Evolutionary Naturalism & the Normativity of the Mental

By

Joshua Scott Beattie

A dissertation submitted in partial satisfaction of the requirements for the degree of

Doctor of Philosophy

in

Philosophy

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor John Campbell, Co-Chair
Professor John Searle, Co-Chair
Professor Alison Gopnik

Fall 2012
Abstract

Evolutionary Naturalism and the Normativity of the Mental

by

Joshua Scott Beattie

Doctor of Philosophy in Philosophy
University of California, Berkeley
Professor John Campbell, Co-Chair
Professor John Searle, Co-Chair

I examine the following plausible and oft-defended line of thought:

(1) Rationality is a normative phenomenon.
(2) Facts about intentionality essentially depend on facts about rationality, so intentionality is a normative phenomenon too.
(3) Normative phenomena resist naturalistic treatment.
So, (4) the normative aspect of intentionality precludes a fully naturalistic account of the mind.

I believe the conclusion is mistaken, and I attempt to show this by means of a clarification and assessment of claims (2) and (3).

Unlike many of those who resist the above line of thought, I think (2) can be adequately defended. Its defense depends, in my view, on sufficiently appreciating the dynamic character of intentionality. That is, as worldly conditions and conceptual repertoires change, so too does the intricate causal network in which intentional states figure. A purely causal-dispositionalist theory, however, is unable to say anything about these changes and why they take place; it must simply be updated to take account of them. It is here that the essential role of rationality considerations comes in: the development of an intentional state’s causal profile tracks the development of its rational profile. In other words, the dispositionalist would have to await a verdict, so to speak, on a given connection’s rational status before knowing whether to include it in a state’s defining causal repertoire. (An analogy: a purely dispositionalist approach is akin to defining a biological species in terms of its current structural features, when a species is in fact a dynamic entity whose characteristic structural features change over time).
If (2) is defensible, then the problem must lie in claim (3). That said, I do not attempt to overturn (3) by offering, as others have, a naturalistic account of normativity or rationality as such (e.g. expressivism); in my view, the prospects for a naturalistic account of intentionality are independent of this more general issue. I focus, instead, on two key ways in which (3) is underspecified: first, it needs to be said precisely what is required for a “naturalistic” treatment of some phenomenon, and second, it needs to be said precisely how normativity comes into play in a given instance. Once these details are properly filled in, I claim, (3) is false for the particular case of intentionality.

On the first point, I argue that the standards for naturalistic explanation must be less demanding than full reductionism but more demanding than the “naïve” naturalism of many anti-reductionists. I claim that evolutionary continuity provides the proper naturalistic measure: a phenomenon must be shown to have arisen through a seamless course of evolutionarily explicable transitions. This yields a robust explanatory requirement, but one that is flexible enough to handle cases to which reductionism is ill-suited (like intentionality). Moreover, this brand of naturalism quite clearly calls for an interdisciplinary approach, as I believe naturalism should. In this case, that approach depends on continuing investigation into the evolution of cognition and the brain.

As for the second point, I think the tendency to give (2) a strongly metaphysical reading must be resisted. That is, intentional states should be seen not as fundamentally normative entities (as when beliefs are equated with states of commitment), but as causal entities that are essentially approximations to a dynamic normative ideal. The involvement of that ideal is enough to rule out straightforward reductions – the failure of dispositionalism has already been mentioned, and similar considerations can be used to support a broader anti-reductionist thesis – but the more modest naturalistic demand of evolutionary continuity can still be met. This requires an account of our capacity to construct and utilize normative models in our thinking. If it can be shown how we acquired the ability to run through various courses of reasoning and behavior off-line, and to have the results of that feed back and modify our cognitive dispositions, then we can make sense of our having intentional states, i.e. dynamic causal states that approximate to and are in some sense guided by a normative ideal. I cannot claim to tell the whole naturalistic story, but I hope to have made the project look tractable.
Contents

Acknowledgements ii

Chapter 1 Establishing the Normativity Thesis: Vertical Norms 1

Chapter 2 Establishing the Normativity Thesis: Horizontal Norms 22

Chapter 3 The Dynamics of Intentionality & The Argument Against Naturalism 48

Chapter 4 Evolutionary Naturalism 67

References 85

Appendix Assessment-Sensitivity & the Naturalistic Fallacy 90
Acknowledgements

I am grateful to the chairs of my dissertation committee, John Campbell and John Searle, for being such interesting philosophers, insightful critics, and friendly people. I appreciate their support of my work over many years.

Thank you also to the numerous others who made the U.C. Berkeley Philosophy Department such a wonderful place to spend those many years.

And finally, immense thanks to my family – you were with me, as always, on this, and I hope you will be with me on all things to come.
Chapter 1

Establishing the Normativity Thesis: Vertical Norms

1. Introduction

It is commonly maintained that intentionality is normative in a fundamental sense, and that, as a result, no naturalistic account of the mind is possible.¹ So in addition to all the specific objections that can be leveled against extant naturalistic theories like functionalism, informational semantics, teleosemantics, etc. – theories that are causal/dispositional in one way or another – there is, according to this view, a more incisive and far-reaching claim that establishes their inadequacy. Rather than pointing out fairly technical problems with specific naturalistic theories – e.g. using counterexamples to show that a theory doesn’t capture the right extensions for our terms/concepts,² or arguing that the view is implicitly circular/question-begging³ – this normativity-based objection offers a general diagnosis for why such technical problems are bound to arise. Because of this, it issues in a firm negative prognosis as well: it says that no naturalistic theory will succeed in providing an adequate account of the mind.

This position clearly has serious consequences, but its central claims are unfortunately somewhat obscure. There are two such claims: (1) that certain intentional mental phenomena are normative, and (2) that this fact constitutes a decisive objection to naturalism in the philosophy of mind. My focus in the next

¹ Prominent examples include Bilgrami (2004), Brandom (1994), Kripke (1982), McDowell (1994), Millar (2004), and Zangwill (1998). This point is sometimes made when the focus is conceptual content, sometimes when it is propositional attitude individuation, but I think it constitutes one common issue.
² See, for example, Godfrey Smith (1989) and Boghossian (1991), among many others.
³ See, for example, Loewer (1997) and Baker (1991), among many others.
two chapters will be on the first claim – call it the Normativity Thesis – and my
aim is to zero in on its most defensible interpretation. The second claim – call it
the Normativity Objection – will play a role initially only in the following way: a
claim must present at least *prima facie* grounds for the Normativity Objection if it
is to count as a possible interpretation of the Normativity Thesis. (This is meant to
ensure that the thesis is treated as a substantive claim and not a trivial, terminological one). Once the Normativity Thesis has been clarified, I will assess
the Normativity Objection more fully in chapters 3 and 4.

A good place to begin is to ask what *kind* of norm is fundamental to
mentality, according to the Normativity Thesis. We can usefully distinguish two
types of norm that are candidates.\(^4\) The first can be thought of as *atomistic* or
*vertical* norms – when discussion centers on satisfaction-conditions (including
truth-conditions) as a central feature of intentional states, it is this first type of
norm that is at issue. These are norms that, to put it crudely, could apply in
principle to a mental state even if it were the only intentional item in existence. In
other words, these norms do not concern the relations between a given mental state
and other such states, or between a given mental state and intentional actions
performed on the basis of it.\(^5\) What they concern is just the relation between the
mental state and the worldly state of affairs it represents (a relation the word
“vertical” is meant to capture, though it does so only imperfectly). If the right
kind of relation obtains, the state has a positive normative status – it is correct, or
fulfilled, or satisfied – while if that relation does not obtain, it has a negative
normative status – it is incorrect, or unfulfilled, or unsatisfied.\(^6\) The earlier crude
claim can now be made slightly more precise: a vertical norm is one whose
assessment depends on a state’s own intentional content but not that of any other
state or action.

Vertical norms are arguably the kind more often implicated in claims of the
Normativity Thesis. They are front and center, for instance, in discussion of the

\(^4\) Similar distinctions can be found in Zangwill (1998, 2005) and Brandom (1994).

\(^5\) This way of putting things is too crude because of (at least) intentions – whether an intention is satisfied
or not depends directly on what actions are performed on the basis of it. (This also in part accounts for the
inadequacy of the term ‘vertical’ mentioned in the next sentence). The intended contrast is between this
kind of normative connection between a mental state and an action (in which the latter constitutes
satisfaction of the former) and the kind in which a mental state or cluster of mental states makes an action
reasonable or unreasonable, or otherwise tells for or against it. The former connection is, more or less,
Wittgenstein’s (1953) ‘internal relation’ between intention and its fulfillment, while the latter is ‘external’
at least in the following sense: we can be ignorant about which actions are sanctioned in this way, and can
learn about such sanctions by means of empirical inquiry.

\(^6\) Just what relation is required depends on the kind of state and its particular content. Mere co-occurrence
of a state and its represented state of affairs (or occurrence at appropriately related times) is enough to
confer positive status on beliefs and desires, while perceptions and intentional actions plausibly require a
further causal connection between the two things.
theories of mental content due to Fodor, Dretske, Millikan, and others. In this context, “the normativity problem” is often used\(^7\) as a name for what is also called “the misrepresentation problem” or “the disjunction problem.” Regardless of the label, the issue here is whether naturalistic accounts can manage to individuate states capable of having both of the two sorts of normative status described above. In other words, the question is whether such accounts can establish that a state has a certain content that might but also might not match the world. Failing to establish this, it is often thought, would be failing to capture an essentially normative aspect of mental content. And since that essentially normative aspect depends on no more than a given state and the bit of the world it purports to represent, this is an example of the Normativity Thesis based on vertical norms.

In this chapter, I will argue, largely following Dretske (2000) and Rosen (2001), that this common way of rendering the Normativity Thesis is mistaken. Many important mental states do have satisfaction-conditions, of course, but that is not best thought of as a distinctly normative feature of them. And I do not deny that naturalistic accounts of content must meet the kind of challenge just outlined; it is just that doing so is not a matter of capturing something fundamentally normative about content.\(^8\) This means that someone hoping that normativity could provide a general (as opposed to case-by-case) reason to reject naturalism will have to look elsewhere.

Dretske and Rosen both conclude their assessment of the Normativity Thesis after considering this vertical construal of it, but there is another gloss the thesis might be given. One might take it to be a claim about the second type of norm instead, which can be thought of as relational or horizontal in character. Horizontal norms relate a mental state either to other mental states, as when a belief justifies or provides reason for another belief, or to actions, as when a desire or intention commits one to doing that which one believes necessary to fulfill it. Clearly, then, an assessment involving horizontal norms requires seeing how two or more things on the intentional level fit together, not simply whether one such thing matches the part of the world it represents. This type of norm has its own presence in the discussion surrounding the Normativity Thesis – most prominently in work of a Sellarsian or Davidsonian bent – and in fact Dretske and Rosen both say things in the course of their discussions that could be taken to suggest its relevance (though neither takes up the issue).\(^9\)

In Chapter 2, I will argue that this kind of norm can indeed underwrite a legitimate version of the Normativity Thesis. However, the argument I will offer

\(^7\) E.g. in the introduction to Millikan (1993).

\(^8\) Fodor rejects the idea that normativity has anything to do with this challenge as well. See Fodor (1990: 135-136).

differs in important respects from other defenses of this claim in the literature\textsuperscript{10}, and this will turn out to have a significant effect on the relevance of the Normativity Thesis to the naturalistic project (i.e. on the viability of the Normativity Objection).

2. Historical sources for the Normativity Thesis

The Normativity Thesis arguably has a long history, some saying it can be traced back as far as Kant\textsuperscript{11} or even, in a confused and inchoate form, Descartes.\textsuperscript{12} To get an initial fix on the claim, though, it is helpful to look briefly at two issues outside the philosophy of mind – the naturalistic fallacy in ethics and anti-psychologism in logic. These serve as paradigms for normativity claims and their anti-naturalistic implications, and the issue of normativity in the philosophy of mind is at least implicitly, and usually quite explicitly, seen as parallel to these paradigm cases. Discussing the latter will shed light, then, on exactly what the Normativity Thesis is meant to say – but it will also lead to some initial questions about whether intentionality is really a parallel case. Having glanced at some of the deeper background, I will then examine at somewhat greater length one of the earliest and most influential explicit versions of the Normativity Thesis, that found in Kripke (1982). That discussion, too, will serve to bring out some of the obscurities surrounding the basic claim.

2.1 Two paradigms

The naturalistic fallacy, as I am understanding it, is shorthand for the claim that normative facts are never established or necessitated by descriptive facts alone.\textsuperscript{13} This claim is most familiar from ethics.\textsuperscript{14} For example, it is often pointed

\textsuperscript{11}E.g. Brandom (1994)
\textsuperscript{12}E.g. McDowell (2004)
\textsuperscript{13}As can be seen here, I treat “normative” and “descriptive” as classifications for facts and properties, not (except derivatively) bits of language or concepts. So by “normative conclusion” I mean a conclusion about a normative matter, not a conclusion couched in normative language. This decision is motivated largely by ease of exposition, but it also goes along with what, I take it, my main dialectical target would accept.

Using the terms this way may also help to avoid a potential confusion. A normative/descriptive distinction (or something close to it) also arises in the dispute between cognitivists and non-cognitivists (e.g. expressivists) in meta-ethics, and there the main issue is language rather than facts/properties. That debate concerns whether or not normative claims – where that is understood as a linguistic category – are descriptive in the sense of being (ordinarily) truth-conditional. I will simply grant that they are descriptive in this sense and that their truth-conditions involve normative facts. (Denying this, as the expressivist does, is generally taken to bolster the case for naturalism, so it would not be a good option for the proponent of
out that it does not follow from the fact that humans do engage in a certain selfish behavior (or that there is a genetic basis for it, or that it has been selected for evolutionarily, etc.) that we should engage in that behavior. Or, a bit more controversially, it is maintained that an action is not morally right simply because it leads to the most pleasurable outcome. The same kind of claims could be made in epistemology to the extent that its conclusions involve what we ought to believe, how we ought to reason, etc. The truth of the fallacy could be disputed, of course, but I will simply grant its truth in this relatively crude form. In any case, this provides an initial sense of which properties are normative: being right or wrong, good or bad, what one ought to do or ought not to do, etc.

It only takes one so far, though, toward understanding the Normativity Thesis. The central claims in the philosophy of mind do not at first glance concern what anyone ought to do or what is morally right or wrong. So any attempt to apply the naturalistic fallacy to the philosophy of mind faces a crucial challenge. It must be shown that certain mental properties are similar enough to those properties we already accept as normative to be classified that way themselves, or it must be shown that certain mental properties, when closely examined, essentially depend in some way on those properties we already accept as normative. Much of this chapter will be devoted to showing how difficult it is to meet this challenge.

We can turn now to anti-psychologism in logic. This position maintains that psychological facts about how people reason cannot explain the truths of logic. It is often taken to rest on the following line of thought: logic concerns the way we ought to reason, not the way we actually do, so the former cannot be reduced to or otherwise fully explained by the latter. In that case it would be just another example of the naturalistic fallacy, a case of the normative resisting full capture by the descriptive sciences. But it is important to see that anti-psychologism does not commit one to this view. The main claim is only that a specific set of descriptive facts – the set of facts concerning how we actually

---

14 Moore (1903) and Hume (1740/1978) are the classic early sources.
15 Janvid (2004) argues that epistemology is in fact not akin to normative ethics in this way, but is instead more similar to philosophy of mind as I characterize it in what follows.
16 In the Appendix, I provide a detailed exposition of the naturalistic fallacy and propose a novel way of understanding how it functions. In short, my contention is that the fallacy is underwritten by the assessment-sensitivity of the notions that are taken to be irreducible.
reason – is insufficient for explaining logic. It is consistent with this, first of all, that some other set of descriptive facts can do so. But more importantly, even if no set of descriptive facts can explain logic, it might still be the case that logic is itself descriptive in nature. This is, after all, what Frege himself, a paragon of anti-psychologism, seems to have thought – according to him the laws of logic are descriptive truths concerning the relations and properties of abstract objects.\textsuperscript{17}

This is worth mentioning for two reasons. First, it serves to warn us against making a parallel mistake in the philosophy of mind. That is, when it is claimed that how we actually apply a given concept (i.e. a specific set of descriptive facts) cannot account for that concept’s truth-conditions, it is important to keep in mind that this would not establish that the latter is \textit{normative} in any sense. Normativity claims cannot be derived from the bottom up, as it were; they have to be defended more directly, on their own terms.

The second lesson bears on attempts to provide a direct defense of that kind. We have seen that Frege took logic to be a descriptive enterprise, but he could (and did) still think that it concerned the way we ought to reason. How could that be? Well, as he put it, “from the laws of truth there follow prescriptions about asserting, thinking, judging, inferring,” adding elsewhere that the laws of logic are “authoritative for our thought if it wants to attain truth” (both emphases are mine). This shows how closely normative facts can cleave to descriptive facts. Consequently, great care must be taken not to confuse something’s being normative in nature, on the one hand, with its being a descriptive feature to which normativity cleaves – what I’ll call a \textit{normative target} – on the other. The danger of this kind of confusion is particularly acute, I think, when it comes to claims of the following form: X is a \textit{norm} for Y (e.g. logic is a norm for reasoning). If the truth of a claim like this sufficed to show that X was normative, too many things would qualify and the Normativity Objection would lose all plausibility. Given the right activity in the Y-slot (e.g. wall-building), basically \textit{any} descriptive fact (e.g. sturdiness) can fill the X-slot.

So this is another important aspect of the challenge facing the defender of the Normativity Thesis: it must be shown that the putatively normative mental properties are not in fact \textit{descriptive} properties serving as normative targets.

\subsection*{2.2 Kripke and the Normativity Thesis}

\textsuperscript{17} See Frege (1893), Frege (1918). It is important to mention, though, that following Frege in his Platonism is not the only way to follow him in thinking that logic is descriptive in nature. One could think, e.g., that logic describes relations of truth-preservation among structured propositions without being committed to any particular metaphysics.
Much of the contemporary discussion of the Normativity Thesis derives from Kripke (1982). The explicit focus for Kripke – and, perhaps to a lesser degree, for Wittgenstein – is linguistic meaning, but as Boghossian (1989) makes clear, Kripke’s major points carry over to mental content as well.\(^{18}\) Chief among these points is a version of the Normativity Thesis. Kripke goes on, of course, to use this as the basis for a very strong version of the Normativity Objection, one that threatens to lead to meaning skepticism if no acceptable answer to it can be found. My concern here, though, will not be with this application of the thesis, but only with the precise content of the thesis itself.

Unfortunately, Kripke’s text is not as perspicuous on this point as one might hope. Here are two much-quoted passages in which the Normativity Thesis seems to be getting spelled out:

(K1) An answer to the skeptic must satisfy two conditions. First, it must give an account of what fact it is (about my mental state) that constitutes my meaning plus, not quus. But further, there is a condition that any putative candidate for such a fact must satisfy. It must, in some sense, show how I am justified in giving the answer ‘125’ to ‘68+57’...[it must] show that only ‘125’, not ‘5’, is the answer I ‘ought’ to give.

(Kripke 1982: 11, my emphasis)

(K2) Suppose I do mean addition by ‘+’. What is the relation of this supposition to the question how I will respond to the problem ‘68+57’? The dispositionalist gives a descriptive account of this relation: if ‘+’ meant addition, then I will answer ‘125’. But this is not the proper account of the relation, which is normative, not descriptive. The point is not that, if I meant addition by ‘+’, I will answer ‘125’, but that, if I intend to accord with my past meaning of ‘+’, I should answer ‘125’.

(37, emphasis in original)

There is plenty that is confusing here. It is clear enough, in (K1), that the terms ‘justify’ and ‘ought’ are meant to bring out content’s normative dimension, but it is less clear exactly how it is to be understood. Kripke says explicitly that the norms his discussion concerns are not mathematical but metalinguistic.\(^{19}\) So it is granted that, for the plus relation, ‘125’ is the correct answer, and the question is whether a person’s mental state (or past uses, or dispositions, etc.) can justify or make proper giving that answer. But if what is sought is justification from a third-person perspective, this seems to reduce to the problem of establishing what the person means (i.e. what Kripke calls the first condition an account must meet). Kripke admits, in the sentence immediately before (K1), that the following conditional is true: “if I meant plus, then unless I wish to change my usage, I am

\(^{18}\) For this reason, I will generally permit myself to formulate Kripkean points in either of these ways, i.e. as concerning the meanings of terms or the contents of mental states or concepts.

\(^{19}\) See pages 8-11.
justified in answering (indeed compelled to answer) ‘125’.” If it is stipulated that a person has no intention to change usage, it looks like that person’s justification for the answer ‘125’ rests solely on the fact of their past meaning (if there is one). If an account could provide such a fact, it would have done all it needs to do – the further normative condition does not appear to add anything substantive.

Consider an analogy: suppose that up until now I have had the habit of brushing my teeth every day at 10pm. If I want my current behavior to be in accord with my past habit, then I am justified in brushing my teeth – indeed compelled to brush my teeth – today at 10pm as well. This is perfectly compatible, though, with the habit being wholly dispositional in nature. Given one’s intention, what one ought to do is to enact the same disposition that one has had in the past. But for all that is said in (K1), meaning could be exactly parallel, in which case a dispositional account of meaning could in principle satisfy both proposed conditions. This is just another way of pointing out that the putatively normative second condition is idle, given the way Kripke sets things up – justification of the relevant kind comes for free once the constitutive first condition is met, regardless of the form the constitutive account takes.

Perhaps what Kripke has in mind in (K1), though, is justification from the first-person perspective. That is, a proposed meaning fact must serve – or at least be able to serve – as one’s deliberative basis for answering in a particular way. This reading is suggested by two things: (a) earlier in the passage Kripke frames the issue in terms of our confidence in the answer ‘125’, and (b) the main meaning-constituting fact on the table at this point in the text is a straightforwardly guiding conception, something akin to a set of instructions that we consciously follow. Still, the relevance of this first-person concern is dubious as well. After all, there doesn’t seem to be any necessary connection between what one relies on to justify a given belief and what makes that belief true. My reason for thinking someone is a U.C. Berkeley student – what makes me confident about it – could be that I have seen the person in several lectures on campus, but having that property does not constitute being a U.C. Berkeley student. Similarly, the facts that ground one’s confidence with respect to meaning (or understanding a concept) need not be the very same facts that individuate that meaning (or conceptual content).

Moving to (K2), an initial puzzle arises because of the way Kripke formulates the question at the start of the passage. He asks about a relation that is clearly descriptive in nature – the relation between one’s meaning and how one will respond to a given question – and then writes as if that very relation needs to be treated normatively. The point he is trying to make surely be put another way. Kripke’s basic question in this passage is better thought of as this: what follows – i.e. follows with necessity – from facts about one’s meaning (e.g.
that one means addition by ‘+’)? And the answer is that nothing descriptive follows, while something normative does. That is, while meaning something doesn’t guarantee the actual production of a certain response, it does guarantee that one ought to produce a certain response, or that a certain response would be justified. While this claim is explicitly about what follows from meaning facts, there also seems to be an implicit message here about what can go into an account of meaning in the first place. The idea seems to be that when one is seeking to explain meaning (as opposed to, say, predicting behavior), talk of what one will do has to be replaced with talk of what one should do.

A pressing question is what exactly the normative ‘should’ is getting at in this passage. Is it that one should say ‘125’ because that is the correct answer to the question, or because, as the development of Kripke’s text seems to suggest, that is the answer that accords with one’s past meaning? Could a mathematically incorrect response still accord with one’s past meaning? If not, why not? If there are two different norms in play here—one mathematical, one metalinguistic—why must they coincide? Another question derives from the distinction (mentioned in 2.1) between being normative and being a normative target. Do content facts simply entail normative truths—a construal that goes naturally with my reconstruction of (K2)—or, what is quite different, are content facts themselves normative facts, as perhaps (K1) suggests? Finally, it remains unclear how to apply the distinction I made in the introduction—are the norms that Kripke is bringing to light vertical in nature, solely relating a contentful state or a meaningful linguistic act to the portion of the world it is about? Or are they horizontal norms, relating one such state to other states and actions that have different (but related) contents? I will look at examples of both these interpretive approaches, and the preceding questions will come into play accordingly.

3. Vertical norms

3.1 Kripke and vertical norms

Boghossian, in his influential survey of this subject matter, interprets Kripke’s Normativity Thesis in a way that cuts through the textual uncertainty, taking it to be a straightforward thesis involving vertical norms. On his reading, “the normativity of meaning turns out to be…simply a new name for the familiar fact that…meaningful expressions possess conditions of correct use.” In essence, Boghossian avoids the problem I raised for (K1) by making the conditional that precedes that passage irrelevant. On his reading, being in accord with one’s prior meaning is not what is crucial, so the second condition that an answer to the skeptic must satisfy is not to provide justification in that sense—i.e. the sense in
which, as we saw, one could also justify brushing one’s teeth at 10pm because it accords with one’s past habits. Rather, the second condition simply states that a proposed meaning fact must somehow establish the conditions of correct use for that term or concept.

There is a lot that can be said for this interpretation. We can make sense of Kripke’s claim that the issue is metalinguistic and not (in his example) mathematical by noting that the question of correctness conditions is prior to the question of correctness. It is natural to read Kripke as granting, for the purposes of setting up the Normativity Thesis, infallibility with regard to calculation (or any other kind of concept application).20 So whatever a person means – ‘plus’ or ‘quus,’ ‘table’ or ‘tabair,’ ‘green’ or ‘grue’ – it can be stipulated for whatever instance is at hand that they apply the term or concept correctly. The challenge is to say, based on the person’s past usage and present psychology, just what it is they mean. This requires privileging one ‘correct’ answer over other ‘correct’ answers, or in other words, establishing that one set of correctness conditions is in play rather than some other set. As already noted, this problem is fundamentally about present usage and not the relation between present usage and past (which Kripke admits is just a presentational heuristic); it is also clear that it cannot be settled by a mathematical proof, or by an examination of an object’s shape or color – these activities are aimed at confirming correctness and presuppose a fixed set of correctness conditions.

Moreover, this interpretation provides a way of understanding the explicitly normative terms Kripke uses to formulate the challenge, e.g. in the second condition of (K1). On Boghossian’s reading, when Kripke talks about what is justified or what we ought to say, he is simply talking about what is correct. We can spell this connection out a bit by looking at the contrast case. When there are no fixed correctness conditions, an utterance (or mental act) can be no more than an “unjustified leap in the dark,” a “brute inclination” (##). This sort of event can be described and perhaps causally explained, but it falls utterly outside the scope of normative assessment.21 Once conditions of correct use are on the scene, however, certain questions concerning justification or what we ought to do follow

---

20 Later on, of course, he doesn’t grant this; when Kripke assesses dispositionalism, for instance, our fallibility with regard to concept application plays a crucial role in his argument against it.

21 Some care needs to be taken with this point. Even blurting out a meaningless sound can still be normatively assessed, e.g. as a successful (or at least justified) attempt to serve an independent purpose, like easing the tension in the room. This, though, is best thought of either as a vertical assessment of the intention (see footnote 6) or as a horizontal assessment of how the action fits in with one’s other mental states. What is at issue at this point in the text is a direct vertical assessment of the event itself, and this makes sense only if the event is meaningful, i.e. governed by correctness conditions. (This clarification applies to the next sentence of text as well).
directly. We can see here the tight link between correctness conditions and Kripke’s talk of normativity that Boghossian’s interpretation demands.

I will now set aside any questions of interpretational plausibility and focus, for the rest of this section, on Boghossian’s sharpened formulation of the Normativity Thesis: the normativity of mental content is a matter of its involving vertical standards of correctness.

The question to be asked now is whether this undeniable feature of contentful mental states – that they have conditions of correct use – is a genuinely normative feature. As Rosen points out, we have little to go on in answering this question beyond some paradigm cases of normativity to use as points of comparison. We know that the normative is to be contrasted with the descriptive, but something could be non-descriptive in a couple different (though related) ways: it could be prescriptive or it could be evaluative.22 The prescriptive mode is most often signaled by the presence of a ‘must’ or ‘should’ or ‘ought,’ and since these do pepper Boghossian’s (and Kripke’s) discussion, it is reasonable enough to begin there.

3.2 Vertical norms understood as prescriptive

Is there anything prescriptive about correctness conditions? If there is, then the Normativity Objection certainly looms. For if intentional content amounts to or essentially involves a set of truths about what one should do in various circumstances, then it is plausible that any account of content relying on paradigmatically naturalistic resources – e.g. one’s behavioral dispositions or the causal profile of the state (including the selective history of that state-type in the species as a whole) – will fail to establish those essential content-constituting truths.23

22 Couldn’t the exhaustiveness of these options be questioned? Indeed, isn’t the notion of correctness itself often included in the list of what is paradigmatically normative (see Greenberg (2005), for instance)? Since I take a substantive notion of normativity to be one that plausibly grounds a form of the Normativity Objection, I see paradigm cases of the objection – the naturalistic fallacy in ethics and anti-psychologism about logic – as the best sources for paradigmgs of normativity. It is fair to say that prescriptions and evaluations fit that bill. I would suggest that those who include correctness in the initial list of what is normative either have not yet done the work of arguing for that assimilation or are working with a notion of normativity that is less demanding than mine. Greenberg, for instance, allows that a normative view of content is compatible with reductionism (see p. 312) and so seems to fall in the latter camp.

23 The point is only that such a claim is plausible, which is enough to establish that this construal of the Normativity Thesis is substantive. (Recall that my test for this is simply that the thesis provide prima facie grounds for the Normativity Objection). I am not simply begging the question against naturalistic theories like Millikan’s, which purport to establish prescriptive truths of just this kind. Many philosophers have indeed felt that Millikan’s view fails to establish genuine prescriptions, or at least prescriptions of the right kind; my claim is not that they are right about this, but that they can make a reasonable case for it.
The initial thought is this. A concept has particular correctness conditions – it is correctly applied in cases \( u, v, w \) and incorrectly applied in cases \( x, y, z \). What this amounts to, or what it essentially involves, is that the concept ought to be applied in cases \( u, v, w \) and ought not to be applied in cases \( x, y, z \).

