7

Getting a Moral Thing into a Thought

Metasemantics for Non-Naturalists

Preston Werner

Non-naturalism is the view that normative properties are response-independent, irreducible to natural properties, and causally inefficacious.[1] Sharon Street and Richard Joyce, by arguing that a non-naturalism entails skepticism, have ushered in a virtually insurmountable literature on non-naturalist moral epistemology and evolutionary debunking arguments.[2] These concerns are not new.[3] Extensions of the so-called "Benacerraf problem" for mathematical entities to the moral realm have also attempted to undermine non-naturalism.[4] And there remains no consensus on how—or whether—these epistemological challenges can be met.

A less discussed question for non-naturalism concerns the metasemantic connection between normative beliefs and the normative facts. This under-explored question concerns the metasemantics of normative terms. Ideally, the non-naturalist could remain ecumenical. As has been noted, however, this is not possible.[5] So the challenge is for the non-naturalist to find some independently motivated metasemantic view that fits well with non-naturalism. Let's call this challenge—which is developed in more detail below—the metasemantic challenge.

The widely discussed epistemological challenges and the metasemantic challenge are at least superficially related. They both raise questions about the status of our normative beliefs if non-naturalism is true. Epistemological challengers argue that non-naturalism threatens to make all normative beliefs unjustified. Metasemantic challengers argue that non-naturalism

[1] Ridge (2014). [2] Joyce (2001, 2006), Street (2006).
[3] See e.g. Ruse and Wilson (1986).
[4] See e.g. Benacerraf (1973), Harman (1977), Liggins (2010), Enoch (2011: ch. 7), Clarke-Doane (2017).
[5] Schroeter and Schroeter (2014), Suikkanen (2017).

Preston Werner, *Getting a Moral Thing into a Thought: Metasemantics for Non-Naturalists* In: *Oxford Studies in Metaethics Volume 15.* Edited by: Russ Shafer-Landau, Oxford University Press (2020).
DOI: 10.1093/oso/9780198859512.003.0007

makes it impossible for us to have beliefs about normative properties and facts at all, since it is impossible for our beliefs to fix onto non-natural properties as their referent.

This chapter focuses on the second challenge. My suggestion is that non-naturalists endorse an epistemic account of reference determination of the sort recently defended by Imogen Dickie, with some modifications.[6] An important implication of this account is that, if correct, a fully fleshed out moral epistemology will simultaneously rebut metasemantic objections to non-naturalism. Thus, the two challenges in effect amount to one.

In section 7.1, I recap two influential epistemological challenges to non-naturalism, as well as the less discussed metasemantic challenge. In section 7.2, I first review a few standard metasemantic theories, illustrating why they spell trouble for the non-naturalist (7.2.1–7.2.2). I then discuss Laura Schroeter and François Schroeter's "normative connectedness model" to see whether it does any better; I conclude that it does not, and for reasons that will generalize to any internalist metasemantics for non-naturalism (7.2.3–7.2.4). In section 7.3, I assess whether the doctrine of reference magnetism may supplement any of the previous accounts, especially given that some recent work on metasemantics for moral realism has attempted to make use of them for responding to a variety of objections. In section 7.4, I develop my positive epistemic theory, which is well-suited to avoid all of the problems with the traditional metasemantic theories when extended to non-natural normative properties. Section 7.5 draws out three ways in which the epistemic theory accords nicely with widely discussed features of non-naturalism.

7.1. EPISTEMOLOGICAL AND METASEMANTIC OBJECTIONS TO NON-NATURALISM

The so-called Benacerraf problem goes back at least to Paul Benacerraf's (1973) "Mathematical Truth."[7] While Benacerraf was concerned with the case of mathematical knowledge, a similar sort of problem arises for non-naturalist moral realists, according to which moral properties are not causally efficacious.[8] This parallel was first noticed by Gilbert Harman (1977). As in the mathematical case, even while causal conditions on knowledge

[6] Dickie (2015, 2016).

[7] Field (1989), Cheyne (2001), and Liggins (2010) for discussions of this point and developments of Benacerraf's problem.

