, M. (2003). Representationalism and the Transparency of Experience. In: Brie Gertler (Ed.), Privileged Access: Philosophical Accounts of Self-Knowledge (pp.XXXX). Burlington, Vermont: Ashgate Publishing (Epistemology and Mind Series).

Tye, M. (2006a). Another Look at Representationalism about Pain. In M. Aydede (Ed.), *Pain: New Essays on Its Nature and the Methodology of Its Study* (pp. 99–120). Cambridge, Massachusetts: MIT Press.

Tye, M. (2014a). 'Transparency, qualia realism, and representationalism'. *Philosophical Studies*. 170: 39-57.

Vayrynen, P. (2008). 'Some Good and Bad News For Ethical Intuitionism', The Philosophical Quarterly 58: 489–511.

Vayrynen, P. (2018). “Doubts about moral perception”. In R. Cowan, & A. Bergqvist (Eds.), *Evaluative perception*. Oxford University Press.

Ventham, E. (2021). ‘Attitudinal theories of pleasure and *de re* desires’. *Utilitas*, *33*(3), 361–369.

Werner, P. (2019). Which Moral Properties are Eligible for Perceptual Awareness? *Journal of Moral Philosophy, 17(3)*, 1-30. doi:10.1163/17455243-20182801

Zheng, R. (2016). Why Yellow Fever Isn’t Flattering: A Case Against Racial Fetishes. *Journal of the American Philosophical Association*, *Vol 2(3)*, 400-419.