Dretske and Rosen both argue, convincingly in my view, that correctness conditions are not prescriptive in this way. Taking one of Rosen’s examples, a person might have reason to cheer up a frustrated child by blurting out the wrong answer to an easy addition question. The point of this example is that the correctness of a certain response is, as Rosen puts it, “practically neutral in itself” (xx), and it is only in conjunction with other motivating factors that it has any bearing on what one ought to do. To that extent, having a correctness condition is just like any run-of-the-mill descriptive feature.

An analogy: my brown pants have a certain same-color-as condition. This condition is met in some cases of putting on socks \( (u, v, w) \) and not met in others \( (x, y, z) \). If I want my socks to be the same color as my pants, then I should put on brown socks (i.e. bring about \( u, v, or w \)). But that my pants have the same-color-as condition that they do, and that the socks I am currently looking for meet that condition, is presumably as descriptive as it gets. This brings to light a recurring problem with Boghossian’s (and Kripke’s) prescriptive-sounding formulations of the Normativity Thesis – they are almost always in conditional form.\(^{24}\) In such cases, the ‘ought’ most plausibly comes from the desire or intention in the antecedent, not from the correctness conditions involved. There is, then, no reason to think of the correctness conditions as normative – they can perfectly well be thought of as normative targets instead.

One could respond to this line of argument in the following way. One could begin by noting that the person in Rosen’s example, while perhaps having reason to blur out the wrong answer, does not have reason to believe the wrong answer. This could be what makes mental states like belief special. Unlike speech acts or acts like putting on socks, it is constitutive of belief that it has a particular aim – namely, being true. So a belief, simply as such, ought to be true; the ‘ought’ here is automatic, not derived from any antecedent goal. And so a set of correctness conditions, making a belief the particular belief it is, should be seen as a set of prescriptions concerning when that belief should (and when it should not) be present.\(^{25}\)

Moreover, the response could continue, Rosen’s example shows only that a person sometimes ought, all things considered, to do or say what is incorrect. But


\(^{25}\) There has been a great deal of discussion about how exactly these prescriptions would need to be formulated – whether as conditionals or biconditionals, wide or narrow scope, etc. See, for example, Hattiangadi (2007). For the purposes of my argument, those details can be left aside.
that is compatible with there being one thing (at least) that tells in favor of the correct response – namely, the fact that it is correct. This consideration can be outweighed, of course, but it is still prescriptive in nature and comes directly from the correctness conditions involved. Putting these two points together, one could even grant that at times a person ought, all things considered, to believe something false, though there will also be an outweighed prescription, resulting from belief’s internal, constitutive aim, to believe what is true. Indeed, there must be such an outweighed prescription, on this view, for that prescription is what establishes that a belief of the given kind is present at all.

Dretske suggests, by way of reply to this, that what is constitutive of belief is simply that it can be true or false, not that it aims at the former. What makes the stronger constitutive claim seem at all plausible is just that we have very general purposes that make the truth of beliefs valuable. For instance, we want to make accurate predictions, since that will tend to serve whatever other purposes we have, and having true current beliefs is valuable for making such predictions. So even if we cannot imagine a single case where the prescription to believe what is true is wholly lacking – not just outweighed by other considerations but completely absent – that may show only how deep-seated our general truth-valuing purposes are.

It can hardly be denied that we do have such general purposes, and that we aim at truth at least in part in order to serve them. The question is whether this wholly accounts for the prescription to believe what is true. The current construal of the Normativity Thesis, which requires an essential connection between the facts of truth/falsity and the aims or demands of cognition, must say that is not the whole story. According to this view, the prescription stands on its own, independent of our purposes, and just happens to coincide with those purposes on a regular basis. What grounds are there for this claim, though? It is not clear, first of all, what explanatory value the claim has, for the more parsimonious view – that we have very deep purposes, difficult to imagine absent, which engender the prescription – can seemingly account for all the available data. More importantly, to the extent that we can imagine a radical change in our purposes, what stops us from imagining that the prescription to believe what is true falls away?

Note that this case is quite different from that of moral prescriptions. It is a common claim that moral prescriptions would remain in force no matter how our purposes

---

26 Incidentally, while I frame the discussion in terms of our “psychological” purposes – however that is to be understood – the claim of purpose-independence at issue here must extend to “biological” purposes as well. Even if, as organisms, we (or certain of our activities, features, systems, etc.) can be said to have purposes separate from our psychological concerns – resulting from, say, the evolutionary history of our species – the defender of the Normativity Thesis must take the prescription to believe what is true to be independent of them as well. Otherwise, assuming that biological purposes can be accounted for in causal terms, the route to the Normativity Objection would be closed off.
general purposes were to change. This stems from a felt demand for objectivity when it comes to what we morally should and should not do. If moral prescriptions are not to shift around depending on people’s aims and inclinations, then they must be seen as constraints on our purposes and so potentially in conflict with them.

When it comes to cognition and belief, though, this does not hold in parallel fashion. It is clear, I think, that we do not take the prescription to believe what is true to be anything like a constraint on our general inclinations or purposes; at best, the claim would have to be that it acts as a constraint on our cognitive inclinations. The latter claim is probably true in some sense – I will say a bit more about this shortly – but it cedes the dialectical ground. A prescription could constrain a narrow range of our purposes and still be held in place, so to speak, by other overarching purposes – the prescription to eat less meat, say, might be held in place by the general purpose of promoting sustainable resource use. It is only because moral prescriptions are constraints on purposes per se that they remain constant no matter which purposes change and how radically. So a putative analogy with moral prescriptions does nothing to discount Dretske’s suggestion that the prescription to believe what is true is contingent on our general purposes.

It would also be mistaken to argue against that suggestion on the grounds that the constraints on cognition are decreed by the higher court of logical law. On this line of thought, the prescriptions governing belief would be purpose-independent because logical standards themselves are purpose-independent. Now I have already mentioned (in 2.1) that it is contentious to think of logic as fundamentally normative rather than descriptive, but we can set that point aside. The immediate problem here is that logic, even if normative, involves nothing like the vertical norms presently at issue. If there are any essential logical prescriptions, they are directed not at truth itself but at truth-preservation. So logic cannot provide any basis for the idea that belief has its own built-in aim.

The upshot is that we are left with no reason to view correctness conditions as themselves prescriptive in nature. It is far more natural to see them as descriptive but invariably caught up in conditional obligations, as normative targets at which we are always (at least in part) aiming.

An important lesson can be drawn from this dialectic, however. It has started to become clear that what motivates the Normativity Thesis is not the unimagininability of belief without its own truth-directed aim, but rather the unimagininability of believing (or cognition, more broadly) without general truth-

27 There is a demand for objectivity as applied to beliefs, but it is quite different in nature. Objectivity for beliefs is a matter of perspective-independence rather than purpose-independence, and that can be accounted for simply by understanding truth and falsity in a certain way. There is no need for a prescription to play any role here.
reliant purposes. In other words, “truth is the constitutive aim of belief” is best seen as shorthand for “aiming at truth is constitutive of believing.” This latter formulation, unlike the former, implies that if a creature shows, for example, no propensity to gather and check on evidence, no patterns of inference that accord with truth-preserving principles, and no sensitivity to the varying outcomes of cognition, then that creature simply cannot be a believer. Of course, if a creature fails these tests for being a believer, it immediately follows that the creature cannot have beliefs, and this may in part explain the common slide to the shorthand formulation of the point. But it is important to see that the underlying thought, expressed more clearly in the longhand formulation, cannot serve to defend the Normativity Thesis as currently construed; that underlying thought concerns whether or not certain horizontal norms are satisfied – e.g. whether a belief is rationalized by the observations or other beliefs on which it is based – and so spills into the subject of Chapter 2. Looking ahead very briefly, I will end up accepting a version of this (longhand) view, but it too allows for multiple interpretations and needs to be understood in a particular way.

To sum up, there are two basic problems with the claim that correctness conditions are prescriptive. First, it is far more compelling to see them instead as normative targets, i.e. descriptive features toward which prescriptions are directed as a result of our independent purposes. And second, when the view is examined

28 I intend for ‘aiming at truth’ to be understood in a de re sense here, i.e. as an umbrella term for all sorts of purposive behaviors (like those described in the next sentence of text) that in fact tend to produce or maintain true beliefs, even if the behaviors are not explicitly directed at that end. In so doing, I mean to back away from any seeming commitment to a highly intellectualized construal of ‘purposes’ in the preceding discussion. It may be enough for being a believer to engage in purposive behavior that is as if guided by explicit purposes of a certain kind (such as wanting to make accurate predictions), or where the purposes are thought of as biological functions; difficulties may attend this kind of approach, but I intend to remain neutral about that dispute at this point.

29 Even if this is correct – and I think it probably is – it leads immediately to a very difficult question: how much is required? It is clear that not every episode of cognition has to manifest such a propensity, pattern, or sensitivity. And that raises another very difficult question: in what sense, if at all, is ‘aiming at truth’ an across-the-board requirement? I hope to address this second question, at least, in what follows (see Chapter 2).

30 Boghossian (2005) argues that horizontal norms like the ones mentioned above are, in a sense, subsidiary to the primary vertical norm to believe only what is true – the former, he says, are simply means to satisfying the latter (just as examining a company’s business fundamentals is a means to satisfying the overarching norm to “buy low and sell high” when investing). There main problem with this suggestion is that if the vertical norm really were the fundamental one, then any means to satisfying it would be normatively acceptable. But satisfying it by, say, disregarding the evidence and believing something that just happens to be true anyway, or by adopting a general policy of believing only a few things about which you are completely certain, is in fact normatively unacceptable. The true beliefs themselves would be fine, of course, but the cognizer – the one to whom the norm applies on Boghossian’s proposal – would be deficient nonetheless.
more closely, it becomes clear that whatever crucial prescriptions are involved are in fact horizontal in character rather than vertical.

3.3 Vertical norms understood as evaluative

Perhaps we should think of vertical norms along the lines of the other normative paradigm: evaluative properties like good or right. An initial reason to think this is that they enter into superficially similar claims. For example, “she is a good person” appears to have the same structure as “that is a true belief.” And the fact that evaluative properties also resist identification with paradigmatically naturalistic ones – that is, they are subject to the naturalistic fallacy – makes it plausible that the Normativity Objection could be sustained on this approach. The basic idea would be that correctness conditions for a concept amount to a real valence scale: to say that a particular application is true (or false) is not simply to describe a correspondence (or lack thereof) between two things, but to categorize the application as positive and commendable (or, on the other hand, defective and regrettable). And so, it could be argued, no amount of pure description of the situation could yield the right kind of verdict, the positive or negative evaluation that makes correctness conditions what they are.

On the face of it, there is a problem with this idea. One might think, especially in light of the discussion in the 3.2, that we find some concept applications positive and commendable precisely because they satisfy a descriptive correspondence demand. If we have good reasons to prefer beliefs that meet certain descriptive conditions (as we might have reasons to prefer socks that are the same color as our pants), we will surely praise and endorse the ones that do. In other words, while the truth or correctness of a concept application may regularly carry with it an evaluation as positive and commendable, the assessment of its truth is not the same as that evaluation. So the move to assimilate correctness conditions to evaluatives goes too fast.

However, the defender of the Normativity Thesis could respond as follows. Treating truth as a descriptive correspondence relation works only because the contents of mental states for granted. But that just skips over the crucial step, for normativity enters the picture at the point of determining a state’s content or correctness conditions. One might deem the subsequent question concerning the satisfaction of those conditions – i.e. whether that contentful state is true or not – a descriptive one, but that descriptive issue is not wholly independent of normativity. So the situation is not at all like that involving socks and pants, where the same-color-as relation is clearly independent of normative evaluation;
on the contrary, it is the evaluative mode and it alone that provides the principle for identifying mental states with specific contents in the first place.\footnote{This may sound like an odd suggestion, since it holds that a certain relation – in this case, correspondence – is descriptive even though it essentially depends on a normative framework. Since I will go on to argue that this suggestion does not work, I can only do so much to motivate it, but perhaps an analogy will help. Imagine establishing a certain goodness-condition for performance on an exam – say, that the performance is good so long as it exceeds the mean score. It seems to me not entirely unnatural to say that, once that condition is in place, the fact that a performance on the exam is good is a descriptive feature of it – it is simply a matter of how it compares to the mean. The exam has, one might say, a good-making feature that is descriptive. But the fact that having that feature counts as being good depends on the evaluative standpoint that established the goodness-conditions in the first place. It might help, in applying this sort of picture to the case of cognition, to carefully distinguish (as I have not so far) truth and correctness. We might say, following Rosen, that truth is the correct-making feature for belief (and satisfaction for desire, etc.). So one might then hold that truth is a descriptive matter – a matter of correspondence between content and world – but the fact that correspondence of that kind counts as correct depends on the underlying evaluative framework that established the correctness-conditions in the first place. Indeed, the evaluative aspect seems to be even deeper in this case, because there is no way of characterizing one of the relata – the content of the state – without the establishment of correctness-conditions. (The same was not true of the exam score).}

This way of understanding normativity’s role fits better with what the standard naturalist alternatives are seen as doing. Those views too are aimed primarily at explaining how a state gets a given set of correctness conditions and only indirectly at the subsequent assessment of a state’s truth or falsity.\footnote{Two examples of proposed content-determining relations are (1) a nomological connection between representation-type and world that is not asymmetrically dependent on any other such nomological connection (due to Fodor), and (2) an abstract isomorphism between representation-type and world that has accounted for past evolutionary success (due to Millikan). Contentful state tokens are true if they stand in the corresponding making-true relations: (1a) being caused by a state of affairs the type of which stands in the relation named in (1); (2a) being isomorphic (in the same way) to a state of affairs the type of which stands in the relation named in (2). It can be seen that in both cases the content-determining aspect of the account is primary and the assessment-determining aspect secondary (and tightly linked to the first).} Of course, much more would need to be said on this last point to show how an evaluative framework might allow contents to be identified. But the preceding discussion has at least clarified the issue with regard to the location of normativity. What we should focus on, then, is the following question: can we accept the claim made above that normativity (of the evaluative kind) enters the picture right when contents or correctness conditions do? In other words, can we make sense of vertical norms, such as correctness conditions, as essentially evaluative in nature?

Unfortunately, it is difficult to say exactly what it is that makes something evaluative in the richly normative sense (as opposed to the sense in which we ‘evaluate’ whether or not our socks are the same color as our pants). Rosen suggests that full-blown evaluatives have an “‘internal connection’ with reasons for action.” But if that is right, then the Dretske/Rosen argument from section 3.2 would still apply. As long as the link between correctness conditions and our
reasons for action can be seen as external – however tight – there is no reason to think of those conditions as already evaluative. It should be noted too – recalling the earlier discussion of anti-psychologism – that it might still be true that correctness conditions cannot be accounted for reductively. They could be irreducible while still being descriptive in nature. I will say more about this issue in section 3.4.

Another problem for the idea that vertical norms are evaluative is that paradigmatically evaluative properties seem to have a distinctive characteristic, brought to the fore in Thomson (1997). As she puts it, “all goodness is goodness in a way.” That is, any instance of the property good is really an instance of a more specific property good for ... (or good to ... or good at ...etc.). ³³ It might be thought that this actually tells in favor of the current proposal, since correctness-conditions also vary depending on whether, say, beliefs or desires are at issue. So, one might say, there is the property of being correct for (or as) beliefs, the property of being correct for (or as) desires, etc. But this parallel collapses on closer examination. Thomson’s point is not merely that we have to specify the class of thing being assessed as good³⁴, but also the way in which it is being assessed as good. In other words, the property correct might be akin to the property big – in that something could be big for a Shih Tzu but not big for a dog – but it is not akin to good in the crucial respect: it does not make sense to ask in what way a belief is correct, i.e. what it is correct for …, correct to …, correct at …, etc.³⁵

Finally, to make matters worse for this construal of vertical norms, it seems that richly normative evaluatives carry a whiff of the subjective about them that correctness and satisfaction conditions do not (or at least not to nearly the same degree). I do not intend to take any firm stand on the objectivity or lack thereof of properties like good and right, but it is at least a contentious issue. A prima facie case can be made that paradigmatic evaluative properties are similar to properties like tasty, which certainly depend to a large degree on a given subject’s perspective.³⁶ To say the same thing about correctness conditions is to defend a much more radical position. Even those views about content that most emphatically stress the interpreter’s perspective, and relatedly, content’s

³³ Some of Thomson’s examples: “good for use in making cheesecake,” “good to eat,” “good at hanging wallpaper,” etc.
³⁴ As suggested by Geach (1956), for example.
³⁵ What about the property of being right? Thomson argues that this rests, in a somewhat complicated way, on the property of being good (specified, as it must be, in a particular way), but the details need not detain us. I think it is dubious that a maneuver of this kind could be applied to correct as well, but that would require further argument.
³⁶ One way of getting at this point, at the level of language, is to treat claims involving predicates like “tasty” as assessment-sensitive in the sense of MacFarlane (2007)).
indeterminacy, generally see that indeterminacy as like the shallow indeterminacy of arbitrary scales of measurement,\footnote{See Davidson (2001: 65, 145).} not the deep indeterminacy that might be thought to arise from differing tastes, differing cultural/moral backgrounds, etc.

These last two considerations are clearly not decisive. The point Thomson makes about good, while suggestive, does not necessarily apply to evaluative properties generally. And many philosophers are prepared to deny that good and right are in any way subjective, regardless of what gets said about tasty. Since that obviously does not preclude thinking of the former properties as evaluative, correctness conditions could be treated in exactly the same way. That said, the defender of a vertical-norm construal of the Normativity Thesis needs to provide a positive reason for seeing correctness conditions as normative, and the immediately preceding considerations instead reveal fairly prominent disanalogies. Correctness conditions have neither the same specifiability nor the same prima facie whiff of subjectivity as the paradigmatically evaluative features on which they are being modeled. That, in conjunction with the earlier point that there is no direct or internal connection between content facts and reasons for action (or even reasons for belief), suggests that no positive reason to see correctness conditions as normative is forthcoming.

In light of this, it seems both more natural and more parsimonious to consider vertical ‘norm’ a misnomer, since the features it picks out are ultimately descriptive in nature. If there is to be a workable version of the Normativity Thesis, then, the norms involved will have to be horizontal norms. In the next chapter, I will turn to examples of the thesis that have precisely this form.

4. Coda

There are two points worth making before moving on. First, I want to reiterate that the conclusion reached in this section – that correctness conditions are not normative – is not ipso facto a reductionist claim that correctness conditions can be fully accounted for in non-semantic terms. For one thing, horizontal versions of the Normativity Thesis have not been ruled out. A view of that kind might allow us to see why correctness conditions are irreducible even though the conditions themselves are not normative in nature – the normativity would come in at one remove, as it were. But setting that possibility aside for a moment, it must be kept in mind that while normativity can be an explanation for irreducibility, it is not the only one possible; contents and meanings might just be ground-floor descriptive features of the world or, perhaps more plausibly,
“emergent” in some sense that precludes reduction. The Normativity Thesis, as I understand it, is an attempt to account for irreducibility by revealing an important independent feature of mentality. I have argued against one version of the thesis understood in this way. The use of the term ‘normative’ in this context may sometimes slide into mere synonymy with ‘irreducible,’ but that usage is highly misleading. If that is how the claim that “intentionality is normative” is to be understood, then it is far less interesting and explanatory than it purports to be.

The second point requires a brief return to Kripke. It might be thought perverse that I have not addressed Kripke’s argument(s) against dispositionalism about meaning. After all, is that not where his claim that meaning is normative gets fleshed out? Rejecting his version of the Normativity Thesis would seem, then, to require a rebuttal for those arguments. I will not delve into Kripke’s anti-dispositionalist arguments in any detail, but I will say this: those arguments either employ an attenuated notion of normativity like that described in the previous paragraph, or they mislocate the source of normativity in the manner described in 3.2 and 3.3.

The dispositionalist position that is Kripke’s direct target maintains that the content of a concept is fixed for a given thinker by all of and no more than that thinker’s dispositions to apply it. This position rules out any possibility of error and so cannot be right. But if the claim that content is normative merely points out that content inherently involves the possibility of error, it is clearly true but also not particularly interesting (or at least not as interesting as the Normativity Thesis is often taken to be). This claim may not equate normativity with irreducibility, as was discussed above, but it could be seen as doing something even more trivial: equating normativity with irreducibility to crude dispositionalism (or any other view that fails to divide one’s actual or dispositional applications into two classes, the right ones and the wrong ones).

Of course, more sophisticated versions of dispositionalism have been proposed that aim to make just such a division and so to provide for the possibility of error. However, it is often thought that the Kripkean Normativity Thesis is deep enough and far-reaching enough to constitute a principled objection to these proposals as well. This requires a far more substantive and interesting construal of Kripke’s claim that content is normative, but it is difficult to make a construal of this kind plausible, as preceding discussion has shown.

A suggestion by Boghossian can serve as a final illustration of this. First, he says, suppose there was a dispositionalist account that managed to privilege a class of dispositions that successfully fixed the right extension for a concept (while avoiding circularity). Dispositions falling outside that privileged class could then

---

38 See Humphreys (1997), Kim (2006), and O’Connor (2005) for some recent discussion surrounding the notion of “emergence.”
be considered correct or incorrect depending on whether they produced applications to the same things that the privileged class did. How would a crucially normative aspect be lacking here? Perhaps in the following sense: knowing these dispositional facts would not motivate us, would not give us a sense of what we ought to do, in the way that grasping content does. In other words, if informed that we were disposed in conditions C to apply ‘eagle’ to all and only eagles, that fact by itself would not seem to push us in any particular direction with respect to future use or appraisal. Since facts about content do push us in particular directions in these respects, a suitable account would have to make that intelligible.

Boghossian says it is not obvious that all dispositional accounts would fail to meet this ‘intensional’ requirement, as he calls it, and his criticisms of actual proposals focus on the prior extensional requirement instead. But even if the intensional requirement were an insurmountable obstacle for dispositionalist accounts, that would not support the view that correctness conditions are themselves normative. The most it would show is that facts about intentional content enter into our normative thinking in ways that facts about dispositions do not. But that could just as well be true of two coextensive yet non-identical descriptive properties. The fact that two covarying properties have different normative profiles, i.e. engender different conditional obligations in the context of our background purposes, may give us good reason to think that neither property is reducible to the other, but it does not ensure that one of the two properties is normative.

Kripke’s anti-dispositionalist arguments cannot, then, be used to support the vertical construal of the Normativity Thesis.
Chapter 2

Establishing the Normativity Thesis: Horizontal Norms

1. Introduction

The Normativity Thesis holds that certain fundamental aspects of human mentality are normative in nature. That quite general formulation needs to be specified, however, before the claim can be properly assessed. In the previous chapter, I argued that a construal of the Normativity Thesis according to which it concerns so-called “vertical norms” – correctness conditions for concept application, the “norm of truth” for belief, etc. – cannot be sustained. It is plausible, certainly, that correctness conditions and the like are essential aspects of human mentality, but the claim that they are normative aspects does not survive scrutiny.\(^{39}\)

In this chapter, I will examine the possibility of interpreting the Normativity Thesis as a claim about horizontal norms. The norms at issue here will be relational, holding between different mental states, between mental states and pieces of evidence (if the latter are different from other mental states), between mental states and actions, etc. To give them a general label, these are “rational relations.”\(^{40}\) And this version of the Normativity Thesis claims, in short, that rational relations are both normative in nature and essential to making some mental states the states they are.

This construal of the Normativity Thesis faces a problem that is basically the reverse of that confronting the earlier construal (involving vertical norms).

---

\(^{39}\) That is, they are not ‘normative’ in any sense robust enough to provide a \textit{prima facie} case for the Normativity Objection, i.e. the claim that naturalism is ruled out.

\(^{40}\) This phrase is to be understood broadly, and not, for example, as referring strictly to norms of formal coherence. I pick up on this point below.
That is, while it is fairly straightforward – though not completely trivial – to establish the normative bonafides of rational relations, it is more difficult to show that these relations are essential to intentionality, as is required. The bulk of my argument is meant to defend the following, somewhat qualified claim: rational relations are essential to mentality if relations of another kind – viz. horizontal causal relations – are essential to it. I will have a bit to say in favor of the antecedent, but that discussion enters into a very large controversy about holism that I cannot hope to settle definitively. Nonetheless, I believe that establishing the conditional is important in its own right. For one thing, the acceptance of the antecedent – the claim that the identity of certain key mental states depends essentially on causal relations those states stand in to various other states, to perceived events, and to performed actions – is already very common. And in any case, establishing the conditional helps to delineate the terrain one confronts when theorizing about the mind.

Before getting to the main argument, though, I begin with a brief section directed at the relatively more straightforward task of establishing that rational relations are indeed normative in nature.

2. Rational relations as normative relations

Some will surely feel that rational relations (e.g. justifying, providing reason for, rendering incoherent, etc.) belong to the set of paradigmatically normative notions (like good and right), thus making a defense of their normative nature unnecessary. However, there is in fact enough room for dispute about this – and enough room for terminological confusion – that a few comments are in order.

First, it is important to note two things: (1) the sense of ‘normative’ I am working with is fairly broad, deriving primarily from an intuitive contrast with the (purely) ‘descriptive’ and allowing for both prescriptive and evaluative brands; and (2) my use of ‘rational relations’ is similarly broad, including relations between evidence and belief, between intention and action, etc., as well as relations holding between mental states themselves. This means that arguments such as those due to Kolodny (2005, 2008) and Gluer and Wikforss (2009), which can be seen as denying the normativity of rational relations, can be sidestepped in the present context; those arguments rely on a more restricted sense of ‘normative,’ a more restricted sense of ‘rational,’ or both.

Kolodny, for instance, makes very clear that his use of ‘rational’ is limited to relations holding between different mental states, and that his use of ‘normative’ is strictly prescriptive – i.e. tied to the “first-person standpoint of deliberation” and the “second-person standpoint of advice”. The conclusion he
argues for can be put like this: there is no general prescription to maintain formal coherence, as such, among one’s mental states; there are instead particular prescriptions to believe and do particular things (given the state of the world), and these things will, as a matter of fact, have formal coherence as a general property. As Kolodny puts it, coherence turns out to be a byproduct rather than a telos in its own right. If correct, this point is nonetheless compatible with the truth of the Normativity Thesis. This is because Kolodny’s argument explicitly allows (a) that rational relations of the narrower sort still have an evaluative dimension and (b) that rational relations of the broader sort – e.g. those holding between mental states and evidence – can (and do) have prescriptive normative import as well.

Gluer and Wikforss take a related, but slightly different tack. They claim that rational relations are not normative because they cannot serve as deliberative rules guiding one’s behavior. Their basic idea is that one responds to reasons directly – upon becoming aware of the evidence, one straightaway believes things that are based on it, for instance – and there is no room for using one’s understanding of rational relations in an explicit or implicit decision about what to believe or do; put another way, while we clearly do not always accord with the requirements of rationality, we are not able to intentionally violate them either. For this reason, they conclude, it is better to see rationality as constitutive of thought and not as a prescriptive, motivating factor.

Gluer’s and Wikforss’s argument, unlike Kolodny’s, can be read as applying to rational relations on the broader understanding (i.e. including relations beyond that of coherence among one’s own mental states), but it also clearly involves a highly restricted notion of ‘normativity.’ The prescriptive, guiding notion of normativity they operate with is strictly first-personal, ignoring both the “second-person standpoint of advice” and the “third-person standpoint of evaluation.” But surely Gluer and Wikforss would not deny that it is possible to offer advice of the following kind: “you ought to conclude that the professor did it, for the evidence all points in that direction.” It may not be possible for the recipient simply to decide to believe what is recommended – they will probably have to look again at the evidence to see if there is something they are missing – but the advice itself is genuinely normative nonetheless. And similarly, it is always possible to evaluate how well someone lives up to rational standards, though it may not (and perhaps cannot) take the form of attributing questionable decision-making to them. These two models of normativity are enough to make the Normativity Thesis, as currently construed, potentially viable.

The discussion so far has been largely terminological, but there is an underlying question about the normativity of rational relations that still needs to be dealt with: put simply, why couldn’t rational relations be thought of as descriptive in nature rather than normative? For instance, a belief or piece of evidence
justifying or providing reason for another belief might simply mean that it raises the latter’s probability of being true; that one’s beliefs and desires make sense of a particular action might simply mean that the action maximizes the satisfaction of one’s set of preferences. These are quite clearly descriptive properties of the items in question. Moreover, our evaluations of other people’s rationality might simply be factual evaluations – as when we determine how many questions someone got right on an exam, independent of any judgment about whether they did well or poorly – and our advice could be instrumental advice, the ‘ought’ deriving from an assumed desire on the part of the recipient to have true beliefs.

A response to this kind of proposal can begin with an appeal to something like Moore’s open question argument. This strategy is spelled out in Gibbard (1990). Gibbard holds that rational relations are normative insofar as characterizing something as rational essentially involves endorsing it. In other words, our evaluations of others’ rationality are not simply factual in nature – they are (at least in part) determinations of whether we do or do not endorse their reasoning, their actions, etc. In order to show that this is so, Gibbard adduces a few key examples – the prisoner’s dilemma, for instance – and argues that questions about the rationality of particular courses of action in such cases persist even when the descriptive facts – e.g. that one course of action maximizes expected utility – are settled. This implies that something is built into the concept of rationality that outstrips whatever facts serve in proposed descriptive analyses of it. Gibbard claims that this extra thing is endorsement; whatever it is, though, it sets rationality off as a normative rather than a descriptive notion.