[8] E.g. Heathwood (2015: 3), McGrath (2014: 186), and Scanlon (2014). Oddie (2005) is an exception to this general rule.

have fallen out of favor, there is still a widespread sense that explaining our epistemic access—in terms of responsiveness to the moral facts—is a serious concern for the non-naturalist moral realist.

Another more specific sort of undercutting defeater for moral knowledge comes in the form of evolutionary debunking arguments (EDAs). At the most abstract level, the idea of an EDA is to provide a genealogy of our moral beliefs or belief-forming mechanisms in terms of the fitness-enhancing evolutionary nature of those beliefs or belief-forming mechanisms. Such an explanation, the debunker argues, will make no appeal to any mind-independent moral facts. And the lack of explanation here strongly suggests skepticism which serves to undercut any justification we initially had about stance-independent moral facts.

I won't say more about how the details of these arguments are best worked out. Instead, let's turn now to the metasemantic challenge. David Enoch provides a concise statement of the problem:

> [According to non-naturalists] the word "good" (to pick one example) refers—at least in some of its occurrences—to the property *goodness*. And this property is, on [non-naturalist views], causally inert and response-independent. How is it, then, that our word manages to latch onto that property—rather than some other property, or perhaps no property at all?[9]

We'll consider more specific metasemantic theories in the next section. For now, just notice the intuitive thought behind the worry. I have a cat named Zooey. As a result, I have a bunch of beliefs about Zooey: that he's lazy, has a cute nose, is mean to other cats, etc. Somehow these beliefs all latch onto *Zooey*, rather than my partner, other cats, or anything else.[10] It's the task of a metasemantic theory to tell us exactly how this works. But whatever the details are here, it's intuitive that my Zooey-beliefs bear all sorts of relations to *Zooey* that don't appear to hold in the normative case, if non-naturalism is true. *Zooey* is frequently part of my perceptual experiences, there is a causal chain between my Zooey-beliefs and *Zooey*, my linguistic community's patterns of "Zooey"-tokenings converge on *Zooey*. But it is unclear that any of these sorts of relations hold (or even could hold) between our normative beliefs and any non-natural properties. So there is a *prima facie* metasemantic challenge for non-naturalists. And this *prima-facie* challenge is made more formidable by noting that many mainstream metasemantic theories make reference to non-natural properties (as they're traditionally construed, at least) not just challenging, but impossible.

[9] Enoch (2011: 177).

[10] I use italics, here and throughout, to refer to properties and objects in the world, as opposed to words or concepts.

One clarification is in order. Metasemantic theories are ambiguous between two possible theoretical inquiries. First, a metasemantic theory could be attempting to answer the question of how words latch onto properties and objects. Second, a metasemantic theory could be attempting to answer the question of how concepts—that is, psychological representations—latch onto properties and objects. The metasemantic challenge to non-naturalism has been framed both ways, without clear differentiation.[11] I will be speaking in terms of mental content and psychological concepts in what follows. All of what I say could in principle be extended to linguistic content as well.

7.2. NON-NATURALISM AND TRADITIONAL METASEMANTICS

Three popular families of metasemantic views are (a) causal/teleological theories, (b) conceptual role semantics, and (c) neo-descriptivism. I briefly explain why these traditional views are not promising for the non-naturalist (7.2.1–7.2.2), before discussing Schroeter and Schroeter's "connectedness" model in more detail (7.2.3–7.2.4).

7.2.1. Causal and Teleological Theories and Non-Naturalism

Causal theories of content all share a commitment to the thesis that concepts get their content in virtue of some causal relation(s) that hold between tokenings of the concept and some property, object, or individual in the world. More formulaically:

> SCT: For any concept C and any property F, C refers to F iff tokenings of C tend to be caused by F.

SCT is open to interpretation between a range of different causal relationships. However, none of these details need to be considered for the present purpose. Any causal theory of content will be a non-starter for the non-naturalist, since the non-naturalist denies that normative properties are causally efficacious. In fact, the causal theory of content has been exploited by naturalist moral realists to ensure that the moral terms pick out natural (causally efficacious) properties.[12] We can set aside causal theories of all

[11] See e.g. Wedgwood (2007), Enoch (2011) (and the quote above), and Schroeter and Schroeter (forthcoming).