The important point is that this argument is very much akin to the paradigmatic case of the naturalistic fallacy in ethics, far more so than any of the vertical norm proposals from Chapter 1. In those cases, we found that the alleged normative property – being true/correct – could be seen as a descriptive feature to which a separate endorsement was (invariably) appended. But in the present example, the argument is designed to show that the key property – rationality – is at the level of endorsement, not that of the descriptive feature on which the endorsement is based. The open questions Gibbard highlights are of the following kind: “here is a certain descriptive property a mental transition has…but do I deem that a good property to have, one that I endorse?” And that fits the paradigmatic pattern of argument for a normative/descriptive gap.

Two further points can be made. One is that rationality assessments, unlike truth assessments, can be seen as having the direct connection with reasons for action that is distinctive of normative notions. If, as Gibbard suggests, rationality assessments involve endorsement, then it follows that finding something to be

---

41 As I showed in Chapter 1, no compelling case can be made for the claim that beliefs as such are supposed to be true, as opposed to the claim that we have deep-seated reasons for wanting them to be true.
rational will mean seeing it as something one ought to do oneself or something one ought to advise for others. But setting Gibbard’s specific view aside, the discussion in Chapter 1 showed in a general way that the thought that belief (qua belief) must aim at truth is better understood as the thought that a believer (qua believer) must meet certain rational requirements. As such, the case can be made that there are ground-floor reasons to do or think what accords with rational standards (as opposed to there being ground-floor reasons to have – out of nowhere, as it were – a true belief). The second point is that normative notions, as mentioned in the previous chapter, seem to have – or at least can be plausibly argued to have – something subjective about them, or anyway something that allows for persistent disagreement. Rational relations have this characteristic as well. The distinctive way in which open questions come into Gibbard’s argument reflect that fact.

At this point, then, I take it as established that rational relations are normative – that judgments of rationality are evaluative in a way that purely descriptive judgments are not – and the question now becomes whether those normative relations are in fact a fundamental aspect of human mentality. While everyone can agree that our mental states stand in rational relations, and that certain normative facts are engendered by our being in such states, is it necessary to go a step further and claim that our mental states are in some sense constituted by rational relations? For the Normativity Thesis to be true, this question must be answered affirmatively. I will now argue that, with an important qualification, that is indeed how the question should be answered.

3. The constitutive role of rational relations

3.1 The basic approach

It is difficult to argue for the constitutive role of rational relations directly. The most well-known attempts to do so come from an interpretationist perspective, claiming that facts about the mental states a creature has are constituted by facts about how others can most profitably interpret its behavior. Since others must utilize rational principles to profitably interpret other creatures’ behavior, it follows that these rational principles help constitute facts about the mind. Dennett and Davidson are the foremost examples of this approach (though their formulations of interpretationism may not be quite so crude), but the approach is controversial, to say the least, and I intend to steer clear of it.

For those who take the more common – and in my view, more antecedently plausible – approach to mental states, treating those states in a realist, interpretation-independent fashion, it is less clear how it might be established that
standing in certain rational relations is constitutive of those states. After all, standing in particular rational relations could easily be seen as the result of being a particular kind of state, not part of what makes it that state in the first place. And there are, of course, plenty of candidates for the constitutive basis of mental states that do not involve rational relations – e.g. brain states that are causally related in the right way to the objects and properties spelled out in the state’s content.

There is, though, a promising line of argument, propounded in different ways by Bilgrami (2004), Zangwill (1998), and Wedgwood (2009) – and inspired, in most cases, by Kripke (1982) – that defends a more qualified claim. The argument can be seen as supporting the following conclusion: if one adopts a functionalist approach to explaining content-bearing mental states – which is to say, if one accepts that horizontal relations of some kind (viz. causal ones) are necessary in accounting for such states – then one must, in the end, accept a constitutive role for rational relations. In other words, purely causal functionalism is an unstable position: an account of mental states must either include normative rational relations or not involve horizontal relations of any kind.

I believe that a successful argument for this qualified claim can indeed be made, but that previous attempts have left out an absolutely critical component. What is needed, as I will explain in more detail below, is a shift from treating mental states as static entities to treating them as fundamentally dynamic in nature. I will work toward a successful version of the argument by briefly examining the three examples mentioned above.

### 3.2 Previous attempts

The basic idea behind this sort of argument is fairly simple: an individual’s causal dispositions prove too variable and undependable to serve on their own as the constitutive basis for having particular mental states, and normative relationships are just what is needed to plug the gap. Bilgrami (2004) presents this basic idea in the starkest and, in my view, most problematic terms. He claims not just that normative relations must come into play somehow in accounting for intentional states, but that intentional state are themselves normative entities. They are, he says, “commitments to think various things and do various things” (Bilgrami 2004: 128). So, for instance, a belief that there is a table in front of one is fundamentally a commitment to believe other things, such as that there is something in front of one (to name one of many). And the desire to help the poor

---

42 By way of analogy, it is surely intuitive that an action is wrong, when it is, because of the type of action it is – misleading someone in order to benefit oneself, say – and not that its being wrong is part of what makes it that type of action in the first place.
is fundamentally a commitment to do various things, such as giving money to charities or advocating certain public policies, etc.

This identificatory step is precipitous, in my view, and in any case risks begging the question. What is clear at the outset is that beliefs, desires, and other intentional states *engender* commitments of the kind mentioned, and what is needed is an argument showing that those rational relations must be part of an acceptable account of intentional states’ instantiation conditions.

Still, the argument Bilgrami offers is worth examining, for it might work even if the assumption that intentional states are identical to normative commitments is jettisoned. That argument is a fairly straightforward version of Moore’s open question argument. That is, Bilgrami suggests that one can see the failings of a causal dispositionalist approach by noting that certain questions – e.g. “I have a desire to help the poor, but do I have a disposition to give money to charity, or to advocate certain public policies, etc.?” – do not have trivial (affirmative) answers. This shows, as Bilgrami puts it, that the intentional state is “something over and above the disposition[s]” (128). Since parallel questions involving rational relations – e.g. “I have a desire to help the poor, but am I committed to giving money to charity or advocating certain public policies, etc.?” – are not similarly open, those relations can step in to play the constitutive role that causal dispositions are unable to.

While this is indeed suggestive of a key asymmetry between the causal relations a mental state stands in and the rational relations it stands in, there are two major problems with Bilgrami’s approach. The first, which I mention here briefly, is that the open question argument operates at the level of conceptualization and meaning, not that of metaphysics. That is, while it might show that how we think of intentional states is different from how we think of (a

---

43 This presentation of the argument is different from – and, in my view, an improvement on – Bilgrami’s own in three ways. First, I have made it clear that the dispositional portion of the question must be put in terms of a disjunction or cluster, whereas Bilgrami makes it seem as if a single disposition is what matters (or at least that the questions can be meaningfully posed about individual dispositions): it is unrealistic, though, to think that a dispositional account would focus on just one specific disposition as the basis for an intentional state. Second, Bilgrami has the questions going in the opposition direction, i.e. starting from the fact that one has certain dispositions. That form of the question may be open too, but his discussion clearly focuses on the (antecedently more plausible) claim that having certain dispositions is not necessary for having the corresponding intentional states, and I have altered the questions to reflect that. Third, because he identifies intentional states with normative commitments, he puts the non-dispositional portion of the questions in terms of ‘oughts’ rather than in terms of intentional states (e.g. “I have this disposition to f, but ought I to f?” (p.128); I have already said that I find the identificatory move hasty, but putting the question in this way has other adverse effects too: it introduces an important ambiguity in the term ‘ought’ (moral vs. prudential vs. rational, etc.), and it encourages a misleading sense of the question’s openness that arises from the fact that we can reflect on and reassess our commitments and change them going forward. That is, we may change our minds, deciding to go in a direction quite different from that of our current dispositions. But that kind of ‘openness’ for the question is not the one that matters.
set of) dispositions, there may be underlying property identities or constitution relations all the same. This limitation has been pointed out many times before, and Bilgrami does say some things meant to extend the argument so as to allow for a metaphysical conclusion, but I do not find his suggestion compelling. This is primarily because his suggestion relies once again on the unnecessary assumption that intentional states are equivalent to normative commitments.44

The more important problem, though, is that the central intuition this approach depends on – that the answers to the relevant questions really could go either way – is unsure at best. Bilgrami senses this problem as well, noting how odd it would be if there were a complete disconnect between the facts regarding a person’s intentional states and the facts regarding what they do or are disposed to do. To alleviate this concern, he suggests that at least one kind of second-order disposition is essential for having a given intentional state: the disposition to cultivate relevant first-order dispositions when one comes to see that they are lacking. So while one might have, for instance, a genuine desire to help the poor and yet not be disposed to give money to charity, or to advocate certain policies, etc., one must at least be prone to modify that behavior upon learning that those actions produce the desired result (and one’s current actions do not).

With dispositions of this kind in the picture as well, it becomes far less plausible that Bilgrami’s key questions are actually open ones at all.45 It might still be the case that no specific disposition can be inferred from the presence of a given intentional state, but is it really possible for an individual to have none of the relevant dispositions? Bilgrami argues that the questions are indeed still open, but once again I fail to find what he says compelling. First, he claims that the key second-order disposition can be specified only by means of normative concepts – as he puts it, it is a disposition to “accept criticism for not having lived up to the commitment to p [and q, and r, etc.]” – but it is not at all clear why this should be

44 What Bilgrami offers is a version of the Property Dualism Argument, best known from discussions of phenomenal states such as that in White (2002). His claim is that an a posteriori identity is not actually a genuine possibility in the present context. Such identities require that an entity be thought of via different modes of presentation – that is what makes it possible to answer the open question in the negative and still be rational – but the very same kind of open-question difficulties will arise for whatever property is thought to serve as the mode of presentation. I am not at all sure this argument is effective even in its original form and setting, but (a) it is almost certainly not effective against more modest naturalistic claims than identity (e.g. supervenience and psychophysical laws), and (b) it is almost certainly not effective if intentional states are thought of not as commitments but as states engendering commitments. The latter property is not clearly normative in the way Bilgrami needs to guarantee that the same problem crops up at the level of modes of presentation. (Compare: in many contexts, driving a car at an unsafe speed plausibly engenders a commitment to slow down. But surely that fact alone does not preclude that driving at an unsafe speed could be identical to, say, driving over 80mph in such-and-such particular circumstances).

45 The plausibility is even less, in my view, if this second-order disposition is seen not as absolutely essential, but as just one important disposition in a whole cluster of relevant dispositions – though perhaps one that is particularly important in the absence of other, first-order dispositions. More on this below.
so. The formulation I offered in the previous paragraph, for instance, is free of such notions. Bilgrami anticipates this sort of response, and after making a few comments about how complicated the spelled-out second-order dispositions would be — a fair point, but not a particularly threatening one — he simply falls back on the original intuition and says a new question with more detailed dispositional content would still be open. This intuition is far too questionable to provide an adequate basis for the Normativity Thesis.

Zangwill (1998) arrives at the same sort of conclusion but by slightly different means. He recognizes that a sophisticated dispositional functionalism — one that relies on clusters of dispositions (not all of which need to be realized) as the explanatory basis for intentional states — can get around the standard problem a functionalist view faces, for no single disposition need be thought of as necessary or sufficient for the state in question. It is not enough, therefore, to point out that certain specific questions are “open” in the Moorean sense or to note in some other way that the connections between intentional states and particular bits of behavior are generally rather loose. But Zangwill believes that this sophisticated form of functionalism is flawed nonetheless, and that a necessary role for rational relations can be demonstrated. He offers two main considerations in an effort to accomplish this: (i) rational relations have an explanatory priority over causal ones when it comes to intentional states, and (ii) dispositional functionalism is unable on its own to provide a basis for rational causation.

Zangwill’s first suggestion stems from the following line of thought. While a dispositional functionalist account states that intentional states are constituted by the presence of (a sufficient number of a batch of) key causal dispositions, it is seemingly far more intuitive to think those causal relations are explained by the nature of the states. In particular, Zangwill suggests, we might think they are explained in the following way: intentional states are constituted by a set of rational relations, and since people are (for the most part) rational, it follows that their causal dispositions will (for the most part) match their states’ rational profiles. So, for example, rather than thinking of a disposition to believe tomorrow is Thursday as part of what makes it the case that a person believes today is Wednesday, it is more natural to think of that disposition as resulting from the belief — in particular, resulting from the belief in virtue of its rational properties and the fact that the person meets basic rationality standards.

---

46 My formulation would need to be spelled out for individual cases, of course; “relevant first-order dispositions” is merely shorthand for whatever specific dispositions are thought to be potentially constitutive of the intentional state in question.

47 It is here, in particular, that Bilgrami seems to fall into the confusion discussed at the very end of footnote 4.
Zangwill sums up the explanatory priority of rational relations like so: “the causal properties of [an intentional state] are contingent on its rationality, whereas the rationalizing properties of [the state] are not contingent on its causal properties” (193). This is a highly suggestive claim, and I will eventually argue for one that is a close parallel, but Zangwill’s version is not quite adequate as it stands. For one thing, it is not clear why similar questions of explanatory priority would not arise for an account of intentional states in terms of rational relations too. Would we not intuitively judge of a given intentional state (e.g. the belief that it is Wednesday) that it makes sense of or rationalizes others (e.g. the belief that tomorrow is Thursday) because of its underlying nature?

This might be seen as a specific problem for Zangwill’s proposal, but it might also be seen as an indication of a more general fact – namely, that intuitions of explanatory priority are just not that dependable. As such, dispositional functionalists can resist the argument, saying that potentially misleading intuitions of that kind pose no major obstacle for a constitutive account like theirs. Zangwill himself grants this dialectical limitation and so provides further argument, bringing the second consideration mentioned above into play.

Zangwill’s further argument starts from the claim that “functionalism is not just a theory about what it is for a mental state to be instantiated; it is also a theory of...what it is for mental states to have rational causal efficacy” (186). In other words, the array of causal dispositions laid out by the functionalist plays two roles: it tells us (i) that, if enough dispositions in a certain cluster are in place, a certain intentional state is present, and (ii) that, if a particular intentional state is present and a particular disposition in its defining cluster is manifested, we have a case of rational causation, i.e. a cause that is also a reason. So, for example, if the disposition to donate to charity is part of the cluster constitutive of a desire to help the poor, then the realization of that causal connection (from desire to donating) is, ipso facto, a case of rational causation, or doing something for a reason. However, Zangwill goes on to claim, it is possible for there to be manifested dispositions of this kind that are not instances of rational causation – when there are deviant causal chains, for example – from which it follows that the second role of a functionalist account cannot be discharged.

I question this argument’s starting point. In my view, functionalism’s primary aim is to account for the instantiation of intentional states. If that could be accomplished, there would be other options for explaining – naturalistically or otherwise – the rational properties that accrue to those states. For instance, one could join together a functionalist account of what constitutes intentional states and an expressivist account of the normative judgments concerning their rational

---

48 Zangwill recognizes this point in his (2005).
relations. And if that is correct, and instantiation is the sole issue on the table, then the possibility of deviant causal chains is irrelevant; such a chain involves an intentional state and one of its standard results, so it cannot serve as a counterexample to the constitutive claims made by the functionalist.

Still, Zangwill’s discussion leaves one with a heightened sense that there is an important asymmetry between an intentional state’s rational relations and its causal relations, such that the latter depend crucially on the former. It is just that the precise nature of this dependence has not been nailed down.

In his (2009), Wedgwood offers two more considerations in favor of a constitutive role for rational relations. The first is an improved version of one of Bilgrami’s suggestions, while the second is a slightly different take on Zangwill’s idea of explanatory priority. I take these up in turn.

Wedgwood states at the outset that “it must be something about the thinker's dispositions...that determines which concepts figure in the thinker's thoughts” (7). But like Bilgrami, he argues that a specification of those dispositions necessarily involves normative concepts or properties. Bilgrami’s claims were based on an antecedent assumption that intentional states are equivalent to normative commitments – or at best an overhasty argument to that effect – but Wedgwood makes no such assumption. Instead, he points out that any statement of plausibly determinative dispositions will have to include a key normative qualification – namely, that there are no defeating conditions. For example, it is not the simple disposition to give money to charity that partly constitutes a desire to help the poor, but the disposition to do so as long as there is no reason to think, say, that all charities are corrupt, or that the services they provide are counterproductive, etc. To an extent, this is just a familiar point about the holistic nature of functionalism, but what Wedgwood emphasizes is that normative facts – facts about rational connections – provide the contours of the holism.

Functionalists have a ready-made reply to this point, of course. They will say that it is certainly much simpler to describe the relevant dispositions using qualifiers of the kind Wedgwood suggests, but it is nonetheless possible in principle to spell them out by means of complicated disjunctive descriptions, eschewing any mention of normative facts about rational connections. Functionalism is not meant to be a practical theory, after all, just one that assuages some fairly esoteric metaphysical worries.

Wedgwood does little to rebut this functionalist suggestion, which leaves the debate, as so often, at an impasse: one side claiming that it is most natural to

---

49 This is probably because he takes a unique approach to these issues, as can be more clearly seen in Wedgwood (2007). He argues first for the normativity of intentionality, and then uses independent arguments for the irreducibility of intentionality in order to reach the conclusion that normativity is also
understand things in terms of normative properties, the other pointing out that there could be a more naturalistic – though perhaps less natural – alternative all the same.

Wedgwood’s second suggestion brings this argumentative territory into even sharper relief. Like Zangwill, he highlights a certain explanatory priority that rational connections have over causal ones. In this case, though, it is not that the rational connections explain why the causal connections hold; it is that they explain why certain causal connections are in the constitutive cluster for a given intentional state and others are not. This point was implicit in Zangwill’s discussion as well: the only causal dispositions that actually matter when accounting for the presence of a certain intentional state are those in which cause and effect are rationally connected as well. For example, tending to lead to a belief that tomorrow is Thursday is part of what constitutes having a belief that today is Wednesday precisely because the former belief is justified or rationalized by the latter.

As Wedgwood explains, this priority holds because it is difficult to see how an irrational disposition could be essential to making it the case that a certain intentional state is instantiated. As he puts it, such states – or more precisely, their conceptual and attitudinal ingredients – are fundamentally cognitive abilities or tools, not cognitive defects (19). And even though certain irrational dispositions are so common as to be generally expected – the psychological literature on reasoning fallacies provides many examples – it would be a serious stretch to say that those dispositions are part of what constitutes certain intentional states. After all, if an individual (or population) lost that particular disposition while the others were all held constant, we would not feel that something was missing from the state itself – if anything, it would seem that the individual (or population) had the state in a purer or more complete form. That is not how it is with the dispositions that are in fact in an intentional state’s constitutive cluster, though; when such dispositions are systematically lost or changed, it at least tells against – though perhaps not decisively – the state’s instantiation.

These matters are made somewhat murky, of course, by the nature of the functionalism at issue. Given that it is based on clusters of dispositions, only a sufficient number of which need to be present in order to ground the presence of a given intentional state, it is clear that dispositions can come and go without having a dramatic impact on the basic intentional facts. This makes room, as we have seen, for relatively isolated instances of irrationality. And while that is a positive aspect of the account, it makes it much harder to test the impact of a single disposition in the way I have suggested. For even if it is correct that an irrational
disposition could always be lost without the intentional state it is associated with being lost too, the functionalist can reply that that is true of nearly any disposition taken in isolation. Moreover, the state would be had in a “purer” form without that disposition only in the sense that the cognitive role it plays would be more rational overall. But the functionalist can justifiably ask why that matters; to assume that it does begs the question at issue, which is whether rational connections have any priority in determining the causal dispositions in the cluster. It seems that the functionalist could stand firm and claim that any kind of causal disposition can play a role in the account. A common but irrational disposition might very well be part of a constitutive cluster – its loss (like that of most single dispositions) would just make very little difference to the larger intentional facts.

If that is right, then the explanatory priority of rational relations might seem to vanish. But I do not think that actually follows. For even if a functionalist account were to involve causal dispositions of all kinds – rational and irrational – in its constitutive clusters, some dispositions would presumably have to be weighted more heavily than others in determining whether a state is instantiated. That is, a constitutive cluster would have to have some more central dispositions and some more peripheral ones. And in order to maintain the priority claim, all that would be required is that rationality has a positive impact on the relative centrality of different dispositions.

The previously mentioned considerations can now be brought to bear in a slightly modified form that makes them nearly impossible to deny. It is, following Wedgwood’s suggestion, at least a fundamental aspect of cognitive states that they are abilities rather than liabilities, so irrational dispositions cannot play a major role in constituting them. Similarly, as I suggested, the implications of gained or lost irrational dispositions will always be more minimal than the implications of (at least some) gained or lost rational dispositions. And that is enough to establish an important asymmetry between rational relations and merely causal relations in an account of the mind.

With this priority claim in place, the question becomes: what can the functionalist make of it? Is there any way to account for this asymmetry without making explicit reference to rationality?

This challenge is still not quite enough to bring about the desired result. For one thing, there might actually be a fully naturalistic way to account for dispositions’ differing degrees of centrality – perhaps, as Wedgwood mentions, something along the lines of Fodor’s asymmetric dependence account. But even if no such account is forthcoming, I believe the functionalist can legitimately refuse to face the challenge at all. What is required of a functionalist account,

---

50 See Fodor (1990).
ultimately, is just a list of the relevant dispositions in different clusters along with their relative weights – once that is in place, nothing more needs to be said. The functionalist can admit that we rely on rationality judgments in coming to understand what goes into the lists, but metaphorically speaking, the lists are all that matter.

In essence, what the functionalist can do is treat rationality as a ladder that gets kicked away in the end. Rational relations, on this view, are an integral part of how we conceive of intentional states, how we get a grip on the precise causal connections that constitute those states, but they need not be seen as an integral part of the constitutive basis itself. This has been an ongoing theme with these attempts to establish the (horizontally construed) Normativity Thesis: it may be possible to show that there are open questions when rational relations are not mentioned, or that it is simpler to describe a state’s core causal connections in normative/rational terms, or that we most naturally understand a state’s causal dispositions as consequent on discharging requirements of rationality, etc., but the functionalist can insist that all these things are beside the point. As long as the functionalist makes no claim to be analytically reducing the concepts of different intentional states, what matters is just that there are naturalistically acceptable conditions sufficient to account for the presence of those states. And that possibility has not been ruled out.

There is, however, a key fact about this dialectic that has been overlooked: the functionalist can adopt this strategy only by assuming that the nature of intentional states is fixed. If, on the other hand, the nature of those states is dynamic, i.e. changing over time, then the strategy falls apart. Consider the metaphor of the ladder that gets kicked away. If a ladder can be used once to get to a stable resting point and then kicked away, then the parallel conclusion is fine – the ladder is not an essential part of being at that place, there might have been other ways to get there, etc. But if the resting point is not stable, and one must go back and use the ladder again and again, getting to slightly different points at different times, then it is not at all clear that the same can be said. Rather, the ladder appears to be an essential part of an ongoing process.

Recognizing that intentional states have dynamic natures, in my view, provides the key to making a successful argument for rational relations’ constitutive role.

3.3 Making the argument work: the dynamic nature of intentional states

I have been examining an argument for the claim that having horizontal causal relations in one’s constitutive account of intentional states necessarily brings with it a constitutive role for rational relations too. In my view, a
successful argument of this kind goes as follows. If one accepts that horizontal causal relations are part of what makes intentional states the states they are, then it follows that the nature of those states is dynamic (i.e. changing over time). And because intentional states are dynamic, it is possible to reformulate the idea of explanatory priority in the following way: over time, changes in a state’s rational relations have a direct impact on changes in its constitutive cluster of causal dispositions, but changes in a state’s causal connections have no such direct impact on its rational relations; in other words, a state’s constitutive causal relations track its rational relations. This means that rational relations must figure in an acceptable constitutive account of intentional states.

I turn now to a defense of this argument’s two main claims.

First, why is it that, once one has adopted a basically functionalist approach, the nature of intentional states must be seen as dynamic? The simple answer is: because an intentional state’s causal connections, in the aggregate, undoubtedly change over time. These changes result from a number of sources, including (i) certain kinds of worldly change, (ii) increases in knowledge, (iii) the emergence of new concepts, and (iv) conceptual change. I will discuss each of these in more detail shortly, but there is an obvious objection that should be dealt with beforehand.

The objection is that a functionalist can easily grant that some of an intentional state’s causal connections change over time while denying that the dispositions in the state’s constitutive cluster change. This kind of reply would be nearly automatic for those who believe there are just a few key dispositions – or even only one – that go into each constitutive cluster. Whatever plausibility this view seems to have comes, I think, from an all-too-common focus on logical constants as a paradigm case. It does perhaps seem possible to pick out a handful of dispositions that are absolutely crucial for possessing a concept like ‘and,’ but it should hardly be surprising that things are more complicated when it comes to concepts like ‘bison’ or ‘game’ or ‘loves’ or ‘expertly’ or countless, countless others. And once things get a bit messier, it is very hard to see how a line could be confidently drawn between the constitutive causal connections, which can never change, and the ancillary ones, which can.

51 This is the main example addressed by Wedgwood (2009), among many others.
52 My suggestion here clearly has an affinity with that in Quine (1951) regarding the analytic/synthetic distinction. Mine is much more modest, though, as long as one accepts the possibility of constitutive dispositions that do not reflect analytic connections. That is, Quine lays out a picture on which any given concept figures in a whole network of different judgments, and all of those judgments are potentially revisable. But even if that is incorrect, and there is sometimes an unrevisable core, my suggestion could still stand. All my claim requires is that for some concepts there are some dispositions in its constitutive base (which is to be found within the overall network) that are revisable.
One might be concerned that if too much of this messiness is allowed into
the very nature of intentional states, to the point that components of a constitutive
base can be gained or lost over time, then there is actually no option but to accept
a highly counterintuitive account of conceptual change. That is, one will be forced
to say that fairly slight changes in the causal behavior of a concept result in the
vanishing of the original concept and its replacement by a brand-new, slightly
different concept. After all, it might be said, if there really is a single concept
that persists through various changes, that must be because an essential,
constitutive core remains constant while inessential properties change.

This line of thought is not compelling, however. It rests on an overly
simplistic essentialism that fails to make room for genuinely dynamic entities.
Consider a biological species – an analogy I will have occasion to develop in more
detail shortly. A biological species, on one very common picture, is a dynamic
entity, changing over time as a result of various evolutionary processes. And
though there will always be some properties that stay the same from one stage of
the species’s evolution to the next, there is no core set of essential properties that
must stay the same through any of its possible evolutionary changes. Yet those
properties, in the aggregate, do still play a crucial role in determining the identity
of the species – if there are enough changes in those properties, it will be clear that
a new species has evolved (or that the original has split into two different ones).

This analogy is by no means perfect, but it is good enough, in my view, to
establish the possibility of a single entity persisting through change even though it
has no core of absolutely essential properties. If that is so, then there is no
obstacle to including a wide range of causal connections in an intentional state’s
constitutive cluster, and given the difficulty in most cases of finding one
disposition (or a small set) that defines a given state, there are often good reasons
to do so.

With that initial objection out of the way, I will now briefly discuss some
things that commonly lead to changes in an intentional state’s causal connections.
The first such source is changes in the world that bear on the inferences one can
justifiably make when a certain concept is involved. This then results in changes
in the causal dispositions associated with certain intentional states. Here are a few
examples:

- At one point, a belief that something was a dog would naturally dispose
  one to believe it ate a diet consisting primarily of meat; today (at least in
  many contexts) it would not. This is due to changes in the living

53 This is very much like the anti-holist claim that two people will rarely share the same concept because of
slightly different causal behavior in the two cases.
conditions of many dogs, especially their changing relationship with humans, etc.

- At one point, a belief that someone was from India would naturally dispose one to believe the person spoke no English; today it would not. This is due to changes in Indian society resulting from British colonialism, etc.
- At one point, a desire to travel from Boston to London would naturally dispose one to board a ship; today it would not. This is due to the advent of flying, etc.
- At one point, a desire to strike it rich by finding gold would naturally dispose one to go to California; today it would not. This is due to the fact that most of the gold in California has already been found, etc.

Clearly, examples like this could be multiplied ad infinitum. They show that worldly changes can lead to a change in an intentional state’s causal network over time. As long as there is no principled way to exclude these potentially changing causal dispositions from an intentional state’s constitutive cluster, it follows that intentional states have a dynamic character.54

Increases in knowledge can also give rise to alterations in an intentional state’s network of causal dispositions. The entities themselves (i.e. the intentional objects) do not change in such cases; rather, new connections between states are established (or old connections destroyed) as we learn new facts about those entities. Here again are a few examples:

- At one point, a belief that something was water did not dispose one to believe anything about its molecular composition; today it naturally disposes one to believe it is made up of H₂O molecules.
- At one point, a desire to improve a patient’s health would naturally dispose one to apply leeches; today it would not.
- At one point, a belief that something was a planet would dispose one to believe it had a circular orbit; today it would not.

And so on. Clearly, none of these causal dispositions are universal in nature; in some cases, it may be only a minority of people who are disposed to think or act in the way suggested. But it is nonetheless true that there have been substantial changes in the aggregate causal relations of those intentional states – and

---

54 A reminder: in saying that these states have a dynamic character, I am not saying the states themselves change (i.e. become different states) as a result of worldly changes any more than a species becomes a different species because one or two of its common properties change as a result of evolution.
particularly among the experts whose use plausibly grounds that of the population as a whole. The examples suffice to show, then, that the causal relations an intentional state figures in can change over time as a result of new gains in knowledge, and this too contributes to the dynamic nature of those states.

The third source of change is slightly different. Rather than altering causal dispositions that have previously been in place, the emergence of new concepts is a matter of new causal clusters being created. It is evident from the examples above that new concepts can also have important effects on causal networks already in existence – the concepts of hydrogen and oxygen, for instance, provide a new and important connection for intentional states about water – but the origins of these new clusters, these new nodes in the holistic intentional web, cry out for explanation in their own right. Ongoing scientific development provides a rich vein of new concepts, of course (e.g. covalent bond, gene, positron, etc.), but innumerable others arise in broader culture (e.g. suburb, mortgage, hip hop). The observation that new concepts emerge is in no way direct evidence that intentional states change over time, but it helps to supplement the overall picture of them as historical and dynamic entities.

Finally, in addition to those effects due to changes in the world and due to changes in our knowledge, there are the effects on intentional states’ causal networks due to changes in concepts themselves. Now, conceptual change is a slippery idea – it is often difficult to distinguish it from (a) changes in the meanings of terms (rather than the underlying concepts) and (b) cases of new concepts replacing old ones (rather than changes in a single concept); and even if it is distinguished from these things, conceptual change is often very closely intertwined with changes in the world and in our knowledge. But I hope the following examples can at least make it out to be a plausible possibility:

- At one point, discovering that someone regularly engaged in experimental inquiry (with data collection, etc.) would naturally dispose one to believe the person was a philosopher; today it would not.
- At one point, a belief that something was fine art would naturally dispose one to believe it was representational; today it would not (or at least not to nearly the same extent).
- At one point, a belief that two people had the same set of symptoms would naturally dispose one to believe they must have the same disease; today it would not (because the symptoms might result from entirely different causes).

55 This example is based on Thagard (1996). It is stems from the idea that diseases were once classified qualitatively, i.e. primarily by their surface characteristics, before subsequently being classified mechanically, i.e. by their underlying causal basis. Even if one doubts the specifics of this example, it at
These examples each exhibit a different kind of change: in the first, the range of application of the concept ‘philosopher’ narrows; in the second, the range of application of the concept ‘fine art’ broadens; and in the third, the basis for categorization for the concept ‘disease’ shifts. In each case, it is arguable, despite concurrent changes in the world and/or our knowledge, that the main change is in the concept itself – some properties that play a role in categorization are gained, lost, or given different relative importance. This has a decided impact on the nature of intentional states involving those concepts, for it changes the causal networks in which they are embedded.

All these factors, then, contribute to the dynamic nature of intentional states (on the assumption, of course, that horizontal causal relations are a crucial part of that nature). This leads directly to the second main claim of the argument: that rational relations have a kind of explanatory priority over causal relations. In order to see this, take any intentional state whose overall dispositional profile (in an individual or a population) alters over time. Those alterations could be

(i) inconsequential
(ii) serious enough that the state’s constitutive cluster of dispositions alters (though the concepts involved remain those same concepts), or
(iii) drastic enough that the state’s constitutive cluster alters because the involved concepts themselves change.

The question is, what determines which type of change it is that takes place? In particular, the key boundary is between option (i) and option (ii) – what determines whether there are changes in the state’s constitutive cluster as opposed to changes in how that same cluster is satisfied?

One reason a particular change might be of type (i) would be that it involved adding or subtracting an irrational (or perhaps rationally neutral) disposition. For example, say that one intentional state – a belief that someone has received a positive mammogram result, perhaps – gradually gains a tendency to give rise to another state – a belief that breast cancer in that individual is highly likely. As this change takes place, does the new disposition become part of

least seems perfectly conceivable that a shift of that kind – from a category determined superficially to one with a “hidden essence” – could take place.

---

56 See Carey (1992) for a slightly different but congenial classification of conceptual change.
57 I would contend that most intentional states are in this category (given a long enough time period).
58 This hearkens back to the discussion in Section 3.2.
59 This is a well-known example of the base rate fallacy, in which a positive result on a test known to be accurate leads to an unlicensed (i.e. greatly overestimated) conclusion about the probability of the event in
either state’s constitutive cluster? Or is it an utterly dispensable addition to the constitutive clusters already in place? What is required here, I contend, is an assessment of the new connection’s rational status.

If, as in this example, it is an irrational disposition that is gained, then that change is entirely inconsequential vis-à-vis the two beliefs’ constitutive bases. After all, those who do not have the new irrational disposition meet the qualifications for having the beliefs to the exact extent that they did (or would have) before—it is not as if they would now be missing a component, even a small one, such that their status might potentially be downgraded, e.g. from having the belief to lacking them or being borderline.

If, on the other hand, the new disposition reflects a new rational connection, then there might very well be a change in one or both beliefs’ constitutive clusters. It would no longer be clear, in such a case, that lacking the disposition would have no effect on one’s status with regard to having those beliefs. One’s status might remain the same, of course, because there are so many other factors at play in the overall determination, but it would certainly be conceivable that a lack of the new disposition could make a difference.

So, the indispensability of rational relations in a constitutive account of intentional states comes from this explanatory priority: changes in the rational relations an intentional state bears to other things explain the changes in its constitutive cluster of dispositions, while changes in a state’s causal connections, considered merely as such, do not. Given that intentional states have dynamic natures, then, any attempt to provide instantiation conditions for them without mentioning or requiring judgments about the rationality of their various connections is hopeless.

Now, there are certain ways a functionalist could try to resist this conclusion. For instance, when it comes to changes in the world or in our knowledge, it could be maintained that the possible changes are in fact already accounted for in the original network specifying the state. This is because the relevant network would have countless conditionals specifying how a state would combine with other possible beliefs. Here is one example: the network for the belief that a person is from India would include a clause saying that this state would tend to combine with another—the belief that many Indians speak English—to produce a belief that the person likely speaks English. This clause would be in

---

question. Given how common this fallacy actually is, it would be more realistic to start with the disposition in place and imagine that it is lost over time. The present example is chosen for ease of exposition.

60 In the specific example at hand, these clusters would presumably include dispositions to reach weaker but better-supported conclusions, e.g. that the likelihood of breast cancer in that individual has increased. The excessive and unjustified inference in the text is a well-known example of the base-rate fallacy.

61 This could be the case if either the mammogram changed so that it had very, very few false positives, or if the overall prevalence of breast cancer in the population increased dramatically.
the network from the start, and the changes in the world over time – British colonialism, etc. – would, by affecting other beliefs in the background, cause different clauses to kick in.

One might respond here along the lines of Brandom (1994), insisting that the inferences in a content-determining network cannot all be construed in a formal (rather than material)\(^{62}\) way, which is what this example seems to suggest. But I will not delve further into that issue, for in any case that sort of tactic will not help the functionalist with the other two sources of change, the appearance of new concepts and conceptual change. It could hardly be suggested that a state’s original, defining network included clauses anticipating all possible new concepts – that would stretch the notion of functionalism’s ‘in principle possibility’ beyond recognition – and conceptual change is more than a matter of change in background beliefs.

Take the earlier example of the concept *disease*. The alleged change is from a concept privileging surface features to one privileging underlying mechanics. (And again, if this example is not historically accurate, it at least represents a certain possibility). This is a change in how the concept itself operates, and it is simply not plausible that a shift in background beliefs – beliefs about the nature of the concept? – could be seen as accounting for it.

There is another kind of objection that could be offered against examples based on conceptual change, though: it could be claimed that there is no basis for the explanatory priority of rational relations in such cases. This is because the transition from one stage to the next – from the superficial notion of disease to the natural-kind notion, for instance – might be utterly arational. That is, it could be thought that in at least some cases conceptual changes happen for no good reason at all; there might just be two (or more) ways that a concept could equally well develop, and (random?) causal factors push it in one way rather than another.\(^{63}\)

This provides an opportunity for an important clarification. The key claim in the argument for explanatory priority is not that rational considerations account for such changes,\(^{64}\) but that they are essential in marking whatever causal changes

---

\(^{62}\) Material inferences are those that depend on the particular contents of the claims involved and contingent linkages between them (e.g. the contingent facts about language knowledge and nationality). The functionalist reply, on the other hand, treats the key inference as one relying strictly on form: (1) person X is from country Y; (2) people in Y tend to speak language Z; (3) therefore person X probably speaks Z.

\(^{63}\) In the case of the concept *fine art*, for example, it might be thought that non-representational work could just as well have been excluded from the concept as it developed over time – and whether it was or was not excluded might have boiled down to the disproportionate causal influence of a key critic, or the cult of personality around a particular artist or set of artists, etc.

\(^{64}\) Though they surely do account for some. Changes in the way a concept like *mentally abnormal* is applied might well result from rational reflection – I owe this example to John Campbell – but the point I am making in the text is a broader one. Rational reflection can be seen as one source for conceptual change among many, and the main argument is based on how those changes get handled, whatever leads to them.
become (relatively) fixed in the process. So even if a transition occurs as a result of purely causal features, the challenge for the functionalist is to say which causal changes make up that change – i.e. what goes into the updated constitutive cluster – and which causal changes are ancillary to it. When the concept disease changed, for instance, it is possible (though of course far-fetched) that two new dispositions simultaneously arose: (a) a disposition to believe, on the basis of a belief that two people have the same symptoms, that they might still have different diseases, and (b) a disposition to believe, on the basis of a belief that two people have the same symptoms, that it would be literally impossible to discover whether or not they have the same disease. On a plausible interpretation of the scenario, the first disposition reflected a rational connection and so should have become part of the constitutive clusters for the beliefs involved, while the second disposition was irrational – stemming from an excessive pessimism about the prospects of science, perhaps – and so should not have become part of the constitutive clusters. Focusing solely on the initial causal networks associated with those states and the causal changes that in fact occurred, however, the functionalist would have no way to make that assessment and, as a result, his or her theory would go off track.

Let us now take stock. The functionalist had resisted earlier attempts to spell out an essential role for rational relations by treating rationality as a ladder that could be kicked away. Yes, the functionalist said, the key dispositions (or at least the most central ones) do in fact mirror rational connections, but that fact need not be mentioned in a constitutive account, the sole purpose of which is to provide intentional states with naturalistic metaphysical grounds. It may even be true that, practically speaking, we have to think in terms of rationality in order to arrive at judgments about which dispositions matter for which states, but it is the dispositions alone – regardless of how they are discovered or conceived of – that do the metaphysical work.

That kind of reply is no longer available to the functionalist. It had appeared available only because the dynamic nature of intentional states was overlooked. Once that dynamic nature is recognized, though, it can no longer be said that dispositions alone serve in the constitutive account. The crucial causal connections do not just happen to mirror rational ones, as the functionalist said before; rather, they go on mirroring them, changing as they change.\textsuperscript{65} Rationality provides the only explanatory basis for a state’s ongoing dispositional changes, so the ladder simply cannot be kicked away.

In the end, the dynamic nature of intentionality brings with it a constitutive role for rational relations. And that means that a horizontal construal of the Normativity Thesis can indeed be sustained.

\textsuperscript{65} I do not intend this to suggest a temporal priority for rational changes, by the way. The priority is explanatory only.
4. Loose ends

It is important to remember that the argument laid out above has a qualified conclusion. It shows that if horizontal causal relations play a role in an account of intentionality, horizontal rational relations must play a role too, thus establishing a version of the Normativity Thesis. But what if an account of intentionality can eschew horizontal relations entirely? The jury is still out on this possibility, but I want to offer a few reasons for thinking it is a relatively bad bet. I cannot hope to settle this large and complicated issue, however, and to a certain extent I am happy to leave the argument in its qualified form – since acceptance of a role for horizontal relations is quite common, the result of the argument is relevant in any case.

An account that does without horizontal relations will generally be based on vertical relations instead. That is, its focus will be on the causal connections between a bearer of content (i.e. a mental state) and the objects and properties spelled out in that content. So, for instance, a belief about stars is what it is, on this kind of view, solely because of a causal link of a certain kind between it and stars. The most prominent examples of this kind of approach are Fodor (1990) and Dretske (1986).

Now, because this sort of account relies in no way on connections with other mental states – or the many connections with non-represented objects, properties, and actions a given mental state will have – it is decidedly non-holistic. Indeed, one motivation for a vertical approach is to avoid what are thought to be the serious problems that come with holism. To the extent, then, that the foregoing discussion has made those problems seem less pressing, it has already provided some support for a view involving horizontal relations. I will quickly mention two examples.

One putative problem is that, at least for many concepts, no particular horizontal connection is necessary for having it. But that fact is accommodated by using clusters of dispositions as constitutive bases, with only a sufficient number of those dispositions required to be present in order for the state to be instantiated.

A second putative problem is that different people (and the same person at different times) will rarely have the same concepts or intentional states because of the differences there are bound to be in the webs of causal dispositions realized by

---

66 One popular position I have not explicitly mentioned is teleosemantics. It is a somewhat curious case in that it does rely on horizontal relations, but not on current horizontal relations as dispositional functionalism does. For that reason, it does not fall within the scope of the main argument’s qualified conclusion. However, some of the considerations against vertical accounts that I offer below apply to teleosemantics as well, which I will note at the time. A fuller discussion meant to rule out teleosemantics as an option will have to be left for another occasion.
each. Again, this is not a problem if only a sufficient number of dispositions in a constitutive cluster need to be present for instantiation. As long as that is true for both people, they will have the same concept or state.

In the background here is the following thought: intentional categories are like many others—scientific categories like “living thing,” cultural categories like “game” (to give Wittgenstein’s famous example)—in that multiple properties are relevant to category membership, none are strictly necessary, and yet the category determinately applies to different instances all the same.

Avoiding holism, then, need not be seen as an advantage in itself. But more importantly, accounts based on vertical relations have serious problems of their own. The problems tend to have a common feature: as a result of restricting their content-determination resources so severely, these accounts are unable to capture many of intentionality’s nuances. Again, a couple quick examples will have to suffice.

First, because they are extensional in nature, vertical causal relations alone seemingly cannot distinguish contents in a way that is sufficiently fine-grained. A single object can be conceptualized or grasped in multiple ways, but a causal relation between an individual and the object is the same no matter what. This problem is commonly illustrated by means of cases of nomologically linked properties, e.g. ‘having a heart’ and ‘having a kidney,’ ‘being a fly’ and ‘being a small, black, flying object,’ etc. It is not clear how an examination of vertical causal relations between an individual and objects could possibly distinguish such properties.

Or consider a different kind of example: our concept ‘water.’ It is accepted by many that, even before the discovery of H₂O, this was a natural kind concept, i.e. one allowing for an underlying essential characteristic that explains the surface features. But even if that is right, it would certainly have been possible for us to have had a different kind of concept—we could have operated with a superficially defined category akin to ‘jade,’ something roughly equivalent to ‘watery-seeming stuff.’ Presumably, though, the causal connections between us and water (and water-imposters) would have been the same in both cases. The differences between the concepts in the two scenarios would seemingly have had to reside in their broader causal characteristics—how they were used, more generally—and that makes adding horizontal relations into the account look like a natural remedy.⁶⁷

---

⁶⁷ This sort of problem arises for teleosemantics as well, but with function assignments as the culprit instead of vertical causal relations. That is, it can be asked how a notion of Darwinian selection could be used to specify one property rather than another when (a) the two properties always occur (or have always occurred) together, and (b) it is the whole package (i.e. the object with both properties) that provides the adaptive benefit. Even if, as the enormous literature on this topic might plausibly be taken to suggest, teleosemantics has a reasonable chance of answering this challenge, the example of the ‘water’ concept is
It appears, then, that a vertical approach has indeterminacy where there should be none. But I would argue that it also lacks indeterminacy where there should be some. I believe it is a virtue of a holistic and flexible view involving horizontal relations that it permits borderline cases. This seems to reflect our best judgment of what is going on when, for example, a person exhibits a fairly large degree of irrationality. A little bit of irrationality is not too troublesome, and we are happy to attribute a determinate content, shareable with others; a huge amount of irrationality tells quite decisively against the person having an intentional state with that content; but in a certain zone in between, we are simply not sure what to say, and that does not strike us as always a matter of lack of information. Moreover, the same kind of borderline judgment seems to be called for in cases of conceptual learning (e.g. in children) and in intentional attribution to non-human animals. A large holistic network as constitutive base provides a nice way to understand the partial grasp of a concept or the stunted realization of a mental state. But if a vertical causal connection is all that matters, things will be too clear cut. If the connection is in place, then the concept is too – fully so, regardless of how it gets put to use – and if the concept is joined together in a certain way with another, a given intentional state is present – fully so, regardless of its connections with other states. This implication is not quite self-defeating, but it certainly fits less well with our intuitive judgments than one might hope.

These considerations all favor an approach that makes use of horizontal causal relations, at which point the main argument kicks in.

One final loose end remains, and it concerns the implications this chapter’s main argument has for naturalism in the philosophy of mind.

To this point, my focus has been on whether the Normativity Thesis – the claim that normative facts or properties play an ineliminable role in an account of intentionality – can be established. In Chapter 1, I argued that the thesis could not be established if it was construed as involving vertical norms, i.e. the representational “norms” of truth, correctness, etc. The problem was that, despite the common label, these are actually not best thought of as normative properties. In the present chapter, I have argued that the thesis can be sustained with horizontal norms, i.e. norms of rational connection, at its heart. That conclusion depends on an antecedent acceptance that horizontal causal relations play a role in an account of intentionality, but acceptance of that is quite common (and I have offered a few considerations in its favor as well).

Establishing the Normativity Thesis means setting up a *prima facie* case for the Normativity Objection, the claim that a naturalistic account of intentionality is still pressing. It is exceedingly likely, after all, that our evolutionary history could have been the same in either scenario (i.e. that with a natural kind concept vs. that with a superficial concept). If that is right, then teleosemantics is no better equipped to solve the problem.
not possible. In this instance, that *prima facie* case stems from two facts: (a) that rational relations cannot be seen as disguised descriptions\textsuperscript{68} and (b) that the need for rational relations stems from the limitations of causal functionalism.\textsuperscript{69} In the next chapter, I will examine that case in greater detail. Taking the horizontal construal of the Normativity Thesis as a given, I will show how it can be used to fill in a notoriously unsatisfying argument schema – due to Davidson – with strong anti-naturalistic conclusion. I will also show, however, that the success of that anti-naturalistic argument depends on how naturalism is understood. This will initiate an examination of different possible construals of naturalism, the bulk of which will take place in Chapter 4. Ultimately, I will argue that when naturalism is understood in the independently most plausible way, the Normativity Thesis outlined here does not preclude an account meeting its standards.

\textsuperscript{68} This was shown in Section 2.

\textsuperscript{69} This was shown in Section 3.3.
Chapter 3

The Dynamics of Intentionality and the Argument against Naturalism

1. Introduction

A distinctive feature of intentional states is that they figure in rational relations – beliefs are warranted by the evidence (or aren’t), desires help to provide reason for an action (or don’t), and so on. This is a fairly uncontroversial observation about mentality, but many have found it tempting to go a step further and to claim that intentionality constitutively depends on the norms of rationality. On this sort of view, standing in certain rational relations (telling in favor of x, being justified by y, etc.) is an essential part of what makes particular intentional states the states they are.

Now, much can be said about why one should or should not take this first step, and the Chapter 2 was devoted to exactly that issue. After showing how other attempts to establish the constitutive claim have fallen short, I offered a way to make a successful case for it. This case required an important recognition about intentional states – that they have dynamic natures – as well as the quite plausible, commonly held assumption that horizontal causal relations have a role to play in a constitutive account of intentionality. Together, those two facts are enough to show that rational relations must enter into such an account as well.

In the present chapter, I want to bracket that specific argument for the constitutive claim – at least at the start – and focus on the next argumentative step it is tempting to take: to argue that a constitutive dependence on the norms of rationality precludes a naturalistic account of intentionality. This second step is tempting because rational connections seem different in crucial respects from the things in the naturalist’s basic theory-building toolkit (e.g. facts about the brain or about behavioral dispositions). As McDowell (1994) evocatively puts the point, if
intentional states can only be understood as occupants of the “space of reasons”, while naturalism deals strictly with the “space of natural law,” then any thoroughly naturalistic approach to intentionality is apparently ruled out.

Unfortunately, even if one goes along with the first step (at least for the sake of argument), the legitimacy of the second step is not obvious. As I have sketched it so far, the argument from rationality’s constitutive role to an anti-naturalistic conclusion is elliptical, and the most common ways of filling it in do not, in my view, manage to make the argument valid. What is needed is a clear link between premise and conclusion, and I believe the best candidate for that link is the same previously overlooked feature of intentionality that was so crucial in the previous chapter. What I will argue, in brief, is the following: it is because intentional states are dynamic (i.e. changing over time), and because the norms of rationality play an ineliminable role in accounting for those ongoing changes, that one cannot use naturalistic resources alone to provide a satisfying account of intentionality.70

An emphasis on the dynamic character of intentional properties stands in contrast to standard approaches to this issue. Attempts to naturalize intentionality generally try to provide (naturalistically specified) sufficient conditions for the instantiation of an intentional property at a given time; it is implicitly assumed, seemingly, that whatever the grounding relationship is, it will remain stable over time. And attempts to use facts about rationality to resist such naturalization have followed suit in this regard, arguing, for example, that static instantiation conditions cannot be specified for intentional states without relying on non-naturalistic rationality facts. I contend that a shift in perspective is called for: it is important to recognize the dynamic character of intentional states both for its own sake – i.e. as part of a fuller understanding of the nature of intentionality – and because it is only then that the anti-naturalistic implications of rationality come into focus.

2. The Davidsonian argument (and two unsuccessful attempts to make it work)

The basic argument I am concerned with here is, clearly enough, a broadly Davidsonian one – an early and very well-known presentation of the argument can

---

70 This chapter and the previous one will have shown, then, that the dynamic nature of intentional states is the crucial factor in establishing both the constitutive role for rationality and its anti-naturalistic implications. That it does double duty in this way is not surprising, perhaps, but it is not completely trivial either – it might well have turned out that different considerations were fundamental at the two main stages of the overall argument. For example, rationality’s constitutive role might have been based on interpretationist grounds, say, while the anti-naturalistic implications stemmed from the naturalistic fallacy or indeterminacy issues. See below for more details on these latter options.
be found in Davidson (1970). My aim is not to provide an accurate interpretation of Davidson’s own views, but to see if an argument of that basic shape can be made to work.

Davidson’s version of the main premise is put in terms of a “constitutive ideal of rationality” that governs the application of intentional predicates. But this claim need not be thought of as tied together with other distinctively Davidsonian – and perhaps dubious – assumptions, such as a strong interpretationism about the mind or a commitment to the indeterminacy of translation. For our purposes, the argument’s main premise (the Rationality Premise, for short) can be taken to be the following:

(RP) Our conceptions of intentional states/properties essentially involve facts about rationality or rational connections.

This premise is somewhat schematic, of course – the project of spelling the argument out is, in part, the project of clarifying and specifying the nature of the involvement mentioned – but it is worth noting now that the premise is conceptual or intensional in nature. That is, it concerns how we understand or think about intentional states, and not (at first blush) their underlying metaphysics – more on this point shortly.

Davidson’s version of the conclusion states that there are no strict psychophysical laws. Again, this claim can be generalized slightly – most importantly, so that the argument can be seen as applying to functionalist/dispositionalist forms of naturalism as well as those that seek direct connections between the intentional and the physical (in most cases, the brain). But there are two features of Davidson’s anti-naturalist conclusion that are to be preserved.

First, the conclusion is quite strong: the essential dependence of intentionality on facts concerning rationality is taken to rule out empirically discoverable connections between the physical (or functional) and the psychological, not just analytic reductions. In other words, the conclusion is metaphysical or extensional in nature – Davidson presumably intended it to apply to the type-identity theories popular at the time, but it is now generally taken to apply to any attempt to provide naturalistic instantiation conditions for intentional states, even those with no pretense of capturing everything every aspect of our concepts of those states. Indeed, one of the main reasons the Davidsonian argument is so potentially powerful – and so difficult to make work – is that it moves in this way from a conceptual/intensional premise to a metaphysical/extensional conclusion.
The second feature I’d like to highlight is this: the conclusion makes explicit that what are ruled out are *projectible* connections, i.e. connections that, unlike one-off token identities (or, arguably, supervenience relations), can be applied to various other cases at other times. Projectibility is clearly a property of scientific laws, but even if we eschew that latter notion (e.g. to avoid any controversies about its precise definition), projectibility is surely also a property of the kind of explanatory metaphysical theory naturalistic accounts aspire to be.\(^{71}\)

As such, the argument’s conclusion (Anti-Naturalism, for short) can be taken to be the following:

\[(AN) \text{ Intentional states do not have instantiation conditions that are both projectible and naturalistically specifiable (e.g. physical, causal, etc.).} \]

I will argue that the role rationality considerations play vis-a-vis the dynamics of intentional states can account for a strong anti-naturalist conclusion of this form, while other proposals for completing the argument cannot. I begin with a discussion of the two most common alternative proposals: the claim that the argument turns on the indeterminacy of intentional state attribution, and the claim that it involves an instance of the naturalistic fallacy.

2.1 *First proposal: indeterminacy*

One way of trying to complete the argument does rely on another of Davidson’s notorious positions: the claim that the attribution of intentional states is irrevocably indeterminate. Of course, many who find the outlines of the argument congenial would prefer not to be committed to that position, but the argument may turn out to hinge on it all the same.

Several commentators\(^{72}\) take Davidson himself to rule out this interpretation explicitly when he says the following:

\[(Q1) \text{ Nor is the irreducibility due simply to the possibility of many equally eligible [translation] schemes, for this is compatible with an arbitrary choice of one scheme relative to which assignments of mental traits are made.} \]

*(Davidson 1980: 222)*

---

\(^{71}\) It is useful to understand the aim of naturalistic accounts along the lines suggested by Boghossian (1991): the aim is to *explain* the supervenience of intentional facts on natural facts, since otherwise the supervenience relation is left as a brute, mysterious connection between two disparate families of properties.

That does sound quite conclusive, but earlier in the same paragraph Davidson also says that

(Q2) the heteronomic [i.e. non-lawlike] character of general statements linking the mental and the physical traces back to this central role of translation in the description of all propositional attitudes, and to the indeterminacy of translation.

(222, my emphasis)

And a short time later he adds that

(Q3) a right arbitrary choice of a translation manual would be of a manual acceptable in the light of all possible evidence, and this is a choice we cannot make.

(223)

So the option of making an arbitrary translation-scheme choice, which in (Q1) seems to render indeterminacy a non-concern, appears in (Q3) not to be a genuine option after all. Perhaps, then, indeterminacy really is the problem; it is because we cannot fix on one intentional interpretation that there will have to be “nomological slack” – or, we might say, “metaphysical theoretic slack” – between the intentional and the physical.

Importantly, there is still a clear role for the Rationality Premise to play on this interpretation, for rationality’s involvement leads to the indeterminacy. Davidson holds that our attribution of intentional states is holistically answerable to an ideal of rationality, and that it is not possible to satisfy that ideal completely. Consequently, “many theories will effect a more or less acceptable compromise, and between these theories there may be no objective grounds for choice.” So the argument, on this interpretation, proceeds as follows:

1. RP
2. This results in indeterminacy
3. Indeterminacy entails “metaphysical theoretic slack”
4. Therefore, AN

As it stands, there is a gap in the reasoning from (1) to (2i) – how exactly is it known that there will not be one option that best lives up to the ideal? – but even if (2i) is granted, questions arise concerning step (3i). The Davidsonian indeterminacy claim maintains that for any given thinker, and any given pool of evidence, there will be multiple complete sets of intentional state attributions – full intentional psychological theories for that thinker – that are equally acceptable. That suggests that there might be two kinds of indeterminacy, one that arises from lack of evidence and one that does not. The examples Davidson offers tend to be
of the former sort, such as a person saying, “that’s a shooting star,” who might (a) mean “star” literally and have a false belief about stars, or (b) mean meteorite by “shooting star” and have all true beliefs about stars.\(^{73}\) As Davidson notes, “additional evidence may resolve this case.”

But, he goes on, “there will always be cases where all possible evidence leaves open a choice.” When discussing this latter kind of case, Davidson often compares the indeterminacy involved to that resulting from different measurement scales (e.g. an object’s length being both 1 (inch) and 2.54 (centimeters)).\(^{74}\) In other words, the indeterminacy is innocuous, entailing no deep incompatibility between the different options. The analogy could be thought of like this: first, there is in both cases a single pattern to be captured (a set of behavioral facts on the one hand, a set of length relations on the other); second, there is a single constraint on how it is to be captured (the ideal of rationality, and the relations constituting the real number system); and third, there are multiple ways of satisfying that constraint while capturing the pattern equally well (different whole intentional theories, and different scales of measurement).

If this view is correct, though, then (Q1) seems perfectly apt. One of the equally acceptable schemes could be selected, and there would be no principled reason why strict, projectible connections could not be found between intentional states \textit{as attributed on that scheme} and physical (or, say, dispositionally characterized) states. And, given the holistic character of those interpretational schemes, any connections found on one of them would presumably be discoverable, in systematically modified form, on them all. So the kind of indeterminacy that remains even after all possible evidence is in does not entail any “metaphysical theoretic slack.”

Indeed, in moving from (Q1) to (Q3), Davidson’s aim seems to be precisely to shift attention from this sort of indeterminacy to the kind owing to lack of evidence. Just before (Q3), for instance, he emphasizes that an intentional psychological theory for a thinker “must be an evolving theory,” one that we adjust “as the evidence accumulates.” That is clearly true, as is the claim in (Q3) that we can never in fact choose a translation manual that we know to be acceptable in light of \textit{all} possible evidence, since we never actually have all that evidence. But it is difficult to see how that is relevant to the main point. As long as there exists a set of interpretation schemes, each of which is compatible with the entirety of possible evidence, then it remains possible that there are connections between the intentional facts, according to any of those schemes, and naturalistic ones.

\(^{73}\) The example is from Davidson (1973).