[12] Boyd (1988, 2003), Sturgeon (1998).

types—they will not provide a non-naturalist-friendly metasemantics without radical alteration in the commitments of non-naturalists.

Close cousins of causal theories of content, teleological theories, are in principle better placed to avoid the above incompatibility with non-naturalism. On teleological theories, "what a representation represents depends on the functions of the systems that produce or use the representation."[13] "Function" here is understood in terms of what the particular state or process was selected for. So, for example, a given concept C refers to *poutine* just in case (and because) it was selected for the advancement of my broader psychological systems' *poutine*-related goals.

Unlike with causal theories, functions need not refer to causally efficacious properties. For example, evolution selects for creatures who are able to grasp mathematical concepts such as addition and division. These concepts play a certain function in allowing us to reason in a variety of ways that increase our survivability. So far so good for the non-naturalist. But despite the slight improvement on causal theories, non-naturalists—or at least the vast majority of non-naturalists—should still reject a teleological metasemantics for normative concepts. To see why, consider how the teleologist will determine the content of a given normative concept. In order to answer this question, the teleologist will look at what best explains the function of normative thoughts. Neil Sinclair (2012) has perhaps done the most to flesh out what a teleosemantics might look like for normative thoughts.[14]

Sinclair argues that moral concepts have an evolutionary function, which is to coordinate interpersonal behavior.[15] On this sort of story, morality evolves in order to solve certain pervasive evolutionary bargaining problems. It should already be clear that this story about the function of our moral concepts, and any similar such story, will be incompatible with non-naturalism. Supposing Sinclair's story is right, the function of moral concepts is to encourage certain fitness-enhancing behavior. A behavior's being (or not being) fitness enhancing is surely a natural property. For this reason, it is unsurprising that no non-naturalists have appealed to teleosemantics in their metasemantic theorizing.[16]

[13] Neander (2012: introduction).

[14] Of course teleologists have discussed weaker notions of normativity for a very long time (see e.g. Millikan 1984, and esp. Millikan 1995). I hope the difference between a normative thought related to an organism's survival and a robustly moral thought is clear enough. Certain kinds of naturalists won't see a difference in kind between these two notions of normativity, but rather a difference of degree. But we can set this aside, since the present chapter is only concerned with non-naturalists.

[15] But see also Millikan 1995: section 5) for a relevant precursor.

[16] Notice here that this forecloses a strategy analogous to third-factor responses to epistemological objections to non-naturalism. Showing a co-extension between the

7.2.2. Conceptual Role Semantics

Conceptual role semantics for normative terms has received the most discussion in the metaethical literature, largely because of Ralph Wedgwood's influential book *The Nature of Normativity*, as well as his subsequent work.[17] The basic idea is explained concisely by Laura Schroeter and Francois Schroeter:

> Like expressivists, Wedgwood thinks the open question argument suggests that the central element in the meaning of moral terms is their action-guiding role. Indeed, according to Wedgwood, grasping the action-guiding role of those terms is all there is to understanding their meaning. But Wedgwood embraces cognitivism: he thinks that the conceptual role of moral terms provides the resources to single out genuine properties as their semantic value . . . the action-guiding role of moral terms, he suggests, suffices to determine which property they pick out.[18]

Wedgwood takes the following to be a conceptual truth:

> (1) Necessarily, if one is rational, then, if one judges "I ought to ϕ," one also intends to ϕ.

As a conceptual truth, anyone with a full grasp of OUGHT is in a position to know (1). More importantly, though, anyone with a full grasp of OUGHT will be disposed to make the relevant transitions from ought-judgments to intentions, whether implicitly or explicitly. So this ensures, at the very least, that OUGHT is connected up with an action-guiding property. But this is not yet enough to connect OUGHT up with a non-natural property, much less the correct one. It only shows that there is a certain relationship between "ought" judgments and intentions to act. Somehow it must be the case that this action-guiding concept connects up with a non-natural property, if Wedgwood's story is to vindicate a non-naturalist metaethics.

A conceptual role semantics for non-natural properties must connect our concepts not just to the prescriptive role of the properties, but also their representational role. We can see this by considering an agent, Anya, who has some concept O*. O* meets the conditions of (1). Whenever Anya judges that she O* to ϕ, and she is rational, she forms the intention to ϕ. But furthermore, suppose that the evidence that Anya relies on to form her O* judgments are facts about causing as much pain and suffering to bunnies

properties our concepts evolved to track and the non-natural properties does not secure reference onto the non-natural.

[17] Wedgwood (2007).

[18] Schroeter and Schroeter (2003: 191).

as possible. Anya may even take the following to be a conceptual truth—or at least something of a platitude:

> (2) Necessarily, if an action ϕ would result in more pain to bunnies than any alternative action, then I O* to ϕ.

We can call (2) the "Input" condition for O*,[19] and (1) the "output" condition. Anya's O* is perfectly coherent, but surely we don't want to say that O* picks out the non-natural property of *oughtness*, unless we're gravely mistaken about what our obligations are.

But it's not just that Anya's concept is normatively problematic. Rather, it's that (1) radically underdetermines the relevant property picked out by our OUGHT concept. As Neil Sinclair (2018) points out in a closely related context, "it is implausible to suppose that a concept refers to a worldly property when competence with that concept is compatible with no disposition whatsoever to be sensitive to any worldly property."[20] The conceptual role semanticist must provide—or at least gesture toward—what I'm calling an input condition on our normative concepts. An input condition is one's disposition to apply the concept in appropriate circumstances. (For example, the input condition for MODUS PONENS will be premises of the form {P, P→Q}.) And this input condition must be close enough to uniform across agents such that the OUGHT concept is shared amongst all competent users of the concept (lest we collapse into relativism).[21]

7.2.3. Neo-Descriptivism, Connectedness, and Non-Naturalism

Descriptivist theories all share the core idea that a concept gets its content by its association with a set of application conditions, which are shared by any competent user of the concept. Call *internalist* any view such that the extension of a given concept for a given agent is determined wholly by (actual or counterfactual) mental states of the agent. The latter feature of descriptivism—that competent users must grasp the application conditions associated with a concept—renders descriptivist theories paradigmatically internalist. Notice that internalism should be initially attractive to the non-naturalist. This is because it allows the normativity of our normative concepts to be built into their meanings. If, as non-naturalists think,

[19] In the Conceptual Role Semantics for logical operators, what I'm calling the input condition is commonly called the introduction rule.

[20] Sinclair (2018: 113).

[21] For compelling arguments distinct from (but related to) the argument of this paragraph, see Lenman (2010).

normative concepts refer to genuinely normative properties or nothing, then we don't have to worry about our normative concepts fixing onto properties which are intuitively not action-guiding.

As is widely known, traditional descriptivist theories have serious problems. So let's just consider the neo-descriptivist theories that evolved to avoid these problems. The key strategy here is to make use of idealization—a subject's concept refers to the property she would pick out once suitably idealized with respect to some relevant set of facts. Neo-descriptivist theories like this have made some appearances in the metaethical literature, most notably perhaps in Jackson and Pettit's (1995) naturalist-friendly "moral functionalism." Perhaps the most sophisticated version of this kind of view is Schroeter and Schroeter's "normative connectedness model."[22] Because this latter model is the most plausible neo-descriptivist theory for non-naturalists, and because the objection I give below will apply to any neo-descriptivist view that makes use of idealization, I will speak in terms of the normative connectedness model for simplicity.

The theory that Schroeter and Schroeter propose to match their connectedness model attempts to shore up the problems with both neo-descriptivist and teleological theories by moving to a fundamentally relational account. As they summarize it:

> The connectedness model relies on a tradition-based [metasemantic] theory: the fundamental units to which semantic values are assigned are not token elements of thought considered in isolation . . . but rather an entire representational tradition . . . On this approach, the metasemantic theory seeks to assign a univocal semantic value to the tradition as a whole, taking into account the understanding, environment, and history of the entire diachronic and interpersonal tradition. Token elements of thought then inherit their semantic values from the traditions to which they are bound.[23]

A token element of thought inherits its semantic value from its whole historical and social use. This captures the idea that a concept's meaning in one person's mouth should have a lot in common with previous uses of the concept and provides a check on a more radical semantic internalism. How, then, does the tradition help to determine the referent of some particular concept? Here is what Schroeter and Schroeter say:

> [T]he idea that the semantic value of normative concepts is fixed by some original baptismal event is highly implausible . . . The correct principles . . . are simply an

[22] Schroeter and Schroeter (2014).