\(^{74}\) For comparisons of this kind, see Davidson (2001: 79, 145, 214).
Of course, we might not be able to discover these connections, simply because we can never hit upon one of the ultimately justified interpretation schemes. The upshot would then be that in practice we must always leave some “metaphysical theoretic slack” because our intentional attributions can shift – as a result of incoming rationality-based evidence – even once we have settled upon a naturalistic description of the situation.

For example, say it has been established that a person (at a given time) is in a physical state consisting of neurophysiological facts \( N_5, N_{14}, N_{76} \), etc., and on our current interpretation, the person (at that time) is taken to have intentional states \( I_9, I_{42}, I_{83} \), etc. We might then go looking for correlations, perhaps speculating that \( N_5 \) is part of the instantiation conditions for \( I_9 \). What premise (3i), on this current understanding, points out is simply that new evidence could lead to a new intentional interpretation that is incompatible with the first one (e.g. one on which \( I_9 \) is replaced with \( I_{20} \)), even though our view of the physical facts is unaffected. If, say, the person being interpreted says or does something unexpected, it might then require us to attribute different intentional states (retrospectively) in order to hew as closely as possible to the ideal of rationality. This ever-present possibility means that we can never take ourselves to have found a firm correlation between the two domains. It is not merely that there might be future disconfirming instances, as is true of any projectible metaphysical principle, but that the allegedly confirming instances so far observed are too unstable – those links can go up in smoke at any time as our intentional interpretations evolve.

This way of interpreting the argument fits nicely with Davidson’s expressed concern that each domain “retain allegiance to its proper source of evidence.” And the worry it spells out does seem to be distinctive of mind-body connections. But there are at least three things that can be said in response.

First, this interpretation of the argument does not license a conclusion with the anti-naturalist strength of (AN). It would not permit one to say that there are no instantiation conditions of the requisite kind, as (AN) purports to do, but only that we are stuck with too little evidence to find out what they are. Of course, this conclusion wouldn’t exactly make the naturalist happy, but it would not undermine his or her basic commitments either.

Second, it is not entirely clear that even this weakened conclusion can be sustained on the grounds offered. In the example considered above, the problem was that the initial candidates for nomological correlation (e.g. \( N_5 \) and \( I_9 \)) disappeared as the intentional interpretation of the subject evolved. But consider the example in another light. Imagine that a rough correlation had previously been observed between \( N_5 \) and \( I_{20} \). In that case, the initial intentional interpretation (which attributed \( I_9 \)) would have seemed to present a counterexample to the correlation, while the later interpretation – based on more and better evidence –
would have turned out to reaffirm it. And perhaps some other cases that seemed to be counterexamples, which was why the correlation was deemed rough in the first place, would end up getting resolved in a similar fashion as more evidence came in. In this way, it is possible that we could see ourselves as making progress, approaching sufficient evidence for a metaphysical connecting principle of some kind.

This scenario might be unrealistic in actual fact, but it’s hard to see why it must be so. And matters are made worse (or better, depending on how you look at it) by the very flexibility of interpretation that Davidson has emphasized. For, continuing with the example, we might start looking at the remaining counterexamples to see if alternative intentional interpretations preserving the putative physical-intentional correlation are in fact available. We would not be “changing the subject,” as Davidson suggests, but simply considering one more factor as we work out intentional attributions in light of the ideal of rationality. In some instances, we might find accommodating interpretations – i.e. interpretations on which the “right” intentional state (in this case, $I_{20}$) is present – that are perfectly acceptable as things stand. In others, we would presumably find conditionally acceptable interpretations – e.g. the person could be reasonably seen as having $I_{20}$ if they also had $I_{72}$, for which there is currently insufficient evidence. And these latter cases could lead to targeted probing for further evidence, and so on.

This process would obviously get very complicated, but that is to be expected. The important thing is that its possibility tells against Davidson’s claim that if “we were to stumble on a nonstochastic true psychophysical generalization, we would have no reason to believe it more than roughly true.” This is surely right for some ways of stumbling upon true psychophysical generalizations, but if this generalization (i.e. the one connecting $N_5$ and $I_{20}$) is in fact a true one, and we stumbled upon it in the way outlined above, then it is at least an open question what we would have reason to believe about it. I would suggest that the following Davidsonian point applies: “the grounds for deciding to trust a statement on the basis of its instances will...be governed by theoretical and empirical concerns not to be distinguished from those of science” (216). Psychophysical generalizations were supposed to constitute an exception to this, but that has now been called into question.

Third and finally, the whole situation changes if the naturalistic base is thought of not as physical properties (as in the previous three paragraphs) but, more plausibly in my view, as causal-dispositional properties. The only reason the evolving-theory model seemed to make trouble for metaphysical connections was because it looked like one set of facts could change while the other did not. But the very evidence that can come in and change our intentional attributions, on the
indeterminacy view, is evidence about causal dispositions. So if causal dispositional facts are what lie on the other side of the proposed metaphysical connection, then there would no longer be any asymmetry.

On this sort of functionalist approach, the example given above would have to be modified. The proposed instantiation conditions for the intentional state \( I_{20} \) would be something like this: if the subject has \textit{most} of the causal dispositions \( C_2, C_{34}, C_{57}, \text{etc.} \), then the subject has \( I_{20} \).\(^{75}\) When the new evidence comes in – i.e. the subject does or says something unexpected, as I described it above – that changes an assessment of the subject’s causal dispositions – perhaps now they are taken to have \( C_{44} \) instead of \( C_{34} \) – and so can change a verdict about whether the subject satisfies the instantiation conditions for \( I_{20} \). The important point here is that putative metaphysical linkages between causal-dispositional facts and intentional facts are not problematized by the ongoing evolution of intentional attribution, and that is because the attribution of causal dispositions evolves right along with it.

I conclude, then, that indeterminacy fails to help the schematic Davidsonian argument go through. The non-evidential kind of indeterminacy, like that due to different measurement scales, is no obstacle; an arbitrary choice of interpretation scheme can be made, and any laws that hold will transfer from one scheme to the next. And the kind of indeterminacy that arises from lack of evidence at best yields a practical pessimism, not a principled anti-naturalist conclusion.

\textit{2.2 Second proposal: normativity and the naturalistic fallacy}

According to the next proposal for completing the argument, it is crucial to recognize that rationality is a \textit{normative} notion. Davidson himself does not emphasize this fact, and indeed some things he says suggest that he does not think normativity is a decisive factor.\(^{76}\) But interpreting the argument as making an appeal to normativity is popular because it seems to offer an independently plausible route to the anti-naturalist conclusion. It taps into the more general thought that there is a naturalistic fallacy, that normative notions are \textit{sui generis} with regard to naturalistic ones, and it treats the Davidsonian argument as simply an application – though perhaps a surprising one – of that general thesis to the domain of intentionality.\(^{77}\)

So on this rendering, the argument goes something like this:

\(^{75}\) This is clearly a crude simplification – there would presumably be some kind of weighting applied to the different dispositions in the list, and the subject would in some sense have to approximate having that set of dispositions, though it probably wouldn’t be as simple as adding them up – but it should do for illustrative purposes.

\(^{76}\) See Davidson (2001: 254).

\(^{77}\) McDowell (1985, 1994) is probably the clearest proponent of this version of the argument; Kim (1985) also emphasizes the normativity of rationality.
Rationality is normative

Normativity is *sui generis* (a.k.a. the naturalistic fallacy)

Therefore, AN

The first concern one might have with this proposal is that it simply postpones the explanatory demand. The naturalistic fallacy strikes many as quite intuitive, but it clearly stands in need of further support. It needs to be explained exactly *why* normativity precludes naturalistic treatment, and none of the traditional suggestions is without difficulties. I will briefly discuss a couple of examples.

One potential avenue of support is noncognitivism about normative judgments – the claim that such judgments express mental states other than belief, e.g. states of endorsement, or contingency plans, or something along those lines. This view makes sense of the naturalistic fallacy, because if normative judgments are not descriptive in nature at all, then their content cannot be captured by any “redescription” in naturalistic terms. Unfortunately, noncognitivism has well-known troubles accounting for the fact that normative judgments can be embedded in standard fashion in larger truth-functional contexts.

Another option would be to claim that while normative judgments do describe and state certain objective facts – i.e. they are to be understood in a cognitivist way – they are distinctive in being inherently motivational judgments. This is generally taken to result from the special kind of fact – normative facts – that these judgments concern. The thought is that upon judging that such a fact obtains, a person is immediately motivated to act in some appropriate fashion. Fully naturalistic characterizations lack that inherently motivational aspect, of course, so the naturalistic fallacy follows accordingly. This proposal strikes many as a fanciful invocation of *queer* properties (a la Mackie (1977)), however, and in any case it proves difficult to specify this idea of immediate or inherent motivation in a way that is plausible but also substantive enough to underwrite the naturalistic fallacy.

We can set these concerns aside, though, granting for the sake of argument that there is a well-grounded naturalistic fallacy, and still the appeal to normativity does not suffice for establishing the Davidsonian conclusion. The transition from (3n) to (4) remains problematic because the naturalistic fallacy, on its most plausible interpretation, rules out *analytic reductions* of the normative to the natural while leaving open the possibility of *a posteriori* discoverable connections between the two. For example, it could turn out that events with certain physical (or otherwise naturalistic) properties are always accompanied by certain normative
facts. Since the Davidsonian conclusion denies precisely this kind of empirically discoverable connection, the argument clearly requires something stronger in step (3n) than the naturalistic fallacy.

Moreover, it should be noted that a direct appeal of this kind to the naturalistic fallacy only makes sense for certain very strong interpretations of RP. That is, RP would have to be understood as implying that intentional properties simply are rational properties or are directly constituted by them, in which case the impossibility of naturalizing normativity would immediately carry over to intentionality too. The premise has been defended in this form (e.g. by Bilgrami (2004) and Zangwill (1998)\textsuperscript{78}), but it certainly need not be taken that way. If instead the premise means that our way of conceiving of intentional states involves recognizing that certain norms of rationality apply to them (as in Boghossian (2003)), or that the norms of rationality serve to outline the set of causal dispositions that go into an intentional state’s instantiation conditions, then the relevance of the naturalistic fallacy is not nearly as clear. Let me briefly consider both examples.

On the first reading of the premise, normative judgments must be part of a total understanding of a given intentional state as the particular state it is. The naturalistic fallacy, in that case, would at best establish that our grasp of intentional states necessarily involves judgments that cannot be construed as concerning naturalistic facts alone. But the naturalist interested in providing metaphysical instantiation conditions for intentional states will not be deterred by this. There can be intentional states present whether they are understood as such or not, and the naturalist’s theory is meant to say when and where those states are instantiated. To put it colorfully, an intelligent being who, for whatever reason, is incapable of making normative judgments could still pick out intentional states extensionally if equipped with the right naturalistic theory. The naturalistic fallacy, applied only to the judgments that go into our total understanding of intentional states, has no bearing on this possibility.

On the second reading of the premise, the norms of rationality – again, perhaps rendered fully irreducible by the naturalistic fallacy – serve as an ideal by means of which the instantiation conditions of intentional states are determined. More specifically, a particular intentional state is instantiated, on this kind of view, when a subject’s actual causal dispositions sufficiently approximate the set of causal dispositions that mirrors the normative ideal. For example, if the norms of rationality imply that a belief that the moon is made of cheese justifies a belief that the moon is edible, then the disposition to transition in thinking from one to the other would be part of the larger set of dispositions making up the instantiation

\textsuperscript{78} It is subsequently weakened in Zangwill (2005) and Zangwill (2009).
conditions for each of those beliefs. Our conceptions of intentional states could be seen as essentially relying on normative judgments, and we would have to use those judgments to fix on a state’s metaphysical instantiation conditions, but the conditions themselves would be causal in nature. As before, a being with no notion whatsoever of the normative could still use the proffered instantiation conditions to pick out intentional states in the world.

It might be thought that this last point goes a bit too fast, for isn’t there a crucial sense in which the instantiation conditions depend on the norms of rationality? The challenge is to show that this dependence is indeed metaphysical rather than conceptual. It may be more-or-less impossible for us to imagine coming up with the right set of dispositions except by using the norms of rationality as a model, but that could be viewed as a truth about our access to the relevant facts rather than the facts themselves.

It has been suggested that the dependence could be spelled out as follows: a state’s instantiation conditions could never be completely specified in naturalistic terms, even in principle, simply because the norms of rationality are so open-ended, allowing for myriad defeating conditions, excusing conditions, caveats, etc. In other words, there would be no way to list out all the different background conditions that could make a particular transition – say, one from a belief that the streets are wet to a belief that it rained recently – irrational.

It is certainly true that the normative principles of rationality applying to a given case are dizzyingly open-ended, and that our imaginations are too feeble to come up with every possible contingency before it actually confronts us. But what this implies – echoing the discussion of Davidson’s view in 2.1 – is that our metaphysical theory will, as a matter of fact, need constant adjustment as more evidence comes in. If we were to give our mythical being a set of instructions based on our best current model, a time would surely come when our own reliance on normative judgments yielded a verdict while that mythical being was left at a loss. But what is not clear is that a metaphysical theory could not in principle reach a point when, dizzyingly complicated as it might be, it would cover all the cases about which we too would reach clear verdicts. The only scenarios still left out of the being’s extension-determining instructions might be ones that would leave us normatively flummoxed as well. So this route to establishing a metaphysical dependence on normative judgments does not quite work.

In the end, the same basic point recurs in these two cases: though the normative principles of rationality come into the picture as an essential part of our thinking about intentionality – as a necessarily accompanying judgment of some kind – it is not possible to establish a link between this conceptual fact and the

---

79 A point like this is made in Wedgwood (2009).
underlying metaphysics. With that, I hope to have made it convincing that the naturalistic fallacy – a straightforward appeal to the normative character of rationality – does not complete the Davidsonian argument in a satisfying way. It is time, now, to offer a proposal that does so.

3. A new approach: the dynamic nature of intentional states

Let me begin with a quick recap. Having accepted, for various reasons, that our concepts of intentional states essentially depend on facts about the rational relations in which those states figure, many philosophers have been tempted to reach a strong anti-naturalist conclusion as a consequence. That is, they have found a broadly Davidsonian argument compelling. Such an argument moves from a premise about rationality’s constitutive role in our understanding of intentionality to a conclusion ruling out naturalist metaphysical theories of intentionality, but its intervening steps are pretty much up for grabs. I have tried to show that the two most common attempts to fill in the argument are unsuccessful. Basing the argument on indeterminacy, as Davidson himself perhaps does, runs into this problem: if the indeterminacy is merely a kind of notational variation, then it has no deep metaphysical implications at all, and if it results from incomplete evidence, then the required ongoing revision of the intentional theory will be coupled with revision of the metaphysical theory too. Basing the argument on the naturalistic fallacy, on the other hand, leads to great difficulties in bridging the fundamental gap between the argument’s conceptual/intensional premise – dealing with how we understand or think of intentional states – and its metaphysical/extensional conclusion – dealing with the states’ instantiation conditions.

Nonetheless, I believe that the Davidsonian argument can be spelled out in an adequate way, and the ingredients for doing so are, in a sense, already contained in or hinted at by the attempts already examined. In particular, the idea that our intentional theory is always evolving is crucial – but the theory should be seen as evolving not only because new evidence continues to come in but because the very entities the theory deals with (i.e. the intentional states) evolve. The basic thought here, though it certainly requires explication, is not meant to be particularly radical. It is simply the recognition that human intentionality is not static: new concepts are continually being introduced and developed, new connections between concepts are continually being discovered, and, perhaps more controversially, the boundaries of concepts are continually being stretched or narrowed or otherwise modified.

To give an analogy, the notion of an intentional state is akin to the notion of a biological species. Just as it has been found necessary, post-Darwin, to shift
from a static essentialism about species to a view of species as historical and
dynamic, I contend that the discussion of intentionality has been implicitly stuck
in a static essentialist mode and needs to change along the same lines. The
argument between those seeking to naturalize intentional states and those resisting
such naturalization has been a bit like an argument between a group who takes a
particular species to be defined by, say, a fixed set of physical and behavioral
characteristics and a group who takes it to be defined by some crucial non-
physical property standing behind those superficial characteristics. But that way
of formulating the options is simply not workable. The identity of a species
cannot be grounded in its common observable characteristics – size, coloration,
habitat, diet, reproductive behavior, etc. – for that simply fails to allow for
evolutionary change. Depending on the selective pressures present – as well as
random processes like genetic drift – any of these features could be different at a
future point in the species’s history. A species is something that changes over
time, and any satisfactory characterization of it must account for that.

Interestingly, though, the analogy can be seen as breaking down – or so I
shall argue – once the transition to a dynamic understanding is made. That is,
while the dynamics of species are explicable in strictly naturalistic fashion, thanks
to molecular genetics and the mechanism of natural selection, the situation with
intentionality is arguably quite different. The norms of rationality, which are
central to the non-naturalistic alternative, plausibly turn out to be essential to any
theory of how intentional states change over time.

Here is where the component from the second attempt discussed above –
that seeking to use normativity to ground the anti-naturalist conclusion – comes
into play. I discussed there a view on which normative judgments about rational
connections were used to pick out the clusters of causal-dispositional properties
required for the instantiation of intentional states. This view strongly suggested
that the metaphysical criteria for intentionality ultimately depended on the
normative, and that there was an in principle impossibility of spelling out the
criteria in naturalistic terms. Given a static conception, these worries could not be
made out satisfactorily – the normative element, crucial though it may be to our
understanding of intentionality, could always be kicked away when the focus
became brute, underlying, extensional metaphysics.

On a dynamic conception, however, this is no longer the case. Set aside all
concerns about incomplete evidence, and consider a highly complex but – for the
time being – extensionally adequate metaphysical theory couched in terms of
causal dispositions and the need to approximate particular clusters of them. In
other words, our normatively blind mythical being could read off from these

---

80 This example is discussed by Rosenberg and McShea (2008) in the closely related context of asking
whether there could be biological laws about species.
causal conditions the existence of any intentional states (albeit while perhaps not grasping everything essential about such states). Here’s the crucial point: such a theory would have absolutely no resources for implementing changes in the outlined causal connections. As a result, it would be utterly incapable of handling any shifts in the cluster of causal dispositions that determine instantiation of a given intentional state. These shifts might stem from (a) changes in worldly facts, (b) changes in our knowledge (i.e. changes in our understanding of the connections between concepts), (c) the emergence of new concepts, and (d) changes in concepts themselves.\(^{81}\) These ongoing changes are a critical feature of intentionality, and a metaphysical theory of intentionality ought to be able to account for them.

This is not a matter of holding the naturalistic accounts to unfair standards. Those accounts are not meant to be practically useful, as I have noted, so the problem is not just that they provide no workable method for predicting the various changes intentional states undergo. The problem, rather, is that they provide an incomplete account of their instantiation conditions. First of all, most naturalistic accounts are presented without any explicit (or implicit) provisions for change whatsoever, strongly implying that the networks of causal connections constitutive of different states are fixed and static. This just shows, as I have said, that the dynamic nature of intentional states has been entirely overlooked in this context. Second, it is not clear how sufficiently specific and informative provisions for change could be introduced into these accounts. The problem is that not all changes in the causal dispositions involving a particular state are changes in that state’s instantiation-determining cluster, so the naturalist would need some way to make this distinction. But the naturalist is limited to brute statements of causal connection, organized into key clusters, and neither the connections nor the clusters by themselves provide any resources for specifying conditions under which those clusters can change.

At this point, it is possible to lay out a version of the Davidsonian argument that works. It goes like this:

1. **RP**
2. Intentionality has an essentially dynamic nature
3. Purely naturalistic conditions (e.g. causal dispositions) do not allow for an adequate explanation of the changes intentionality undergoes over time
4. Therefore, AN

\(^{81}\) See Chapter 3 for a more extensive discussion of these sources of change.
I will now address the premises one at a time to finish fleshing this argument out.

I begin with premise (3). This is probably the premise with the fewest attendant concerns, but it can still be clarified to some extent. Now, I said a moment ago that a metaphysical theory should be expected to account for the changes that occur to dynamic entities. Here’s a potential way to make this point clearer: once changes have occurred, a metaphysical theory ought to be able to explain them retrospectively. Evolutionary theory, for example, generally does this quite successfully. It has two major advantages: it has various mechanisms of change built into it – natural selection, genetic drift, mutation, etc. – and it is not called upon to distinguish changes that matter from changes that do not, because all changes are on a par. I think it is quite clear, though, that a theory which metaphysically grounds intentionality in causal dispositions is ill-equipped for this task. After all, the theory involves something like a (very large) bunch of (very long) lists, and there is just no plausible candidate for a principle of modification for those lists. It is a great deal like the view that species are defined by a set of characteristics – how, looking back, could the members of the original set account for the changes in that set at a later time? I think it is similarly clear for theories based on other naturalistic properties – e.g. physical properties of the brain – that, if the grounding properties are seen as changing along with the intentional states themselves, there are no principles available in the theory to account for those changes. Looking back, an examination of an intentional state’s particular neurological basis at one time would provide no means for understanding why the state had a slightly different neurological basis sometime later.

The more likely naturalistic response is, rather, that the crucial grounding feature of an intentional state in fact remains constant, despite the state’s various superficial changes. This brings us to premise (2). Premise (2) affirms quite the opposite, maintaining that intentional states have no essential core that is necessarily preserved, come what may. This claim should be found immediately plausible by those who think intentional states are to be understood holistically. If it is accepted that the identity of a state depends on a broad network of connections to other states (and events, actions, etc.), then the uncontroversial observation that this broad network changes over time – through newly introduced concepts, newly discovered conceptual connections, etc. – should lead to an acceptance of premise (2) straightaway.

But what of those who are not holists? Well, one large group – probably the majority – of non-holists falls outside the scope of this discussion. This is because they clearly do not accept premise (1), the Rationality Premise, which is the Davidsonian argument’s starting point. As I’ve said, a separate discussion is

---

82 The so-called “major transitions” are, of course, very difficult to explain and very much up in the air. See, for example, Smith and Szathmary (1995).
required to determine whether one should accept that claim. Naturalists such as Fodor and Dretske claim, contrary to RP, that we understand intentional states (or at least beliefs, which are their prime focus) as essentially akin to signals or information-carriers and not essentially as items standing in rational relations.

There are those, however, like Wedgwood, who accept RP but suggest that there is a single fundamental disposition that goes into an intentional state’s instantiation conditions – i.e. one fundamental rational connection that must be exemplified in a subject’s causal behavior for the state to be present. I find this suggestion implausible for the kind of reason that generally supports holism – namely, that it is simply very difficult, for the overwhelming majority of intentional states (or conceptual components thereof), to isolate a single causal disposition that can bear the required weight. This kind of view is generally introduced with logical constants (e.g. ‘and’, ‘or’, and ‘if…then’) as its primary examples, and while it may in fact be workable for that handful of cases, it seems hopeless for most other ingredients of intentional states.

In addition, I think that an attempt to naturalize intentionality by specifying single dispositions along these lines would still run into trouble when it came to the appearance of new concepts. It might be able to shrug off new conceptual connections and other kinds of change, claiming they were inessential as long as the single fundamental disposition remained. But when a new concept arose, it would be unclear how the causal dispositions definitive of all the other concepts could possibly account, even retrospectively, for where the new one came from. Though the analogy is somewhat strained at this point, it would be as if evolutionary theory could account for changes in existing species but not the appearance of new ones. The upshot of all this is that, even on a Wedgwood-style account, a version of premise (2) would be available (focusing on the appearance of new concepts) that would allow the Davidsonian argument to go through.

Finally, a few words about premise (1), or RP. I said at the outset that assessing the Davidsonian argument involved, in part, getting clearer on the precise content of this claim. At this point, it seems that there at least two versions of the premise that work well with the intervening steps I have suggested. The first is the premise understood in the very strong way I mentioned in section 2.2:

(RP1) Our conceptions of intentional states/properties essentially involve facts about rationality and rational connections in the following way:

---

83 Wedgwood (2009) suggests an approach of this kind.
84 It is strained, of course, because new species arise only as a result of changes in existing species (especially diverging changes between two subgroups of a species), while new concepts do not, in any clear way, always arise, from changes to existing concepts.
intentional states are necessarily conceived of as either identical to or partly constituted by rational properties.

Originally, this was deemed insufficient for arriving at the anti-naturalist conclusion because the naturalistic fallacy, at best, ruled out only conceptual or analytic reductions, leaving open the possibility of empirically discoverable naturalistic instantiation conditions. Given the argument as currently outlined, however, such a metaphysical theory would clearly be unable to capture the changes intentional states undergo over time. An empirically discoverable connection between certain rational properties and certain naturalistic ones would not be projectible, as demanded by AN.

The other version of the premise, which I find more independently compelling, is the following:

(RP2) Our conceptions of intentional states/properties essentially involve facts about rationality and rational connections in the following way: rational connections must be used in order to zero in on the set of causal dispositions that determine a state’s instantiation conditions.

This premise lays down a framework for how our notions of intentional states rely on normative rationality judgments. But we cannot, as we could before, maintain that a naturalistic metaphysical theory might rest, stable and untouched, beneath this conceptual surface. On the contrary, steps (2) and (3) of the argument guarantee that we continually have to go back to normative rationality judgments in order to keep up with and understand an intentional state’s changing instantiation conditions. In other words, a metaphysical theory of intentionality must evolve – and not only in the evidential way that Davidson emphasized. But no evolving metaphysical theory of intentionality can entirely dispense with rationality and connections of that normative kind. Consequently, no naturalistic metaphysical theory of intentionality can be even extensionally accurate, considered over the long haul.

4. Conclusion (and a look ahead)

Arguments meant to establish anti-naturalist conclusions in the philosophy of mind face a difficult problem: if they are to avoid begging the question, their premises must not simply presuppose metaphysical differences between

---

What it is that accounts for the changes in rationality judgments over time is a vexing question that I need not go into for present purposes. It is enough that they do change, and that these changes are then reflected in changes at the level of (approximately) required causal dispositions.
intentional properties and naturalistic properties; the only readily apparent option, though, is to make use of conceptual premises, i.e. claims about how intentionality is thought of or understood, and those generally place too few constraints on the underlying metaphysics to establish a conclusion of the desired kind. This problem has become particularly acute as the goals of naturalism have shifted, i.e. as the hope for full-on analytic reduction has given way to more modest ambitions (like a posteriori identity, sufficient instantiation conditions, etc.).

The argument outlined here, though, succeeds where others have failed. It does so by pointing out the dynamic character of intentionality, a distinctive feature of it that has almost invariably been overlooked; by showing how that dynamic character results in an ineliminable role for normative considerations (at the conceptual level) in a necessarily ongoing project of metaphysical theory-development; and by making clear how that leads to inevitable conflict with a demand for projectibility, one of the remaining requirements of even the more modest contemporary brands of naturalism (at the metaphysical level).

But the fact that naturalistic standards have shifted in the past suggests that they could do so again. Or, more to the point, it makes clear that pinning down the demands of naturalism is crucial for assessing the prospects of arguments for and against it. After all, couldn’t the argument spelled out here fail against a version of naturalism with lower standards still? There are indeed philosophers who claim to be naturalists while demanding nothing like the projectible explanatory accounts this argument rules out. Are they really naturalists? If so, why? If not, why not?

In the next chapter, I will attempt to provide some clarity on the issue of naturalism and its demands. I will offer considerations in support of a new – or at least fairly uncommon – understanding of naturalism’s proper explanatory standard: the need to establish evolutionary continuity. And modifying our understanding of naturalism in this way does end up blunting the force of the argument put forth here, and so leads to a much more positive assessment of naturalism’s prospects.
Chapter 4

Evolutionary Naturalism

1. Introduction

In Chapter 3, I began discussing the implications of the Normativity Thesis for a naturalistic approach to intentionality. I showed there that the thesis, once clarified along the lines suggested in Chapters 1 and 2, could indeed be used to fill in a traditional anti-naturalist argument schema. A full understanding of the Normativity Thesis’s implications, though, requires that the demands of naturalism be clarified as well. The discussion in Chapter 3 simply borrowed one well-known (and relatively undemanding) notion of naturalism, but I believe there is another notion of naturalism that is independently superior and, crucially, leads to a different conclusion about the implications of the Normativity Thesis. My goal in this chapter is accordingly twofold. First, I want to argue, in a quite general way, for what I see as the proper standard for naturalism in the philosophy of mind. And second, I want to show that the Normativity Thesis, on its best construal, does not preclude an account meeting that naturalistic standard.

An immediate problem for the first undertaking, of course, is that the term ‘naturalism’ is very flexible in application, so it can seem that a dispute over what naturalism really is must be merely terminological. I think, though, that one basic way the term is used – what is often called its “metaphysical” rather than “methodological” sense – has a core set of widely shared intuitions underlying it, and there is a genuine stalemate regarding what those shared intuitions demand from philosophical accounts of the mind. If this stalemate can be broken in the right way, the result won’t be just a new use for the term ‘naturalism’ but a new and more satisfying way to approach that larger philosophical project. For now, I will leave these underlying intuitions quite vague – their core, essentially, is the

86 See, for example, Papineau (2009) or the introduction of De Caro and Macarthur (2004).
(vaguely expressed) thought that the mind is part of the natural world in the same way that perspiration and proteins are – and will begin by describing the stalemate.

Most of what gets called naturalism is reductive in a broad sense: it seeks to provide instantiation conditions for intentional states using only those notions considered to be scientifically kosher – in general, those derived from physics, chemistry, and biology. I call this “standard naturalism,” and it is exemplified most prominently by the theories of mental content due to Fodor (1987, 1990), Dretske (1986, 1988), Millikan (1984, 1993), Papineau (1987), etc. This approach faces difficulties of two kinds. On the one hand, there are technical concerns – roughly, concerns that a proposed reductive account does not yield the correct results, e.g. content attributions that line up (closely enough) with those of folk psychology. On the other hand, there are fundamental concerns – roughly, concerns that the whole approach is misguided from the start, e.g. that it is doomed to fail because it leaves out something essential to intentionality. The claim that reductions fail to capture the fundamental normativity of intentional states is one clear example.