[23] Schroeter and Schroeter (2014: 12). They use "determination theory" for what I'm calling a metasemantic theory.

idealization of a subject's own reflective methods for refining her understanding of the precise subject matter of her words and thoughts . . . In effect, we can build our [metasemantic] theory from the first-person reflective epistemology of the topic in question.[24]

The difference between the normative connectedness model and traditional neo-descriptivism is that the idealization process includes information about the tokenings of the concept that the subject will count as tokenings of the *very same* concept that the subject sees herself as using. (This is the sense in which the metasemantic theory takes the semantic tradition as prior or at least essential to the determination of the concept's referent.) This provides a non-trivial restriction on an idealized subject's assessment of the referent of a given concept. Her assessment cannot be chauvinistic in the sense that her conceptual tradition was so radically epistemically and reflectively mistaken that their use of the concept was inherently defective. The idealized subject is then given constraints on her interpretation of the concept that are not present in other neo-descriptivist theories.

7.2.4. Neo-Descriptivism, Connectedness, and a General Lesson

Whatever the strengths of the normative connectedness model, and neo-descriptivist models more generally, the idealization strategy will fall into a principled problem for non-naturalists, or at least the vast majority of them. To see why, first notice that non-naturalists accept:

> *Metaphysical Autonomy*. The normative facts are irreducible to the natural facts, in the sense that there is no conceptual entailment or complete metaphysical explanation from the natural facts to the normative facts.

Metaphysical Autonomy is a way of cashing out the irreducibility claim that is so central to non-naturalism.[25] It is arguably closely related to Moore's open question argument[26] and Hume's is–ought gap,[27] as well as to the supervenience objection to non-naturalism.[28] Importantly for our purposes, the Metaphysical Autonomy implies the following:

[24] Schroeter and Schroeter (2014: 13–14).

[25] Thanks to Aaron Elliott for helpful conversation on how to frame Metaphysical Autonomy.

[26] Moore (1903). [27] Hume (1975).

[28] Blackburn (1984), McPherson (2012), Elliott (2014).

> *Epistemological Autonomy*. An agent could be wholly procedurally rational and fully informed about all the natural facts and yet ignorant or mistaken about the normative.[29]

Given the lack of an entailment relation between two domains, it should be clear that knowledge of one domain is not guaranteed to improve knowledge of the other. In practice, this may seem puzzling—of course gathering more information about the natural facts assists us in drawing new and improved normative conclusions! But notice that this sort of knowledge will always be inferred from some purely normative claim, such as that suffering is bad. This is really just a lesson of Hume's is–ought argument reiterated.

Given the autonomy theses, we can show that the normative connectedness model and neo-descriptivist theories more generally face a dilemma. We need a story about what facts get fed into the idealization process. Either this idealization process will include the normative information or it will not. Notice that this idealization process must meet two constraints. First, the information in question should remove any ignorance about the information relevant to determining the concept's referent. This information is relevant because it is required to illustrate the ways that non-idealized agents' concepts can refer to a determinant property even in light of ignorance, in virtue of the fact that their concepts *would* fix on these properties once ignorance is eliminated. This can ensure that a subject's WATER concept refers to *H2O*, even if her non-idealized self is disposed to mistake *XYZ* for *H2O*. Her concept refers to *H2O* just in case she would retract her judgments that instances of *XYZ* fall under her WATER concept under idealization. Second, the information included in the idealization process cannot make reference to the referents of the concept in question, on pain of circularity. So, for example, idealizing for determining the semantic value of WATER should not include facts about *water* (qua *water*), but merely facts about the distribution of clear potable liquids that fill the lakes and rivers, etc. Without this, the idealization process is just smuggling in the fact that *water* is the semantic value of WATER, which is circular.