Now, it is possible that these difficulties can be overcome, but at this point there is nothing like a consensus that they have been overcome. For that reason, it would be nice if, on reflection, naturalism turned out not to require reduction of this kind at all.

Indeed, some of those who resist the reductive project (or despair of its success) have suggested that a suitable naturalism can be preserved nonetheless. It can still be shown that the mind is part of the natural world, they say, even if certain features of the mental – like normativity – preclude reduction by means of the restricted set of concepts mentioned above. I will refer to this approach as “non-reductive naturalism,” and it can be found in the work of Brandom (1994), McDowell (1994), and Tye (1992), among others. The difficulty here is to say exactly how a non-reductive account can manage to demonstrate the naturalness of the mind rather than leave it too mysterious, detached, and special. After all, any work done to show that reduction is not in the offing is likely to heighten the sense of separation between the two levels or domains at issue; it then becomes incumbent on the anti-reductionist to say why those sharply separated phenomena are still equally natural, still unproblematically part of the same unenchanted world order.

Attempts to do this have been unsuccessful, in my view. And that makes this kind of naturalism begin to seem like nothing but a bare and empty assurance.

I can now provide a more substantial formulation of my primary goal in this paper: it is to articulate a standard for naturalism that (a) captures (much of) our intuitive motivation for such a position and (b) is demanding enough to be
satisfying, but (c) falls short of requiring reduction. A version of naturalism having these three characteristics is the best we could hope to adopt.\footnote{Pending successful reduction, perhaps. Though, given the kind of naturalism I advocate, I would suggest that reduction is actually not a preferable alternative so much as one way of satisfying the basic naturalistic demand. I pick up on this point below.}

The relevant standard, I claim, is \textit{evolutionary continuity}. This means that naturalistic explanation should be essentially diachronic in nature, showing, first and foremost, how a given feature has arisen or emerged over time, and not (necessarily) what that feature is constituted by at a given time. This provides a nice gloss on the central (vaguely expressed) naturalistic intuition mentioned above, and when philosophers describe the underlying motivation for naturalism they often do so in explicitly diachronic terms. Moreover, the versions of non-reductive naturalism on offer provide no insight – not even potential insight – on this diachronic question; this, arguably, is why they fail to satisfy as instances of genuine naturalism. And finally, such insight can be provided, at least in principle, even if wholesale reductions are unavailable. In sum, this makes \textit{evolutionary naturalism}, as it can be called, the proper goal to strive for in the philosophy of mind.

With this understanding of naturalism in place, the question about the implications of the Normativity Thesis can be reconsidered. In the final part of the chapter, I will make a case for the claim that the Normativity Thesis, though it blocks certain forms of naturalism (as seen in Chapter 3), is compatible with evolutionary naturalism. Much work remains, of course, to meet that naturalistic standard in full, but the Normativity Thesis provides no principled barrier to it being done successfully.

\section{2. Clarifying the demands of naturalism}

\subsection{2.1 Standard naturalism}

I have said that standard naturalism is reductive in a broad sense. It need not be \textit{strongly} reductive, e.g. in the manner expressed by Fodor’s famous phrase: “if aboutness is real, it must be really something else” (Fodor 1987: 97). A strongly reductive view implies that intentional notions are wholly dispensable, that they are no more than fully naturalistic notions in disguise. But a view can be much less drastic and still fit into the category of standard naturalism. The category includes any attempt to show how intentionality is \textit{grounded in} naturalistic events and properties. It can be seen as having two parts: an acceptance of supervenience – generally taken to be a minimal requirement for naturalism, but insufficient on its own – combined with an explanatory account of
that supervenience. The task for naturalism, on this view, is to say what non-intentional conditions make it the case that intentionality or aboutness is present.

Naturalistic accounts of this kind make claims at two levels of generality: first, at the more general level, they say what kinds of non-intentional facts go into the constituting of content; and second, they show how specific contents depend on specific variations in these more general features. To illustrate this basic structure, and to bring out the broadly reductive character of standard naturalism more clearly, I will briefly sketch the accounts given by Fodor and Millikan.

Fodor’s view is that the relevant non-intentional facts are causal/informational ones. Thus, there will be an intentional state with a particular content, on Fodor’s view, whenever there is a physical state with causal properties of a certain kind – namely, properties of reliable covariance with particular worldly features. Applied more specifically, a state will have ONION as part of its content, for instance, provided that the underlying physical state carries information about onions, i.e. is connected with them via a more-or-less precise causal link. The crude picture here is of some reidentifiable neural structure, firing consistently in the presence of onions, which, in conjunction with other such structures, is sufficient for having mental states that are about onions.

Millikan’s account gives pride of place to a different non-intentional fact – biological function. This too is to be understood in causal terms, ultimately, but causal-historical terms rather than causal-informational ones. What this means is that reliable covariance with some bit or feature of the world is neither necessary nor sufficient for content; what matters is that a state-type (physically construed) has been preserved by natural selection because in the past it was repeatedly – though perhaps only occasionally – associated with a particular worldly feature. So it is a state’s biological function, defined by its past contributions to survival and reproduction, that gives rise to content properties. Returning to the above example, a neural structure can give rise to an intentional state with the content ONION even if onions are not often present when it fires; this will be so as long as neural structures of that ilk have historically been useful when they did manage to fire in the presence of onions.

Despite the differences between them, it should be clear that these accounts share a basic view about what naturalism is and what its demands are with regard to philosophical accounts of the mind. Naturalism, according to this perspective, is a commitment to using only those explanatory resources sanctioned by natural

---

88 Boghossian (1991) helpfully puts things this way; he notes that supervenience, without a naturalistic account like this underlying it, is simply a brute and mysterious connection between two disparate families of properties.

89 Actually, a step has been skipped here. Physical states must first be typed as representations – by means of their functional role – before the question of their semantic content can arise. Only then do their causal covariance properties ground intentional ones.
science; more specifically, it is a commitment to using only those resources that are themselves utilized in natural scientific explanation. It is for this reason that naturalistic explanations of mental features (like intentional content) must be broadly reductive: they avail themselves of only a limited inventory of entities and properties (neural structures, causal relations, biological functions, etc.) and use them as the explanatory basis for a set of features that fall outside of that inventory.

Much of the discussion/criticism of these proposals, both from within the naturalist camp and from without, has focused on various technical difficulties that arise for this explanatory project. In general, the worry is that the explanatory resources at hand cannot serve as a basis for something specific enough and, relatedly, fallible enough to be intentional content. I have largely ignored the many subtleties of these accounts – Fodor’s notion of asymmetric dependence, used to privilege the “correct” causal links over the “mistaken” ones; Millikan’s attempts to overcome functional indeterminacy by stressing representation-consumption – that are designed to deal with these concerns. It is far more important, for my purposes, to examine the fundamental concerns, the suggestions that all these intricacies simply paper over a more deep-seated failure.

This failure could be characterized in a couple different ways. First, it might be thought that these reductive accounts are ultimately question-begging. For instance, it is a common criticism of Fodor’s view that when he appeals to counterfactuals in order to isolate a content-determining causal link – it is, roughly, the one on which all others modally depend – he tacitly relies on the assumption that content is kept fixed across the possible worlds being examined. If that is right, then content properties are simply assumed from the outset, not accounted for by means of more basic notions.

Second, one could argue that even if these reductive accounts avoid begging the question, there is still a gap between intentional facts and naturalistic facts that just cannot be bridged. As we have seen, one common basis for this sort of concern is the thought that the mental is essentially normative. The idea is that mental categories depend for their very identity on there being standards for their proper use. As such, these categories fit into a whole network defined by relations of justifiability and warrant, the items within that network engendering commitments rather than (or in addition to) brute causal effects. If that is right, then it looks quite plausible to assert, as McDowell (1998) does, that these items and categories are sui generis with respect to the domain of natural science – that the properties and relations definitive of mentality are not only absent from natural scientific explanations but so wholly different in kind that there is no “echo” of the former in the latter.
A fundamental objection of this sort is especially forceful when targeting strongly reductive accounts – an intentional state cannot be nothing but a certain complex of naturalistic features, for that leaves its essentially normative aspect out completely – but it can apply to more modestly reductive accounts as well. As Chapter 3 showed, the normativity of the mental, properly understood, can serve in a valid Davidsonian argument ruling out even empirical laws connecting the physical and intentional domains. If that is the case, then a standard naturalist attempt to explain the basic supervenience relation is entirely hopeless.

Whether one is convinced by considerations of that sort or merely driven to pessimism by the various technical concerns that arise, it at least looks plausible that reductions of the above kind are in trouble. And that fact is quite disturbing as long as one has some vaguely naturalistic intuitions – as I have claimed most everyone does – and standard naturalism is taken to capture those intuitions. The question then arises whether naturalism can be reconceived in a way that makes it independent of standard naturalism and its reductive requirements. I will now look at two attempts to do just that.

2.2 Non-reductive naturalism

The basic picture of standard naturalism is this: (1) a philosophical account of some phenomenon is naturalistic only if it shows how that phenomenon is grounded in features (properties, relations, events, etc.) that are part of the explanatory resources of the natural sciences; (2) many important features of the mind are not themselves part of the explanatory resources of the natural sciences; (3) so a naturalistic account of mind must be reductive, showing how those mental features are in fact constitutively connected with features that are part of the explanatory resources of the natural sciences. Any attempt to preserve naturalism while giving up the reductive requirement in (3) must jettison at least one of the premises. I will first look at a proposal due to Tye (1992) on which (1) is maintained while (2) is rejected, and then address McDowell’s (1994) claim that (1) must be given up in the face of (2)’s evident truth.

Tye’s line of thought is quite simple. He begins by granting (1), saying that something is natural just in case it participates in causal interactions falling under the laws and theories of a natural science (Tye 1992: 432). But, he says, when we gather the explanatory resources this makes available, we must recognize that psychology is just as much a natural science as physics or biology. This is so for two reasons: first, psychology is a science insofar as it involves the formulation of ceteris paribus laws (the sort found in most science); and second, it is a natural

---

90 Tye uses the term “physical science” rather than “natural science,” but this difference does not affect the argument.
science insofar as the items referred to in those laws are physically realized (a version of the supervenience claim accepted by all standard naturalists).

In other words, Tye argues that one can accept the basic picture of standard naturalism – the image of a restricted class of entities and properties that must be used to explain everything else – without thereby being committed to the need for reductive accounts of mind. The reason is simply that all the key features of mind are already in that restricted class; psychology makes reference to the contents of mental states in its explanations, and psychology is itself a natural science, so, as Tye puts it, “intentionality is already naturalistic” (438). And it is good that this is so, because “the prospects for successful reductions [of these notions] are bleak” (439).

I think this brand of naturalism comes too cheaply. I will return to discuss this point once I have presented a positive account of naturalism’s requirements, but for now I will simply mention the following. For one thing, Tye does not address the issue of normativity or otherwise confront the possibility that important features of mentality might not fit into a scheme of lawlike generalizations. But even if that possibility were illusory – if, say, the normativity claim were untenable – it would still seem that more was demanded of naturalism than Tye suggests. The two facts on which his naturalism rests – that law-based explanations involving mental entities are possible and that mental entities are physically realized – are plausibly seen as setting the main explanatory task: to show how and why those facts obtain. The standard naturalist attempt to account for supervenience (or physical realization), to say something about which physical conditions necessitate intentional states and why, is one way to try to satisfy that further explanatory need, and I will soon suggest another.

McDowell reaches the standpoint of non-reductive naturalism by a somewhat different path. On his view, as on Tye’s, the domain of the mental is irreducible. In particular, what he calls the “space of reasons” – defined by normative relations – is sui generis with respect to the “space of law” – the basic framework of natural science. In other words, he affirms the truth of (2) in the standard naturalist picture because he sees a fundamental difference between the law-governed character of science and the norm-governed character of mentality.

McDowell can preserve naturalism, then, only by calling (1) into question. It is a mistake, he claims, to identify the natural with the space of law (or its inhabitants). Instead, one must come to see that normatively governed judgments and intentions “belong to our way of actualizing ourselves as animals” (McDowell

---

91 Indeed, the analogy Tye sets up between psychology and other “higher” physical sciences could be thought to break down on precisely this point: perhaps geological entities and processes are deemed unproblematically naturalistic not just because we take them to be physically realized but because we can say something systematic about how and why they supervene on the physical.
The capacity to respond to reasons is part of our “second nature,” and this “could not float free of potentialities that belong to a normal human organism” (84). So rather than accepting the standard formulation of naturalism and arguing that there is more in the scientific inventory than one might think, as Tye does, McDowell grants that natural science is unable to account for what is distinctively mental and instead rejects the standard formulation of naturalism. Mentality is natural, according to McDowell, simply because it is an aspect of our way of life, a feature of our normal development, and we are, after all, just living, breathing creatures.

There is much in this picture that I find commendable. I agree both that our notion of naturalism should be broadened and that this broadening needs to involve a focus on something like biological development. Even within standard naturalism I take it to be a clear advantage of Millikan’s approach over Fodor’s that it emphasizes the biological character of cognition. But while teleosemantic theories such as hers argue that mental categories just are biological categories, that the attribution of content is ultimately an attribution of function, there is a different biologically informed outlook one might have. That is, one might take mentality to be one of our complex biological endowments, one that is natural, as McDowell suggests, simply because we are that way and because of how we got that way.

As an analogy, consider the power of flight. It would be odd to say that the power of flight is reducible to some function or set of functions. Of course it has functions in Millikan’s sense, but what it is is an organismal capacity with its own distinctive principles of operation. Similarly, mentality surely has biological functions, but it too is an organismal capacity with its own distinctive principles of operation. Perhaps those principles are so distinctive that, unlike in the case of flying, its operation is not amenable to fully scientific description, or at least not amenable to fully reductive description. No matter – it can still be seen as natural as long as it is clear how an organism managed to attain such a capacity.

It should now be obvious, though, why McDowell’s view leaves a great deal to be desired as a form of naturalism. In effect, that view says: here are some biological creatures that, simply through normal physical maturation and normal social upbringing, become normative creatures. Therefore, normativity is natural. But how exactly do organisms get that way? What does such maturation and upbringing have to be like, and what did the species history have to be like in order to make such upbringing possible? McDowell fails to give even the slightest hint of an answer to these questions, and for that reason the naturalistic challenge remains firmly in place.

92 It might be thought that these further questions simply fall outside the scope of philosophy. McDowell could be seen, then, as arguing only that naturalism requires no special philosophical work (beyond the
To address these questions is to pursue a new kind of naturalism, which is
diachronic in nature. I believe that this understanding of naturalism captures much
(or most) of our motivating intuitions and involves the right level of explanatory
demand. For that reason, it is the kind of naturalism that ought to be pursued.

2.3 Evolutionary naturalism

The examples of non-reductive naturalism addressed in the previous section
turn out to share a deep assumption with standard naturalism: that naturalistic
explanation is fundamentally *synchronic*. Standard naturalism’s characteristic
description of things in terms of properties on different levels, one level resting on
another, is a dead giveaway that a synchronic model is operative. The guiding
image is a kind of three-dimensional snapshot of all the entities and properties
present at a time, with explanatory linkages both within and between levels. One
or more of these levels have a privileged naturalistic status, while the status of
other levels depends on the explanatory connections that can be made between the
privileged level(s) and the non-privileged ones.

This perspective is also sometimes described using a language of spatial
containment. The question then is whether explanatory connections can be put
into place in such a way as to draw certain properties into a privileged circle.93
These are only metaphors, of course, but they are telling. In any case, the clearest
way to see the synchronic nature of standard naturalism is to look at the kind of
explanation it seeks: it is one in which, for example, intentional features that are
present at a given time are accounted for on the basis of non-intentional features
that are present at that same time.94

Even though Tye and McDowell reject the reducibility of the mental, they
preserve this basically synchronic outlook. Their suggestions can be captured
quite nicely using the three-dimensional snapshot metaphor mentioned above.
Tye essentially expands the base level so that many more things – including
mental properties – can be found in it from the beginning; as a result, explanatory
clarifications he offers, at least). It is certainly true that the remaining questions are not purely
philosophical (to the extent that anything is), but it is a mistake to think that further philosophical work is
not called for. Just to give one example, consider how an answer to those questions would be affected if
“becoming a normative creature” were taken to mean “having a special faculty for intuiting normative
truths” and having mental states relied on uses of that faculty.

93 McDowell is a good example. He describes the standard naturalist project as one in which the space of
reasons is “domesticated” or “appropriated” within the space of law, or in which the former space is shown
to be “just part of” the latter.

94 These non-normative features may include historical ones (as in Millikan’s account) without the
explanation given thereby ceasing to be synchronic, as long as it is *present* facts about the state’s history
that are doing the explanatory work. That there is a different kind of historical explanation will hopefully
become clear in what follows.
connections between levels are no longer required. McDowell, on the other hand, keeps the base level fixed (and narrow), but expands the notion of naturalism vertically, so that it takes in the space of reasons hovering above as well as the space of law down below. He has no problem with the spatial separation, just the allocation of privilege at the start.

The alternative to all of this is to understand naturalistic explanation in *diachronic* terms. Interestingly, this conception is often in the foreground when the motivations for naturalism are explicitly discussed. For example, Dretske, a standard naturalist, says we need a naturalistic explanation in order to make intelligible “how naturally-evolving biological systems could have acquired the capacity for belief” (Dretske 1986: 17). Similarly, Brandom, a non-reductive naturalist, notes that “one need not be an atoms-in-the-void physicalist in order to want to understand better the nature of the transition from a world devoid of semantic properties to one in which they are at least locally rife” (Brandom 2001: 587). A bit later, he makes the same point with regard to normativity: “there was a time when there was...nothing bound by or subject to assessment according to conceptual norms. Later, there was. An important part of understanding ourselves is understanding the nature of this transition” (607).

Clearly, the language in use here is not that of containment but that of acquisition, transition, origin. It evinces a perspective founded on the Darwinian notion of evolutionary continuity, and it is this idea, I think, that is properly thought of as the core of naturalism and the source of standards for naturalistic explanation.

There are other intuitive motivations for naturalism, of course. There is, for instance, the sort of thought that leads up to Fodor’s famous expression of strong reductionism, which is that physics determines all the basic ingredients of the world and anything else real must be constituted from them. But the kind of motivation I have highlighted – an urge to understand transitions in a certain way – is more general on at least two counts. First, it is more widely held. There are plenty of people who accept a broadly Darwinian outlook without committing themselves to a staunch physicalism. Second, it applies to more cases. There are plenty of instances in which the question of reducibility just does not come up but the question concerning origin or acquisition is pressing; consider accounting for the capacity to fly or, to give a cognitive example, the ability to imitate. Other things being equal, it is preferable if naturalism reflects sentiments that are more widespread and explanatory needs that are more wide-ranging.

So an evolutionary notion of naturalism fits well with a great deal of our motivating intuitions, and arguably with those that are most important (because

---

95 Though one with strong tendencies toward evolutionary naturalism, it is worth pointing out.
most general). Its other main advantage is that it sets the naturalistic standard at
the right level, neither overly demanding nor overly permissive. Evolutionary
naturalism allows for naturalistic explanations even when reductive explanations
are not available, so in that sense it is not overly demanding. This is a positive
feature because, as we have seen, (a) reductions tend to run into serious technical
difficulties, and (b) there may be deep, principled reasons for thinking that certain
key mental features, in particular, cannot be captured by any reductive account.
So it counts in favor of an approach that it is in some ways less onerous than
reductionism, as long as it is still substantial and satisfying overall—that way we
are not left banging our heads against the wall for no good reason.

I foresee two kinds of objection. First, someone could say that accepting
lower standards in this way is just a cop-out, a way to avoid both the hard work of
the reductive project and the harsh implications implied by its failure. An
adequate reply to this can only come from the other two components of my
defense: the claim that a less demanding notion actually better reflects our
underlying motivations for naturalism, and the claim that the evolutionary
alternative is still demanding enough to be satisfying (which remains to be
argued). Second, someone could challenge the claim that naturalistic explanations
are really possible without reductions. It must certainly be granted that successful
reduction is one way to carry out the broader explanatory project—that is, one
might show how a higher-level feature could have arisen by revealing that lower-
level entities present before the transition just needed to come together in a certain
configuration. (Indeed, this goes back to the earlier point about increased
generality: evolutionary naturalism subsumes reduction in the sense that the latter
can play a role in the former; evolutionary naturalism can account for the
usefulness and desirability of reduction, but not vice versa). But what if that were
the only way to account for a transition?

An initial reply to this might be: well, maybe that is not the only way. In
other words, there is a question of burden of proof here. Barring further argument,
evolutionary naturalism at least carves out conceptual space for satisfying
diachronic explanations that are not (and do not depend on) synchronic reductions.
But of course a better reply would be to give actual examples of such
explanations, and I will do that in the next section.

It remains to show that, despite eschewing reductions, evolutionary
naturalism is still demanding enough to be satisfying. First of all, it is arguable
that every instance that calls for a reductive explanation calls for a diachronic
explanation as well. After all, the targets of reductions are generally high-level,
complex features or entities, not the sorts of things that have been around from the
beginning of time (like, say, fundamental particles, fields, forces, or what have
This means that a shift to an evolutionary naturalist approach does not require dropping any metaphysical questions that previously seemed important; those questions are reformulated, but a naturalistic demand is still there. This contrasts with non-reductive naturalism, which was unsatisfying because of the complacency it called for on certain subjects. But adding the diachronic explanatory challenges of evolutionary naturalism rectifies the situation. Tye might be right that psychology is fully scientific, but the transition from a world in which psychological laws are not applicable to one in which they are is left mysterious. McDowell may be right that responsiveness to norms is part of the human biological repertoire, but the advent of such a capacity is left completely unexplained. If I am right that accounting for these transitions is absolutely central to naturalism properly construed, then these views are simply not naturalistic.

One last point in favor of evolutionary naturalism: it is genuinely interdisciplinary. It is quite bizarre on the face of it to the think that naturalism, of all things, would be a project to be undertaken solely from the armchair. Yet that is how it has been approached. Sterelny (2003) points out that even Millikan’s account, for instance, which uses a great deal of conceptual machinery derived from evolutionary theory, is not itself evolutionary because it depends in no way on the historical facts of humankind’s development. Now, I believe there can be philosophical contributions to the naturalistic project that are independent in this way, but they do need to be seen as part of a larger package, as serving to help construct a diachronic explanation that relies as well on scientific evidence and reasoning, contingent historical fact, etc.

3. Putting evolutionary naturalism to work

3.1 The demands of diachronic explanation

I have claimed that evolutionary naturalism makes room for fully satisfying naturalistic explanations of a diachronic, non-reductive kind. To say precisely what such explanations require is a difficult task, however. That is because their goal is to account for evolutionary novelties – brand new structure and capacities – and not just shifts in the population-level distribution of already present traits, which is what evolutionary theory’s best developed tools tend to be good for. The problem of accounting for evolutionary novelties is at the center of an open and

---

96 Could certain normative properties like “good” be exceptions? I am not sure how to settle this question, to be honest, but it would strike me as bizarre for even a normative realist to say that the property “good” predated living things.
active research area, with many biologists and philosophers of biology addressing it in recent years.\textsuperscript{97}

My aim here is to outline and illustrate some of the major components of diachronic naturalistic explanation, to break down a few of the primary tasks that need to be performed. This outline will be rough, certainly, but it will be enough to permit an assessment of the Normativity Thesis’s implications for evolutionary naturalistic explanation.

A diachronic explanation of a novel feature can involve up to at least three major task-types. First, if the target explanandum is a complex cognitive/behavioral capacity, as is the case whenever the philosophy of mind is involved, an initial task is to show how the capacity at issue rests on other – generally simpler or more constrained – capacities. As we shall see below, this is not a matter of reducing a capacity to simpler ones, as that is normally understood, but of mapping out a capacity’s preconditions. This accomplishes two things. It breaks down the more specific explanatory tasks (to be discussed shortly) into smaller and more manageable pieces, and it provides the first, albeit fairly abstract, framework for viewing the transition as a step-wise progression over time. It is also this task which is most likely to require a philosophical contribution, for the capacity in question must be carefully and adequately characterized, its preconditions conceptually tested, etc.

A nice example of this first task can be found in Sterelny (2003). Sterelny seeks to explain the advent of human culture, the greatly expanded use of materials – in terms of both the range of materials used and the range of uses to which they were put – that appears in the archaeological record roughly 60,000 years ago. The account runs broadly like this. Human culture rests on two things, a capacity for \textit{cumulative niche construction} (i.e. the ability to modify one’s own selective environment in ways that are preserved and further modified in future generations) and a tendency toward \textit{phenotypic plasticity} (i.e. a developmental program that yields different outcomes depending on different inputs). Cumulative niche construction, in turn, has two preconditions, a capacity for \textit{socially cooperative behavior} and a capacity to engage in \textit{pure imitative learning}. Pure imitative learning, Sterelny says, does not depend on a “theory of mind,” as some have thought, but rather on the ability to \textit{represent abstract behavioral programs}. And so on.

Neither the details of this explanation nor its success are important for my purposes. What it is meant to do is exemplify the initial task of spelling out the hierarchy of preconditions for a particular complex feature. As we shall see

\textsuperscript{97} See, e.g., Wagner (2000), Pigliucci (2008), and Love (2008). The rise of interest in Evo-Devo, the study of interactions between evolutionary and developmental biology, is in large measure a response to this problem.
below, the clarification of the Normativity Thesis undertaken in previous chapters can be seen as the first step in a procedure of this kind.

The second task is to say something about the evolutionarily relevant conditions underlying the feature’s development. This most often involves a description of a selective environment favoring the capacity at issue (or one of the simpler capacities it is based on). For example, Sterelny suggests that cooperative behavior was strongly favored by the hominid shift to a savannah-based habitat, which required that infants be carried across increasing distances. But it is important to note that this is just one available option – other factors, such as architectural constraints, genetic drift, etc., can just as well come into play. Hence, a commitment to evolutionary naturalism does not require any particular stance on the dispute concerning the relative importance of adaptationism. The overarching goal of evolutionary naturalistic explanation is to establish continuity, but that can be done in a number of different ways.

The third task is to provide detail about the genetic and developmental processes that produce the feature at issue. The ultimate measure of continuity is genetic – i.e. a transition must be underwritten by realistic rates and types of genetic change – but because of the way fairly small changes in development can ramify, a minor genetic change can result in quite dramatic change in an organism’s macroscopic features. This is particularly true in the case of cognition, since a change affecting the development of the brain can theoretically yield radically new cognitive abilities.

Clearly, this is an exceedingly difficult task to carry out. The study of genetics is still in its relative infancy. Moreover, there are serious limitations on the evolutionary conclusions that can be reached at the genetic level, as Wagner (2001) points out; even when simple model systems are involved, testing hypotheses about the genetic changes that have led to novelties is hamstrung by the fact that the genetic background – i.e. the other genetic material in the vicinity – is so important. It may be impossible to reconstruct the exact genetic conditions that primed a population to make a certain transition in the past.

As for development, we are similarly just getting started in understanding it. That said, Deacon (1997) provides a good example of the kind of informed speculations that can be offered. Deacon’s aim is to explain the human capacity for linguistic reference, a novelty crying out for naturalistic treatment. (Non-human communication systems have at best very superficial analogues for this phenomenon). He begins with the clarificatory work of the first task, the result of which, very briefly, is that reference depends on learning high-level, abstract associations (between different symbol tokens) and to some degree suppressing

---

98 See Fodor (1996) for the concern that most evolutionary explanation is too adaptationist in orientation.
low-level, more immediate associations (between symbol tokens and objects). The problem, then, becomes one of explaining how humans acquired an ability/tendency to learn a certain way.

At this point, facts about how the human brain develops are brought in to help. It is well known that primates have relatively large brains for their body size, and that this is further exaggerated in humans. But Deacon argues that increased size is too coarse-grained to be very explanatory; far more important are shifts in functional organization. And it turns out that facts about specifically human brain development can account for such shifts. First, Deacon notes that the human increase in brain size is devoted disproportionately to the cerebellum and the cerebral cortex. This can be seen from the earliest stages of embryological development, and it results from enhanced growth in one particular area: the dorsal part of the developing neural tube. (This suggests a role for certain developmental genes whose operation is specific to that region). This is the first key developmental step – a (conjectured) modest genetic change leads to a pronounced macroscopic change, i.e. the dramatically increased size of certain brain regions.

Next, the fact that brain development relies on neural competition for targets means that these disproportionately larger regions will become far more extensively connective. This is the second key developmental step – a pronounced change in the relative size of different regions results in an even more pronounced change in connectivity and function. And since the pre-frontal cortex is known to be associated with tasks requiring suppression of immediately salient stimuli, Deacon suggests that it is the increased operation of that region that allows humans to overcome the learning obstacles associated with linguistic reference.99

As with the Sterelny example, the details of Deacon’s argument are not particularly important. For my purposes, it is important only that it shows how genetic and developmental information can be used to explain, in part, an evolutionary transition. With that, my survey of the three major tasks is complete, and we now have a rough outline of how diachronic explanation works. At no point does Sterelny or Deacon attempt to provide (or imply any need for) a reductive account of their respective explananda, yet we can see how genuine progress might be made in accounting for the advent of those novel features and securing their status as natural phenomena. This both illustrates the basic explanatory standard of evolutionary naturalism and engenders optimism that the standard can be met. In my view, this is the proper way to approach naturalism in the philosophy of mind.

To that end, I turn to the Normativity Thesis one final time.

99 We would of course like to hear more here about how exactly the pre-frontal cortex plays this role, but at this point we are running up against the limits of current knowledge.
3.2 The implications of the Normativity Thesis revisited

I have argued that as long as naturalism is taken to be synchronic in nature, the Normativity Thesis presents a decisive obstacle to it. Normative judgments serve to mark the contours of intentional categories as they change through time, and this precludes a broadly reductive account of intentionality in strictly causal terms. But what if naturalistic explanation is diachronic?

Shifting to evolutionary naturalism puts the question of intentionality’s naturalistic status into a completely different light. What had previously seemed like the main explanatory challenge – to ground particular intentional state-types in underlying causal facts, underwriting an otherwise mysterious supervenience relationship – in a sense just falls away. The main challenge now is to explain how a certain very general capacity – the capacity to have intentional states – has arisen, not to take intentional states one by one and figure out what underlies them. And the first step in meeting this updated challenge is to determine in more detail how that capacity to have intentional states should be characterized.

Clarifying the Normativity Thesis, as I have done in Chapters 1 and 2, turns out to help with precisely this task. I have argued that the only sustainable version of the thesis (a) adopts the basic functionalist framework of constitutive causal networks and (b) adds to it a role for normative judgment in accounting for changes in those networks. This means, in short, that the capacity to have intentional states rests on at least two other capacities. The first is the capacity to have behavior-mediating states that are dispositionally very complex and highly interconnected. The second is the capacity to construct and work with abstract normative models. I will now discuss each of these in a bit more detail.

It has long been recognized that one major difference between human thought and that of non-humans is the extent to which there are multiple avenues into and out from a given representation. That is, our representations are “decoupled” from specific sensory stimulations and from specific behavioral responses. Accounting for the origin of intentional content accordingly requires explaining how it became possible to have these decoupled cognitive states. A great deal of work remains to be done on this topic, but I will briefly mention two conjectures. First, regarding background selective pressures, it has been suggested that the change in hominid habitat (from forest to savannah) made it advantageous to develop different food extraction techniques; this would favor decoupling in

---

100 There is a clear connection here to the topic of linguistic reference addressed by Deacon. In that case, the communicative vocalizations of non-human animals are so tightly bound up with particular stimuli and responses that they are in fact more naturally viewed as involuntary reactions akin to laughter. Reference proper only comes with more decoupled symbol uses.
that a broadened behavioral repertoire would allow for tinkering, experimentation, tool use, etc.\footnote{See Sterelny (2003).} Second, regarding genetic and developmental forces, it seems that Deacon’s points about the changing structure and connectivity of the human brain are again apt; our sensory and motor cortical areas are comparatively small while the pre-frontal cortex – associated with abstraction, planning, suppression of immediate action, etc. – is comparatively huge, and this provides the beginnings of a plausible neurodevelopmental story about decoupled representations.

The second condition – that we have a capacity to construct and work with normative models – is key. It illustrates the great importance, in the context of diachronic explanation, of the first-stage philosophical account. For imagine if the Normativity Thesis were sustainable in a more traditional and direct form. Say, for instance, that it were possible to establish that intentional states essentially are or depend on normative commitments\footnote{As Bilgrami (2004) and Zangwill (1998) do, respectively.} or that they require responsiveness to rationally defeating conditions as such.\footnote{As in Wedgwood (2009)} If that were the case, it would be much more difficult to know how to proceed in the diachronic explanatory project. It would be necessary to explain both the advent of normative properties themselves and how we came to have states with which they were inextricably entwined. It would be a stark challenge indeed.

Fortunately, those versions of the Normativity Thesis do not pass muster. On a proper understanding of the thesis, it is not the case that normative properties are found within the constitutive bases of intentional states, so to speak – it is rather that normative judgment plays an ineliminable role in determining the causal-dispositional ideals that different intentional states must approximate. This alters things substantially. In effect, it reduces the additional explanatory challenge (beyond that of accounting for decoupling) to one that would have been present regardless. After all, it would be clear that we operate with normative models of other thinkers (and ourselves) no matter what transpired with the Normativity Thesis.

What does operating with a normative model involve? It is very closely related to the much-discussed ability to construct a “theory of mind.” In fact, it is a good bet that whatever structural/functional innovations in the brain provide the cognitive resources for one do so for the other as well.\footnote{Again, Deacon’s story about the pre-frontal cortex appears to be a good place to start here, though the need for further detail is as pressing as ever.} That said, it is most natural to think of a theory of mind as relatively transient and local, as constructed to apply to an individual thinker, as based on a fairly limited but directly available set of data, as used predictively, and as incorporating irrational tendencies.
Normative models function somewhat differently. They are more stable, and they serve as something like a background theory of mind, which we use both as a template to aid in construction of the first sort of theory and as an evaluative standard that gives rise to various sorts of critical and/or coordinative behavior. Given that, it is plausible that selective pressures favoring group coordination played a role in the evolution of this capacity.105

If normative models of this kind were not cognitively available, there would be no way to make sense of the distinctive nature of intentionality described by the Normativity Thesis. For we must understand not only how humans gained a capacity to instantiate causally flexible cognitive states, but also how those states could have and be seen as having dynamic natures, persisting through subtle changes over time. Without our having a capacity to construct and work with normative models, intentional states would eventually lose their moorings.

In the end, the Normativity Thesis, correctly interpreted, proves to be amenable to the diachronic explanatory procedure of evolutionary naturalism. This is by no means a trivial result. Before evolutionary naturalism was articulated as an option, I showed that the Normativity Thesis ruled out standard naturalism. And a more conventional understanding of the thesis, if it had been workable, would have made trouble for the diachronic option too. In this way, the clarification of the Normativity Thesis and the clarification of naturalism’s demands have both been crucial. They have come together to yield a view of the mind that is properly attendant to its most distinctive features and satisfyingly naturalistic at the same time.

105 Gibbard (1990) says something similar about normative speech and judgment, more generally
References


Wagner, G. (2001). What is the promise of developmental evolution? Part II: A causal explanation of evolutionary innovations may be impossible. Journal


Appendix

Assessment-sensitivity and the naturalistic fallacy

1. Introduction

The notion of normativity looms large on the contemporary philosophical scene, and a major reason is that normativity is seen as having important anti-naturalist implications. That is, it is widely – though by no means universally – accepted that there is a “naturalistic fallacy,” that the normative is *sui generis* and so cannot be identified with, reduced to, or derived from the non-normative. Among those inclined to accept this claim, there is still no firm consensus regarding exactly why it is true, what its precise scope is, or how it is best argued for. A major goal of this paper is to provide some clarity on these matters.

But set that concern aside for a moment. Whatever its source, the naturalistic fallacy, if correct, yields a powerful conditional: if there is ineliminable normativity in a certain domain, then any strong form of naturalism in that domain is ruled out. It then becomes a matter of *locating* normativity, so to speak, to see where this anti-naturalist verdict applies. This is far from straightforward. It has been argued that normativity is present, and that the anti-naturalist verdict applies, not only in value theory, but also in epistemology, theory of meaning, and philosophy of mind, but these arguments are controversial. As Rosen (2001) has convincingly maintained, the notion of normativity is in many ways too ill-defined, too jumbled, to allow for clear-cut answers to basic classificatory questions involving it. Remedying this situation (to a degree) constitutes a second goal of this paper.

As a way of meeting these two goals, I propose the following necessary (but not sufficient) condition for normativity: claims or judgments involving normative elements are *assessment-sensitive*. That is, for a claim to count as
normative (in a sense strong enough to have anti-naturalistic implications), the truth-value of the claim must depend not (or not only) on the context in which it is produced, but on the context from which it is assessed.

The main support for this proposal derives from its many theoretical benefits. First and foremost, it provides an explanation for the naturalistic fallacy, showing precisely why normativity resists naturalistic treatment, and it does so by means of a unified treatment of the two traditional paradigms of normativity, prescriptives (as focused on by Hume (1740/1978)) and evaluatives (as focused on by Moore(1903)). Second, it offers a principled way of understanding the open question argument, the interpretation of which has been a matter of great dispute; so understood, the argument has the broad scope that is often desired for it (but that has proved difficult to secure), for it rules out a posteriori identities as well as definitional ones. Third, the proposal avoids the problems that attend the other main ways of accounting for the naturalistic fallacy; on the one hand, it maintains the truth-aptness of normative claims and so does not have the embedding problems, for example, that plague noncognitivist approaches; on the other hand, it does not rely on motivational internalism, the controversial idea that normative judgments, construed in a cognitivist manner, are inherently motivating. Fourth and finally, the proposal can be applied to a related issue in the philosophy of mind, suggesting a plausible construal of the much-discussed Davidsonian claim that there can be no psychophysical laws because of the “disparate commitments” of the two realms involved.

So the proposal does an excellent job of meeting the first goal I mentioned. It also helps with the second, as can be seen from the following important example. If the proposal is correct, then the notion of semantic correctness, often thought to be the important normative element applying to linguistic meaning and mental content, is not normative after all. This is because judgments concerning semantic correctness are not assessment-sensitive, so the proposed necessary condition is not met. From this, it follows that the naturalistic fallacy cannot be used as a principled argument against naturalism with regard to semantic content. (Of course, there may be other reasons why naturalism is not viable in that context).

There is another implication that is worth noting. Though normativity still has anti-naturalist consequences on this proposal, those consequences are made to seem shallower, in a certain sense, than they might have initially. That is, in the end it is not normativity per se that gives rise to and accounts for irreducibility, but a characteristic it happens to share with other domains of discourse.

---

106 The qualification matters. The term “normative” is used in various ways, some of which may be compatible with naturalism. The necessary condition I’m proposing is only for being normative in the stronger, anti-naturalistic sense.
2. Assessment-sensitivity

Assessment-sensitivity is a semantic framework that John MacFarlane has developed in a series of papers (see especially his 2003, 2005a, 2005b, 2007), applying it to a number of theoretically troublesome kinds of assertion. The basic idea of assessment-sensitivity is this: the truth-value of a single claim, produced in a particular context, can shift as a result of relevant differences between those assessing the claim. MacFarlane argues that future contingents (e.g. “there will be a pop quiz tomorrow”), knowledge attributions (e.g. “Alfred knows that his car hasn’t been stolen”), claims of personal taste (e.g. “Indian food is delicious”), and epistemic modals (e.g. “Alice might be at the library”) are all assessment-sensitive. He attempts to show in each case that this interpretation provides the only satisfying account of various data concerning such claims – e.g. how we treat disagreements involving them, the circumstances under which they are retracted, etc.

In order to get a fix on this basic idea, let’s look briefly at claims of personal taste. On MacFarlane’s model, when someone says, “Indian food is delicious,” they do not say something that is, as he puts it, “covertly about the speaker” (MacFarlane 2007). In other words, they do not say something more or less equivalent to “Indian food tastes good to me.” This view could be formulated as follows: a term like “delicious” applies truly to something only relative to a standard of taste, and a particular speaker’s claim involving that term implicitly involves his or her standard. This would amount to a standard kind of contextualism – the truth of the claims at issue would depend on features of the context in which they are made (in this case, on a key feature of the speaker).

MacFarlane argues that this interpretation simply doesn’t do justice to the disagreements that can arise concerning claims of this kind. After all, if we were to reply to the original utterance by saying, “no, Indian food is not delicious,” we would take ourselves to be saying something incompatible with it, such that only one of the claims could be correct. But that isn’t the case on the contextualist view sketched above; the two claims would be essentially unrelated, semantically speaking, since they were produced in different contexts (i.e. by different speakers), and they could quite easily both be true (if those speakers have different standards of taste). So this leaves two options. Either (a) our sense that the reply is incompatible with the original utterance is completely mistaken, or (b) predicates like “delicious” do not behave in the suggested contextualist manner.

MacFarlane’s approach is a way of taking option (b), allowing us to hold onto our common-sense picture of disagreement and incompatibility. That said, the approach does adopt a key component of the contextualist view, for it accepts
the idea that predicates like “delicious” apply truly to things only relative to particular standards of taste. The key difference is that it is not the original speaker’s standard that matters to the evaluation of a claim as true or false, but that of the person(s) assessing the claim. So, “Indian food is delicious,” as uttered by A, might be true as assessed by B (given his standard of taste) and false as assessed by C (given hers). Or it might be true as assessed by A at the time of utterance and false as assessed by A later on, if his or her standards have changed in the meantime. Importantly, though, if D says “Indian food is not delicious,” then no matter who is doing the assessing (and when), only one of the two claims – A’s or D’s – could be assessed as true. In that way their incompatibility is preserved, and we can make sense of A and D as disagreeing with one another.

Some will surely think that the problem with this proposal is precisely what it shares with the original contextualist view: the contention that predicates of this kind apply truly to something only relative to a certain standard. The obvious alternative is to say simply that they apply truly or not – Indian food either is or is not delicious – and that’s the end of it. And this objectivist reaction\(^{107}\) will be all the more common/compelling if the issue is predicates like “right,” “good,” and “rational,” rather than “delicious.”

I readily admit that the proponent of assessment-sensitivity is in a tricky spot here. MacFarlane, in a very brief discussion that opens his (2007), suggests that there are features of how we use terms like “delicious” that tell against an objectivist construal – such as that we make judgments involving those terms without the circumspection and deference that is characteristic of other, paradigmatically objective judgments – just as there are other aspects of use – e.g. how we go about disagreeing and issuing retractions – that provide evidence against contextualism. I’m not sure how convincing the anti-objectivist considerations he offers really are, though.\(^{108}\) For one thing, it’s pretty clear that circumspection and deference are not entirely lacking even when it comes to our judgments of taste; many people are more cautious issuing verdicts around so-called experts in such domains, often retreating to safer, overtly subjective judgments instead (e.g. “well, I like it, at least”). And to the extent that circumspection and deference are relatively lacking, this could simply be because it is less well-established who is authoritative about such matters, and we all provisionally take ourselves to be (though some of us must of course be wrong).

---

\(^{107}\) Though the terms are by no means perfect, I will throughout the paper use “objective” as a label for expressions whose semantics is more standard, i.e. neither contextualist nor assessment-sensitive, and “objectivist” as a label for views that take certain terms to be objective in this sense.

\(^{108}\) I should note that MacFarlane himself in no way presents these considerations as fully decisive. His discussion is primarily meant to set the stage for his paper’s main focus, which is the comparison of contextualism and assessment-sensitivity, and he simply leaves the objectivist aside with the suggestion that there is “more to say” against that option.
The underlying trouble is that if these terms really do behave in a non-objectivist manner, then that makes the original account of disagreement less compelling – or rather, it makes it less clear that disagreement is a datum that needs to be accounted for in the first place. Do we really take two claims to be incompatible if hearing one of them (even from many different people) leads to no hesitation whatsoever in uttering the other? Moreover, MacFarlane speculatively suggests that the point of assessment-sensitive discourse is to “foster controversy” as a means to the coordination of our various standards. But this too depends on disagreement being uncomfortable – otherwise there would be no pressure to coordinate – and non-objectivist behavior seemingly belies the presence of any real discomfort. The prospect of unpleasant disagreement would presumably make us at least a little bit cautious in our judgments.

It seems, then, that a viable picture of at least some assessment-sensitive expressions is this: though their use is like that of objective expressions, their underlying semantics is different. In other words, it is simply a fact about certain claims that their truth is relative to the background standards of those assessing them, but that fact does not manifest itself in practice in any clear and decisive way. If that’s right, though, then how does one establish that a term of that kind is assessment-sensitive rather than objective? As far as a general answer to this question goes, I can do no more here than provide a brief and rather sketchy suggestion; when one focuses on the particular case of normative expressions, however, I think a clear line of argument presents itself, which I will spell out in a moment.

First, with regard to the general question, I think Harman (1985) nicely describes the underlying framework for an answer. Harman says there are two basic approaches we can take toward an area of inquiry. On the one hand, we can engage in it from within, taking the conceptual resources it offers and the relevant beliefs we find ourselves with and attempting to achieve reflective equilibrium – maximal coherence, a set of general explanatory principles, etc. On the other hand, we can step outside that area of inquiry and ask how its basic conceptual ingredients fit into our overall picture of the world – this, Harman says, might result in reductionism, or nihilism, or non-reductionist ontological commitment, or it might call for a revisionary semantic proposal like expressivism or assessment-sensitivity.

Assessment-sensitivity does not necessarily fail to show up in practice, though; a plausible case can be made that epistemic modals, for example, are used quite differently from standard objective expressions. Harman is primarily concerned with moral inquiry, but he mentions religious belief and our thinking about the mind as other areas toward which these two approaches can be taken. I take it that the point could be extended to various other areas as well, judgments of personal taste included.
For our purposes, the important thing to note is that this distinction between two approaches provides a way of understanding the split between use-facts and semantic-facts that I mentioned above. It is perfectly feasible, on this view, for the engaged use of a set of terms or concepts to exhibit objectivist characteristics, while the detached or external approach calls for a non-objectivist semantic treatment. As such, establishing assessment-sensitivity will crucially depend on the detached or external approach, and will not come from an examination of our engaged practice alone. I’m inclined to think that, to the extent we find it plausible that terms like “delicious” are assessment-sensitive rather than objective, this indeed results not from differences in how those terms are used so much as differences in how they strike us from something like an external perspective.

To be truly informative, this framework would need to be worked out in greater detail, showing precisely what the external approach involves and how a particular semantic treatment can get established by means of it. My aim here, though, is to defend the specific thesis that assessment-sensitivity is a defining – i.e. necessary – feature of normative discourse. And my argument for this thesis, though compatible with the Harman-style “two-approach” model, does not depend on it. What it depends on is just this: when we reflect on how normativity fits into our overall picture of things, there is a reasonably strong intuition supporting the naturalistic fallacy and its anti-reductionist implications. As such, it is of clear theoretical benefit, other things being equal, if that intuition can be backed up and the fallacy accounted for. My claim is simply that taking normative expressions to be assessment-sensitive does the best job of this, explaining the naturalistic fallacy (and the associated Open Question Argument) far more effectively than an objectivist construal can, while at the same time avoiding the semantic problems that beset the main noncognitivist alternatives.

In a sense, the argument for the main thesis boils down to this. An account of normativity faces two separate, apparently conflicting demands. On the one hand, it should explain certain deep theoretical intuitions, the naturalistic fallacy foremost among them. On the other hand, it should make sense of the semantic facts pertaining to normative claims. The first task pushes one away from objectivism, but, as can be seen with noncognitivism, seems to lead to trouble with

---

111 Harman thinks this is the case with moral terms and concepts, for instance.
112 Nor does it depend in any way on the suggestion that some assessment-sensitive discourse is indistinguishable from objective discourse at the level of practice; my argument would simply supplement whatever case could be made on the grounds of differences in use.
113 Does my argument have no importance, then, for those who don’t have that intuition? If they have no such intuition, no sense that there is a conceptual gap in the vicinity, and so see no advantage in accounting for such a gap, other things equal, then it may not. But if they are driven to deny the fallacy on the basis of general naturalistic commitments, then I hope my argument might be seen as allowing them to have their cake and eat it too. Unfortunately, that last bit goes beyond the scope of the present paper/chapter.
the second. A return to objectivism makes the second task straightforward, but the first remains completely baffling. The claim that normativity is assessment-sensitive is uniquely equipped to handle both of these tasks simultaneously.

I now turn to the positive theoretical explanations that the thesis offers.

3. Explaining the naturalistic fallacy

According to the naturalistic fallacy, there is a principled division between normative properties (and judgments involving them) and natural properties (and judgments involving them).\(^{114}\) These two classes are autonomous with respect to each other in (at least one of) the following senses: (1) there are no property identities between members of the two classes; (2) no property of one class can be reductively explained in terms of properties of the other class; (3) the presence of a normative property cannot be inferred from the presence of natural ones alone.\(^{115}\) Fortunately, I don’t have to adjudicate between these different construals, since the assessment-sensitivity of normative properties nicely accounts for all of them.\(^{116}\)

Assessment-sensitive properties exhibit quite distinctive behavior, since the instantiation of such properties varies along with the standards (of the relevant kind) of those assessing the given situation. An object might have a property as assessed by one person, but lack the property as assessed by another, and neither person would have to be making a mistake (from his or her point of view). This is very much unlike most natural properties, which are objective in nature – their instantiation is a simple (though not always easily decidable) yes-or-no matter.\(^{117}\)

\(^{114}\) These two classes are mutually exclusive by terminological stipulation. They need not be jointly exhaustive, though – supernatural properties, for instance, presumably fall into neither category. And while it is difficult to say precisely what it is that qualifies a property as natural, the ensuing discussion depends on no more than a rough sense of it. Of course, a large part of my aim in this paper is to go some way – though admittedly not all the way – toward saying what it is that qualifies a property as normative.

\(^{115}\) Unlike the other two, this third characterization of autonomy is asymmetric. This is due to the plausibility of principles like “ought implies can,” which would allow certain descriptive conclusions to be reached on the basis of normative premises.

\(^{116}\) I have shifted here from talk about normative expressions to talk about normative properties primarily because it is easier to put the metaphysical points currently at issue in the latter terms. Nothing crucial turns on this shift, however, as the treatment of assessment-sensitivity is strictly parallel.

\(^{117}\) It is important here not to confuse assessment-sensitive properties with very closely related objective ones. For instance, even if \textit{delicious} is assessment-sensitive, \textit{delicious according to so-and-so’s standards} (or \textit{delicious}_A) is not. A single object can be in the extension of \textit{delicious} as assessed by A, and not in the extension of \textit{delicious} as assessed by B – that’s the distinctive behavior just mentioned. At the same time, the object will simply be in the extension of \textit{delicious}_A, and not in that of \textit{delicious}_B, with no qualification needed. But those are different properties. Properties of that kind might play a key role in a contextualist view, but the problems with that approach have already been discussed.
It should be clear, then, why the assessment-sensitivity of normative properties would underwrite (1) above: two properties cannot be identical if they engage in such different behavior. Property identity requires sameness of extension at any possible world, but that requirement will not be met in these cases. With a standard natural property, a given object (at a given world) is either in the extension or it isn’t; with an assessment-sensitive property, that same object (at the same world) might be in the extension on some assessments but not on others. Put another way, a standard natural property can be thought of as a function from worlds to extensions, but an assessment-sensitive property is a function from world/standard pairs to extensions. This difference ensures that a property of the first kind will be distinct from a property of the second.

However, that only shows that an assessment-sensitive property cannot be identified with any objective property. Presumably, most natural properties are objective, but there could be some properties that are both natural and assessment-sensitive. If MacFarlane (2003) is right, for instance, claims about the future (if the future is undetermined) are assessment-sensitive, but those claims undoubtedly still concern fully natural facts. For that reason, establishing (1) requires the following supplementation: two assessment-sensitive properties that depend on different aspects of the context of assessment will also engage in differing behavior, precluding their identity.

This supplementation is exceedingly plausible—the fact that different aspects of the context of assessment are involved in determining the instantiation of two properties means there will always be room for slippage between the two. For any given object, and a given assessment, the object could be in the extension of one of the properties and not in the extension of the other. Once again, full coextension is guaranteed to fail. It is also exceedingly plausible that natural properties, if assessment-sensitive, will depend on the context of assessment in ways that are different from normative properties. It’s difficult to see how a property would be considered natural if it depended on something akin to a standard of taste.

Thus, the assessment-sensitivity of normative properties explains (1). Moving on to (2), the situation is not quite so straightforward. The problem is that it’s not entirely clear what is required for “reductive explanation” of a property. In most cases, this explanatory requirement is treated as stronger than the requirement of identity, permitting one to claim irreducibility even while (empirical) property identities are granted. If that’s right, then (2) simply follows from (1), so no further argument is necessary. But is there some way to see the explanatory requirement as instead a weaker one, such that a property

---

118 This is why assessment-sensitivity is at best necessary for normativity, not sufficient.
119 So-called non-reductive naturalists stake out this sort of position.
could be reductively explained by others even though it is not identical to any of them (or, what is more likely, to all of them in conjunction)?

There is good reason to think this is not possible. After all, it seems to be a minimal requirement of property explanation that one say (in other terms) when and where it is instantiated. There may, of course, be ways of doing this that do not suffice for reductive explanation — noting that salt is instantiated whenever NaCl is, for instance, in which one makes use of an already known empirical identity — but that’s beside the point. What we’re looking for is a way of explaining a property that doesn’t entail (let alone presuppose) any identity claim at all. But if property explanation necessarily involves saying when and where it is instantiated, it looks like identity cannot be avoided. It seems that a property either cannot be explained at all — because it is primitive, perhaps — or it can be shown to be identical to some other cluster of properties.

What about functional definitions? As applied to mental properties, for instance, definitions of this kind are often taken to provide reductive explanations that can nonetheless be coupled with claims of non-identity — that’s why functionalism is an alternative to identity theory. However, it is important to see that while such definitions can be coupled with certain claims of non-identity — generally between mental properties and physical ones — they cannot be coupled with the claims of non-identity that would matter in this context. This is because the mental property being functionally defined is taken to be identical to the cluster of properties that is doing the explaining. So, for instance, believing that p just is being caused by a, b, and c plus causing d, e, and f plus whatever else goes into the functional definition. Once again, reductive explanation brings identity along with it. And if so, then (2) follows directly from (1).

Still, there might remain a nagging suspicion that the very case at hand provides an example of explanation without identity. Let’s look again at the property delicious, accepting for the sake of argument that it is in fact assessment-sensitive. Aren’t we able to explain this assessment-sensitive property in terms of objective ones — e.g. having such-and-such a standard of taste, delicious according to such-and-such a standard, etc.? I think we do in some sense make use of these other properties in giving an account of delicious and saying when and where it is instantiated. In a way, delicious is explained as being constructed from these other properties, each of which captures one portion — one assessment-context component — of the whole.

That said, I think the demand for reductive explanation ultimately cannot be met in this case. The reason is that reductive explanation of a property has important implications for our judgments concerning it. In particular, our

---

120 See footnote 117 for further discussion of this point.
judgments concerning that property can be directly guided by our judgments concerning the other, explanatory properties. For instance, we can find out that something is a college freshman simply by finding out that it is a student and that it is in its first year in school. But things don’t work that way with delicious. Even if we determine both (a) that a certain standard of taste is currently in play and (b) that a particular object is delicious according to that standard, that doesn’t enable us to say that delicious is instantiated. It allows us to say that delicious_current_standard is instantiated, but that’s not quite the same thing.

The basic problem here is that judgments concerning delicious are judgments made from contexts of assessment and with certain standards of taste, not judgments about contexts of assessment and standards of taste. The determination that can be made on the basis of (a) and (b) is one that anyone else could make too, from any other context of assessment. The property involved is, like those proposed by the contextualist, covertly about the judger, and so varies from one judgment to the next. And that means that the single assessment-sensitive property delicious hasn’t been fully captured.

If an assessment-sensitive property were to be reductively explained, it would enable us to determine from within any particular context of assessment that the property was instantiated, simply by determining how things stand with other properties. Perhaps that can be done in some cases, but it won’t be by means of objective properties or by means of assessment-sensitive properties that depend on the context of assessment differently – either way, judgments will potentially diverge as the context of assessment shifts. If normative properties are assessment-sensitive, then, (2) also immediately follows.

The case for (3) is much simpler, and builds on this last point. If one property can be inferred from another, then the two have to be of the same basic kind – either assessment-sensitive (and similarly dependent on context of assessment) or not. If they are of different basic kinds, then their differing behavior across contexts of assessment will break the inferential connection between them. It will not simply follow from the presence of one that the other is instantiated as well.

There are two brief points worth noting here. First, it is not that assessment-sensitive properties can be inferred exclusively from other assessment-sensitive properties; they can be inferred from a group of properties as long as just one member of that group is assessment-sensitive (and similarly dependent on context) as well. This is because that one assessment-sensitive property in the inference base will allow for the whole pattern to be maintained across contexts, the others in the group (presumably standard descriptive ones) serving as a constant background. So this fits nicely with the intuitive judgment that a normative conclusion is viable as long as there is at least one normative premise.
Second, the asymmetry of condition (3) also makes sense on this account. The instantiation of an assessment-sensitive property may well depend (and, I would say, usually does depend) on various objective properties being instantiated as well. Of course, as the context of assessment changes, the instantiation of the former depends on different sets of the latter, which is why the original property cannot be identical with or reductively explained by any particular set. Nonetheless, there might be some properties that are found in all such sets. This makes it possible that, for a given assessment-sensitive property, whenever it is instantiated – i.e. from whatever context of assessment – certain other properties are instantiated too. And this would allow one to infer from the assessment-sensitive property to one that is not (though not vice versa). That, again, fits with common intuitions about normative matters.

It can be seen, then, that the basic claims of the naturalistic fallacy are well accounted for by the thesis that normative properties (and judgments concerning them) are assessment-sensitive. I will end this section by quickly noting how the thesis bears on the two classic paradigms of normativity and the naturalistic fallacy.

Hume was one of the earliest to suggest the fallacy; he remarked that prescriptive judgments express a “new relation” – one involving ought or ought not – that cannot be deduced from straightforwardly natural properties and relations “which are entirely different from it.” Moore made a similar claim about evaluative judgments, saying that they concern properties – like good and bad – that cannot be defined in terms of natural ones. What isn’t made fully clear in either case is exactly how the one sort of property differs from the other, such that the attempted deduction or definition is bound to fail. Assessment-sensitivity provides an answer that can apply to both cases. If the truth of these two sorts of judgment depends on certain standards that vary with contexts of assessment, then the claims made by Hume and Moore are backed up and accounted for. The operative standards might be the same in the two cases or they might not – that’s a matter for further inquiry, regarding the relationship between evaluation and prescription – but the fact that both are assessment-sensitive in one way or another would be enough to underwrite the naturalistic fallacy as applying to them.

In the next section, I’ll focus on Moore’s famous argument for his conclusion, and show how it can be profitably understood in these terms.

4. Explaining the open question argument

I should make clear at the outset that I will not be attempting to show that Moore’s actual open question argument (OQA) should be thought of as involving, even implicitly, the notion of assessment-sensitivity. My claim is only that the
The notion of assessment-sensitivity can play an effective role in an argument of *that basic kind*. This updated version of the OQA establishes the same conclusion but in a much clearer way, and gets around a crucial limitation of the original.

Moore argued, very roughly, as follows. For any normative property, X, and natural property, Y, the question, “this thing is Y, but is it X?” remains open. In other words, the answer to such a question is never *trivially* “yes,” as it would have to be if X were reducible to Y. (Consider the parallel question, “this thing is a first-year college student, but is it a freshman?”) This shows, then, that there is a fundamental gap between the normative and the natural – the conclusion is the naturalistic fallacy.

This argument has struck many philosophers as being onto something important. Not only has the conclusion seemed basically right, but the key argumentative device – the open question test – has seemed like an effective way of reaching that conclusion, a way of revealing something crucial about the relationship between normative properties and natural ones. That said, there are two problems with the argument that are often pointed out.

First, there is the matter of the argument’s scope. As Moore presents it, the argument only rules out *analytic* or *definitional* naturalism. That is, it shows that the meaning of a normative expression or concept cannot be fully captured in non-normative terms. This is presumably all Moore himself wanted the argument to accomplish, and it is by no means a trivial result. However, it provides no reason at all for rejecting naturalism understood, as it often is nowadays, as a metaphysical claim about what properties there are – namely, as the claim that all properties are natural ones. In other words, it is perfectly possible, given Moore’s argument, for normative properties to be identical with natural ones in the same way that salt is identical to NaCl; the OQA may show that the concepts are different in some way, but empirical investigation is still capable of showing that the two concepts pick out just one property, and a perfectly natural one at that.

Second, even when one focuses on the narrower construal of the argument, it is not clear how we know that the key question will *always* remain open. The claim of openness isn’t meant to describe our reaction to *naturalizing* proposals made so far, but to project to all possible proposals. But how can it do that? If the argument is to do more than express a mere hunch, it needs to be supplemented with an explanation of what it is about normative properties that results in this guarantee of openness.

The assessment-sensitivity of normative properties supplies a solution to both of these problems. I’ll begin with the second. Assessment-sensitive properties give rise to Moorean open questions because it is always possible for the context of assessment to shift. So even if an assessment-sensitive property is tightly linked with a certain natural property in the current context, the two might
come apart if a shift were to occur. I contend that it is our implicit awareness of this possibility that accounts for why Moorean questions strike us as open in the relevant (i.e. projective) sense.

A few clarifications are in order here. First, it is important to note that I am not saying that assessment-sensitivity is the only way to account for Moorean open questions. Normative properties could be semantically just like straightforward natural properties – as objective as you like – and still give rise to them. It might simply be a brute fact, something we can give no explanation for, that we take normative properties to be irreducible, or there might be an explanation different from the one I’m offering. (I'll consider one popular proposal in the next section). My claim is just that assessment-sensitivity is one quite plausible way of accounting for Moorean open questions, and that that tells in favor of the thesis that normative properties are assessment-sensitive. The strength of this consideration increases, of course, the bleaker the alternatives look.

Next, I want to say a bit more about the implicit awareness of assessment-sensitivity that I take to account for Moorean open questions. This might seem to be in tension with the discussion up to this point, in which I’ve emphasized the practical indistinguishability of assessment-sensitive and objective discourse. For notice that the mere fact of assessment-sensitivity, if not in any way recognized by us, might be able to explain metaphysical truths like those canvassed in the last section, but would be incapable of explaining the OQA. That argument relies on our sense that these questions remain open, and where would that sense come from if assessment-sensitivity lies beneath the surface? Indeed, one might even think that the picture I’ve sketched so far has the opposite implication, suggesting that someone caught up in a particular context of assessment will judge (or might easily judge) that questions of that kind are closed.

To see why there is no such tension, let’s look again at the Moorean question: “this thing is Y, but is it X?” Yes, the latter clause is an engaged question, one which someone will answer from within a particular context of assessment and with a certain standard, as I’ve said before. The answer to the question may well be “yes,” full stop, and if a different person says otherwise it will be treated as a disagreement. In other words, that question and its answer will be treated in a more-or-less objectivist manner. Moreover, a person’s answer may well depend quite directly on whether various objective properties are instantiated, and may even depend in that way on property Y. (Incidentally, I actually take it to be an advantage of my view that it has all this in place to help make sense of the fact that some people believe, at least at certain times, that these questions are not open and that reduction is possible; that may be more difficult to explain on other views).
The important thing to see is that a definitive “yes” answer has nothing to do with whether the question is taken to be open or not. The key question, really, is a second-order one: “was that answer trivial or, in a way, automatic?” At that point, I think, we are no longer being asked to judge whether X applies, but to reflect on the relationship between X and Y, to step back and assess where our original answer came from. Returning to the suggestion made by Harman, this second-order question requires (or at least inclines) us to take an external perspective rather than an engaged one.

It is at that point that an implicit awareness of assessment-sensitivity comes in. We recognize that, even if a thing is both X and Y, there is more to being X than just being Y. That additional factor, I’m suggesting, is the potentially shifting standard of evaluation on which judgments regarding X depend. And since that potential for shifting is ever present, we sense that Moorean questions cannot be closed simply by modifying or adding to Y; we sense that those questions will always remain open.

This point leads to one last clarification: it is awareness of assessment-sensitivity, and not of standard-relativity more generally, that accounts for the openness of Moorean questions. Standard-relativity, recall, is a feature of contextualism as well as assessment-sensitivity; the difference is whether the operative standards are determined by the context of application or the context of assessment. One might think that the real explanatory work in this and the previous section is done by the more general feature, not assessment-sensitivity specifically.

There is something to this thought. Dreier (1992), for example, argues that a contextualist treatment can make sense of what he calls “the ineliminability of the normative.” However, taking normativity to be assessment-sensitive turns out to provide a far superior explanation of the OQA.121 This is because, according to the contextualist, for any normative claim made in a particular context it is settled once and for all whether the normative property at issue is instantiated or not. As Dreier notes, once a standard of evaluation is fixed by the context of application, a normative assertion is equivalent to one involving perfectly natural properties. Dreier: “Remember that it is not the proposition expressed that is especially normative, on the [contextualist] view, but the way it is expressed” (original emphasis).

This brings us back to the first criticism commonly leveled against the OQA, the one related to its scope. In effect, the OQA has a weak construal and a strong construal. On the weak construal, Moorean questions are open in this sense: someone who is fully competent with terms “X” and “Y” will feel that the

121 This is in addition to the other drawbacks of the contextualist view, such as the problem of disagreement mentioned toward the beginning of this paper.
answer to the question is not trivially “yes.” As already noted, if the OQA is understood this way it has no force against a strictly metaphysical naturalism, for which empirically discovered property identities are perfectly adequate. Contextualism can give a fine account of this version of the OQA. That is because full competence with “X” can be accompanied by lack of knowledge concerning a given context of use. If someone doesn’t know enough about the context to be sure what standard of evaluation is operative, they will be unsure about the answer to the Moorean question as well. So X doesn’t immediately or trivially follow from Y, since additional information is required to establish the link.122

Many people, though, have felt that the OQA has a broader reach, demonstrating that the search for a posteriori identities involving normative and natural properties (on the model of the identity between salt and NaCl) is just as misguided and hopeless as the search for naturalistic definitions.123 This would require a stronger construal of the OQA on which Moorean questions are open in something like this sense: someone can be fully competent with terms X and Y and know all the natural facts related to them and still feel that the answer to the question is not trivially “yes.”

Contextualism cannot underwrite this stronger version of the OQA, but assessment-sensitivity can. On the contextualist view, as long as we know all the natural facts pertinent to a given normative judgment – in particular, the facts about the context that establish a certain standard of evaluation – then we’ll know which fully natural proposition the judgment expresses in that context. Given that knowledge, a Moorean question pairing the two will seem quite trivial indeed. (It would be like the question, “this is NaCl, but is it salt?” asked of someone with minimal knowledge of chemistry).

If normative properties and judgments are taken to be assessment-sensitive, on the other hand, then no matter how much we know about the context we’re in and any relevant empirical correlations that have been discovered, a Moorean question will still strike us as having a non-trivial answer.124 This is because we implicitly recognize that standards might change – our standards might change – such that the answer to the original question will, assessed from a new context, be

122 Rip Van Winkle examples provide the model here, as Dreier makes explicit. For instance, the question, “this happened on June 10, but did it happen yesterday?” is open on the weak construal because someone who has lost track of time would be unable to answer it, even if they are competent with all the terms involved.
123 In any case, many have the sense that this search is hopeless, whether or not the OQA is the source, and it would be useful to have an explanation for that sense.
124 Again, this is best thought of as a fact about our second-order reflection on the question, not a description of our first-order judgment; if we’re just interested in whether property X is instantiated or not, then all that knowledge of natural facts will lead us directly to an answer (within our context of assessment).
different. We implicitly recognize that there is a factor involved in the normative judgment that is not involved in the natural one paired with it, and that factor is not simply additional information that would need to be grasped in order to forge a link between the two.

With that, I hope it is clear that the assessment-sensitivity of normative properties puts the OQA on solid footing – so solid, in fact, that the argument secures the most ambitious result that could be hoped for it. Again, that serves as a reason for accepting my main thesis, since it is of theoretical benefit to have this intuitive argument backed up and clarified. In the next section I will run through some of the advantages this approach has over its main competitors.

5. Two less satisfactory alternatives: noncognitivism and motivational internalism

Up to this point, I have tried to show how the assessment-sensitivity of normative properties and judgments can explain the naturalistic fallacy and the open question argument (in their strongest forms). The need for such explanation has been widely noticed, of course, and other proposals have been offered in an attempt to supply it. I take the two most popular options to be noncognitivism and motivational internalism, and I now want to make clear why the assessment-sensitivity thesis is preferable to them.

The naturalistic fallacy says that normativity cannot be identified with or reductively explained in term of purely natural properties. A noncognitivist approach accounts for this in a very simple way: normative judgments, despite superficial appearances, are not factual judgments about properties or relations at all, so there just isn’t anything there to be identified with or reductively explained by something else. Instead, according to noncognitivism, normative judgments are declarations of approval/disapproval. As it is sometimes put, normative judgments do not express beliefs, like factual judgments do, but other, more desire-like mental states. This would nicely explain the OQA as well: since the first clause of a Moorean question expresses a factual belief, while the second asks how things are with some other, independent mental state(s), the answer to the question will never seem trivial.

125 This is, I think, the standard way to think of noncognitivism, but Gibbard (2002) argues that even a noncognitivist approach is compatible with the claim that a natural property is picked out by normative concepts. In that case, the view is much more similar to what I’ve called non-reductive naturalism in that it allows for a posteriori property identities.

126 Or, in any case, some kind of act that is not truth-evaluable
A full discussion of noncognitivism is well beyond the scope of this paper. I will simply highlight two well-known criticisms of that approach and show that the thesis I am proposing is not vulnerable to them.

The first problem noncognitivism faces is one that came up before in another context – making sense of disagreement. Because noncognitivism denies that normative judgments issue in truth-evaluable propositions, it is unable to account for disagreement as a matter of two judgments that cannot both be true. Attempts have been made to work out a different notion of incompatibility, but they have run into notorious difficulties.\(^{127}\)

I won’t assess these attempts or the objections that have been made to them; I will simply note that the assessment-sensitivity thesis allows one to take the easy way out on this issue. Normative judgments, if assessment-sensitive, are still truth-evaluable; the question of their truth-value is just slightly more complicated. But, as was discussed earlier in the paper, that added complication doesn’t prevent one from giving a more-or-less straightforward account of disagreement in terms of truth-value incompatibility.\(^{128}\)

Noncognitivism also confronts what is known as the embedding problem. If a normative claim is not a factual claim but, say, an expression of approval, then it is far from clear how it can be embedded in larger contexts, such as conditionals, in the same way that factual claims can. Embedding is standardly understood by taking the larger context to be truth-functional, i.e. a function from the truth-conditions of the (embedded) parts to the truth-conditions of the whole. That option obviously isn’t available to the noncognitivist, since an embedded normative claim has no truth-conditions, and trying to see the larger context as a different kind of function is, to say the least, challenging. Consider an example where the antecedent is normative and the consequent is not. What does the conditional function yield – a (different) expression of approval? Could modus ponens be accounted for on that model?

Once again, I’ve done no more than sketch these worries. Still, accepting that normative judgments are assessment-sensitive (rather than noncognitive) permits one to avoid the worries entirely. Assessment-sensitive claims, because they express truth-conditional propositions, will embed just like any other factual claim. Assessment-sensitivity affects the precise nature of those truth-conditional propositions, but not the basic framework.

So here is the overall picture. It is commonly, intuitively recognized that normative judgments are somehow different from run-of-the-mill factual judgments. This pushes many toward some form of noncognitivism. But noncognitivism is a semantic proposal that is quite radically revisionary, and its

\(^{127}\) See Unwin (2001) and Schroeder (2008), among many others.

\(^{128}\) See MacFarlane (2007) for a much fuller discussion.
implications seem to go too far. In other words, it can be objected to on the
grounds that normative judgments are not so different from run-of-the-mill factual
judgments. Assessment-sensitivity occupies the required middle ground. It is in
important ways a less radical semantic revision than noncognitivism, and this
allows it to account for the mostly ordinary behavior of normative claims while at
the same time capturing a crucial element of difference. That element of
difference doesn’t show up much (or at all) in practice – where it would seemingly
have to if noncognitivism were right – but it does show up when we reflect on
underlying metaphysical questions, i.e. precisely when the naturalistic fallacy and
the open question argument come to the fore. So assessment-sensitivity provides
the explanations we need without the unwelcome implications.

The second major alternative is to use the thesis of motivational internalism
as a basis for explanation. Motivational internalism maintains that normative
judgments have a special, direct connection with motivation to perform actions.
So, just to give one simple example, a judgment that something ought to be done
is specially and directly connected with (some) motivation to do it. Now, a more
precise formulation of this special connection is obviously needed – at a
minimum, it must not be mediated by contingent background motivations, since
that would allow all judgments to have it – but at this point it is enough to have the
basic idea.

Motivational internalism, as stated, is perfectly compatible with
noncognitivism, so much so that it is often seen as evidence for it. This is because
noncognitivism, since it treats normative judgments as expressions of desire-like
states, accounts quite straightforwardly for the connection with motivation.
However, since noncognitivism has already been addressed, I will be focusing on
motivational internalism as an alternative proposal. That is, I will treat
motivational internalism as the general claim above (regarding the special, direct
connection) combined with the claim that normative judgments are, semantically
speaking, objective. In effect, the thesis at issue holds that there are objective
normative properties, and that judgments concerning them are inherently
motivational.

This thesis, if correct, could explain the naturalistic fallacy in the following
way: since judgments about normative properties are inherently motivational,
while judgments about natural ones are not, reductive explanation is hopeless;
there will always be a key feature of the former that cannot be derived from the
latter. Similarly, the open question argument makes perfect sense: the additional
motivational feature is relevant to answering the Moorean question, but it clearly
goes beyond whatever is stated in the first clause of it; as a result, the answer to
the question will never seem to follow trivially.
Two drawbacks immediately present themselves, though. First, and very briefly, many people take the core idea of motivational internalism (as I am interpreting it) to be highly dubious. Judgments concerning objective properties, it is thought, simply don’t entail motivations; such judgments interact with our desires and goals, but they do not produce desires or goals by themselves. This is an aspect of Mackie’s well-known argument that normative properties are too queer to be countenanced. At the very least, then, the thesis of motivational internalism is a contentious one. So if it can be shown that assessment-sensitivity does as good or (as I think) a better job explaining the naturalistic fallacy, there will be no pressure to get by, so to speak, with motivational internalism, simply because it is so clearly acceptable in any case.

The second drawback is relevant to this comparison of explanatory capacity: motivational internalism cannot be used to support the stronger version of the naturalistic fallacy, that on which empirical property identities as well as reductive explanations are ruled out. The connection between judgment and motivation might show that there is something distinctive and irreducible about normative concepts, but they could pick out fully natural properties nonetheless. As is well known, two coreferential concepts can play very different cognitive roles; in this case one would play a motivational role in a way the other does not. Setting that point aside, though, and sticking with the slightly weaker version of the fallacy, motivational internalism still has a pretty serious problem. This has to do with the task I mentioned a moment ago: specifying the nature of the special connection with motivation.\textsuperscript{129} The strongest formulations of that connection conflict with basic psycho-behavioral facts. For instance, if the claim is that there is a necessary link between normative judgment and motivation, then instances of weakness of will should not be possible.\textsuperscript{130} Since they are possible, that formulation cannot be right. Yet it’s not clear how one can weaken the thesis without losing its ability to explain even the weaker version of the naturalistic fallacy.

One possibility is to suggest, as Dreier (1990) does, that there is normally a direct link between normative judgment and motivation. This is almost certainly true, but it seems to be true of some naturalistic judgments too. Take the judgment that some item will provide pleasure. That judgment will normally be connected with a motivation to pursue or acquire the item (though of course there will be

\textsuperscript{129} This discussion draws upon Dreier (1990) and Hattiangadi (2007).
\textsuperscript{130} One might respond by saying that there is still some motivation on the losing side in cases of weakness of will; the point I am making would then have to rely on more extreme cases in which there is no motivation to do what aligns with one’s normative judgment (e.g. “desiring the bad” from Stocker (1979)). These cases are certainly quite unusual, but they seem conceptually possible at the very least.
How exactly can one distinguish this case from the normative one? If this case is said to depend on the contingent presence of a background motivation to pursue the pleasurable, the normative case could be said to depend on the contingent presence of a background motivation to pursue the good (or whatever). Notice that the stronger formulation, by contrast, licensed the claim that we are necessarily motivated to pursue the good, something clearly not true of pursuing the pleasurable. But if we look merely at what we’re normally motivated to pursue, all sorts of things appear to qualify.

Here’s another try, due to Smith (1994), Wedgwood (2007), and others: it is necessarily the case that, if one is rational, there is a direct link between normative judgment and motivation. This seems to help with the counterexample I just offered, since it is plausible that one can lack the motivation to pursue the pleasurable without being irrational. But the same kind of problem can arise for this proposal. Take the judgment that doing Z will destroy the universe. Presumably, it is necessarily the case that, if one is rational, there is a direct link between this kind of judgment and motivation (in this case, one hopes, motivation to avoid the action). This kind of judgment may not offer much naturalizing potential, but what about the judgment that doing Z will minimize aggregate suffering? Is this judgment directly linked with motivation, assuming one is rational? The answer is not obvious; some people have certainly thought there is such a link, and simply denying that there is runs the risk of begging the question with regard to the naturalistic fallacy.

These problems exhibit a pattern. The “if one is rational” and “normally” clauses serve to cut down the number of cases to be examined, thereby excluding certain clear counterexamples (like weakness of will) to the strongest formulation of motivational internalism. But by doing so, they also open up the possibility that other judgments will meet the restricted condition. This works against the alleged explanation of the naturalistic fallacy (weaker version), since there is no longer any conceptual guarantee that normative judgments have a feature that naturalistic ones lack. So while these weaker construals of motivational internalism may be correct – the arguments here do not suggest otherwise – they cannot be used to perform the explanatory job that is of interest to us.

In conclusion, the thesis that normativity is assessment-sensitive has clear advantages over both noncognitivism and motivational internalism when it comes to explaining the naturalistic fallacy and the open question argument. Unlike noncognitivism, it allows us to see why there is a naturalistic fallacy while also allowing us to maintain various common-sense views about normative judgments: that they are truth- evaluative, that we can have real disagreements about them, that

---

131 Is this the same notion of normal at work? Perhaps not, but in that case the challenge is to say what is distinctive about the sense of “normal” used in the thesis.
they embed completely normally in larger contexts, etc. Unlike motivational internalism, there is no obscurity in how it accounts for the irreducibility of normative properties, and it is also capable of explaining, as motivational internalism is not, the strongest version of the naturalistic fallacy, which rules out *a posteriori* property identities.

6. Implications

To this point, my aim has been to provide support for the thesis that assessment-sensitivity is a necessary feature of normativity. I have done so by showing that the acceptance of that thesis has substantial theoretical benefits. In particular, it validates and explains the intuitively compelling naturalistic fallacy and open question argument, and it does so more effectively and with fewer associated drawbacks than the main explanatory alternatives. It can also be applied in the philosophy of mind to validate and explain the Davidsonian claim that there can be no psychophysical laws. I now want to examine one way in which the thesis can be put to further use.

We can begin by remarking that a certain more direct defense of the thesis is not feasible. Namely, we cannot confirm the thesis by simply getting a list of all the normative properties and judgments and showing that they are all (plausibly) assessment-sensitive. This is because it is a matter of some dispute precisely which properties and judgments should go on that list in the first place. There are two paradigmatic examples of normativity – prescriptive judgments involving *ought* or *ought not*, and evaluative judgments involving *good* or *bad* – to which the naturalistic fallacy has historically been taken to apply, and it would indeed tell against my thesis if these judgments were clearly not assessment-sensitive. I don’t think there is this disconfirming evidence, but I will not pursue the matter here. I will just assume that an assessment-sensitive construal of these paradigm cases is viable, and that it provides the best explanation for the *sui generis* character they seem to have.\(^{132}\)

But if the thesis is accepted – on the basis of the explanatory benefits I’ve laid out combined with its plausibility as applied to the paradigm cases – it can then be put to use in determining which *other* properties and judgments are normative. If it is questionable whether or not a certain property or judgment is normative, the thesis can serve as a kind of test, offering one clear way to reach a negative verdict. That is, if the property or judgment in question is not

\(^{132}\) Similarly, in the discussion of Davidson’s argument I assumed that *rationality* was indeed normative and that it could be plausibly construed as assessment-sensitive. If that turned out not to be true, the argument obviously wouldn’t go through. My purpose in that discussion, though, was to show only that the argument could be made sense of in those terms, not that it had to be correct.
assessment-sensitive, then it fails to meet a necessary requirement for being normative.

I believe this is precisely how things stand with regard to the property of \textit{semantic correctness}. It is quite commonly claimed that semantic correctness is normative, and that, consequently, a naturalistic account of content is not possible.\footnote{This line of thought traces back at least to Kripke (1982), and is endorsed by …} I think, on the contrary, that semantic correctness fails the assessment-sensitivity test for normativity, so the proposed argument against content naturalism is unsound. This is not to say that a naturalistic account of content \textit{is} possible, just that normativity cannot be the reason it is not.

Why is semantic correctness often thought to be normative? For one thing, it seems to involve a prescriptive dimension, in that a given concept, if it applies \textit{correctly} to Xs, \textit{ought} to be applied to an X and \textit{ought not} to be applied to a non-X. Relatedly, it seems to involve an evaluative dimension as well, since correctness is clearly something positive, an achievement, while incorrectness is clearly something negative, a defect. Both of these claims are problematic as they stand and would need to be refined at the very least, but I will not address them further. I will instead move on and apply the test provided by the assessment-sensitivity thesis. This will show that semantic correctness cannot be normative.\footnote{It is perhaps worth issuing a reminder here that I am concerned in this paper only with a notion of normativity that carries with it anti-naturalist implications. There may be some other sense in which semantic correctness is normative, but if that sense entails no naturalistic fallacy it is beside the point. See Hattiangadi (2007) for a distinction along these lines between two senses of normativity.}

Consider first an ordinary, objective, naturalistic judgment, such as, “Earth’s tallest mountain is less than 10,000 meters tall.” The common view, which I’ve been assuming, is that this judgment is not assessment-sensitive, for its truth value does not and cannot vary as the context of assessment changes. Rather, the judgment expresses a proposition that is, in this instance, true (at this world), and its truth holds for anyone assessing it. Someone who (wrongly) disagrees with this judgment denies that very proposition. Given my framework, anyone who takes there to be a naturalistic fallacy is committed to this common view, since otherwise there would be no contrast between naturalistic judgments on the one hand and assessment-sensitive – in particular, normative – judgments on the other. The question is whether someone can accept this common view about judgments like this and not also accept it with regard to judgments about semantic correctness. I believe they cannot.

Of course, some disagreements we find ourselves in arise due to differences in how the terms involved in the judgment are understood (or, equivalently, differences in the concepts taken to be applied). For example, someone might
disagree with the above judgment because they take the concept mountain to apply to distinctively shaped masses of rock plus 1,000 vertical meters of the air above them (something like a nation’s airspace).

This might give the superficial impression of assessment-sensitivity – can’t both disputants be correct, given where they are coming from? – but it is important to see how this case is different, on the common view. On that view, either two separate propositions are expressed – one involving the ordinary concept mountain, the other involving an idiosyncratic variant – which is why they can both be true, or there is a single proposition at issue – perhaps because facts about social usage establish that the ordinary meaning is the correct one – in which case the person denying the judgment is plainly wrong. Either way, a disagreement due to this kind of meaning discrepancy is not a case of there being a single proposition that is true from some contexts of assessment and false from others.

How, then, do things stand with judgments concerning semantic correctness, such as the following: “the concept mountain applies correctly to that thing” (said while demonstrating Mount Everest, let’s say)? Once again, we cannot take this judgment to be assessment-sensitive simply because disagreements might arise as a result of different ways of understanding the concept being judged. If there is a difference of this kind, then either two separate propositions are expressed by the two judgments, or there is a single proposition expressed and one of the disputants must be wrong. Either way, once the proposition is fixed, the default view looks to be correct: that proposition is simply true or not – either the thing demonstrated is in the extension of mountain (as it figures in that proposition) or it isn’t – and no change in the context of assessment will affect that.135

The support for this default view goes beyond surface appearances too. When it comes to more plausibly assessment-sensitive judgments – we can return to our original example, “Indian food is delicious” – there seem to be good reasons for introducing standards of evaluation that can vary without constituting changes in meaning of the kind we’ve been looking at.136 But no such standards can be introduced with regard to semantic correctness, as a result of the very tight connection between correctness conditions and the meanings of terms. Any variance in standards of correctness will either (a) collapse into a difference in meaning at the ground level, which, as we’ve seen, means there is no assessment-sensitivity anywhere, or (b) push assessment-sensitivity down to the ground level as well, which, as I’ll argue, leads to an unsustainable picture.

135 The meaning of a term might change, of course, in which case later judgments of the same form could be assessed differently, but the truth value of any particular proposition remains always the same.
136 Dreier, in his contextualist treatment of normativity, explains this in terms of Kaplan’s (1989) character/content distinction, and something similar can be adduced here as well.
In some cases, meaning can remain constant even while correctness conditions (extensions) change in systematic ways. This is obviously true of classic indexicals like “I” and “here,” and is perhaps true of assessment-sensitive terms like “delicious” and “good” as well. Standardly, though, a term’s meaning and its correctness conditions go more closely together. And since the concept of semantic correctness is used in stating such meanings, a variance in standards of correctness will result in differing extensions, and therefore different meanings, for the terms involved. That just brings us back to the previous example of a disagreement about the claim, “the concept *mountain* applies correctly to that thing there.” One cannot say that this expresses a single proposition, and that the involvement of correctness allows for different truth values when it is assessed from different contexts – there would no longer be any sense in which the initial noun phrase contributed the same thing to the proposition in both cases. Rather, as we saw, this case needs to be interpreted as involving two separate expressed propositions, each incorporating a slightly different concept. So here the differing standards of correctness seep down into different first-order meanings, and the assessment-sensitivity model fails to apply.

But perhaps there are shifting standards that would not have this consequence. The most likely candidate might be varying standards of strictness. So, to vary a classic example, “the concept *hexagonal* correctly applies to France” could be thought to be false as assessed from some contexts, which involve relatively strict standards of correctness, and true as assessed from others, in which standards are relatively lax. If this were right, though, the same situation would have to apply to the judgment, “France is hexagonal.” In other words, allowing varying standards of strictness with regard to semantic correctness requires that one allow varying standards of strictness with regard to the first-order concepts as well. In which case, if semantic correctness is assessment-sensitive, all concepts are.

So, in the end, really maintaining that semantic correctness is assessment-sensitive would require adopting a position that is, at the very least, quite revisionary. More importantly, from the current perspective, that position is incompatible with the acceptance of a naturalistic fallacy applying to semantic correctness, for it leaves no distinction between the two types of judgment – normative and naturalistic – that the fallacy presupposes.\(^{137}\) This is a serious problem, since making use of the naturalistic fallacy was the whole point of arguing that semantic correctness is normative. The clear implication is that

---

\(^{137}\) But couldn’t the two types of judgment still be accounted for in terms of the different ways in which they depend on the context of assessment? Strictly speaking, yes. However, the way in which semantic correctness would depend on context of assessment seems to be the same as that in which “naturalistic” concepts would. So the distinction still could not be made to apply in the case at issue.
semantic correctness is not assessment-sensitive, and so not normative. The naturalistic fallacy simply does not apply as a result of that feature of terms and concepts.

I want to end by noting, very briefly, a second implication of the thesis that normativity necessarily implies assessment-sensitivity. As I’ve presented it, that thesis serves to underwrite the naturalistic fallacy – it helps to establish the claim that normative properties and judgments cannot be naturalized. However, by shifting the explanatory burden from normativity per se to one necessary (but not sufficient) aspect of it, the thesis in a way mitigates normativity’s anti-naturalistic importance. A way of seeing this point is to note that delicious, for example, cannot be naturalized either, and for precisely the same reasons. In other words, though there is indeed something special about normativity that keeps it at arm’s length from naturalistic explanation, it is not something particularly unique.