So first consider an idealization process which doesn't include normative information. On this view, we feed into our idealization an ideal base-level description of all of the natural facts. While such an ideal base-level description will plausibly ensure a referent for natural properties, such a story won't work for normative properties considered as non-natural. When it comes to

[29] As Hille Paakkunainen pointed out to me, on some views of what it takes to grasp the normative concepts, Metaphysical Autonomy is compatible with the rejection of Epistemological Autonomy (Setiya 2012). I think these views have their own problems, but such a discussion would take us outside of the scope of this chapter, so I set them aside.

a property whose truth conditions are not grounded in an ideal base-level description of the physical world, an idealization with respect to these facts will not remove any ignorance present in the non-idealized agent, and so won't fix the concept onto a determinant property. And *Metaphysical Autonomy* entails that an ideal base-level description of the physical world will *not* (alone) determine the truth of normative facts. As such, given that the idealization base does not include any normative bridge laws, even wholly rational agents with full information about the physical world are not guaranteed to fix on a unique non-natural fact. Theories that rely on idealization of this sort—including neo-descriptivism and the connectedness theory—will all fail for this reason. Schroeter and Schroeter's theory of *de jure* sameness may be able to show that a certain set of token instances of a concept all refer to the same normative property (if they refer at all), but it will leave open the question of *which* property is the concept's semantic value.

This objection suggests a very natural reply. The notion of an ideal base-level description need not be restricted to a description of only physical facts. Such an understanding of what the ideal base-level description of the world would look like is only motivated by a pre-existing commitment to physicalism (the claim that all facts are reducible to physical facts). Furthermore, there is precedent for understanding the ideal base-level description as more than just physical facts. For example, property dualists would want to add phenomenological facts to the ideal base-level description.[30] So for the property dualist, fixing the semantic value of the concept of PHENOMENAL-RED requires acquaintance with a certain phenomenological fact, and so acquaintance with such a fact will be part of the ideal base-level description. So, it may be argued, the non-naturalist should embrace the same sort of strategy. The strategy involves including what I'll call "base-level normative" facts in the ideal base-level description of the world. Thus, idealized agents would have access to the information they need to fully determine the referents of their normative concepts.

A worry about this approach can be seen from considering exactly how this set of "base-level normative facts" might look. An obvious place to look is at the normative facts that metaethicists have argued are fundamental. One influential position in this literature is that *reasons* are the fundamental normative kind, and thus all other normative truths are grounded in *reasons* facts.[31] Others have argued that it is not *reasons* but some other set of normative facts that are fundamental, but nothing I'll say

[30] See e.g. Chalmers (2012: ch. 3).

[31] See e.g. Scanlon (2014).

in what follows depends on which position is right.[32] I speak in terms of *reasons* for simplicity.

Suppose the base-level normative facts are just the facts about *reasons* described using the agent's concept of REASON. Since the semantic value of the concept REASON is one of the very things we need a metasemantic theory to fix, building up the base-level normative facts from claims that involve the concept of REASON would result in obvious circularity. What's required for idealization to work for the non-naturalist is some way of understanding a base-level description of the normative facts that doesn't invoke one of the very concepts it's intended to illuminate. And it's unclear how to do so.

I don't deny that some way of characterizing a base-level description of normative information could avoid this problem. Notice that there is a parallel here in the case of natural facts. We will need a metasemantic theory of the concepts that figure in the base-level description of the physical world as well. At this point, one could go externalist, or perhaps one could interdefine a cluster of fundamental base-level concepts and then Ramsify over those concepts. However this might work in the case of natural facts, it won't work in the normative case.[33] I haven't given an argument to illustrate that avoiding this problem is impossible. But it appears to be a serious problem. What the non-naturalist needs, then, is either a solution to this problem, or an alternative metasemantic theory that avoids this problem as well as the problems for the other theories given above.

7.3. REFERENCE MAGNETS TO THE RESCUE?

Before turning to my positive proposal, it's worth briefly discussing reference magnets, which have a rich history of solving (or purporting to solve) metasemantic problems.[34] More specifically, several metaethicists have also made use of reference magnets to make theoretical progress.[35] The central theoretical role that reference magnets are supposed to play is to constrain reference in cases of (sometimes radical) underdetermination. Suppose a theory of reference entails that a particular concept has multiple eligible referents. A theory of reference may entail, for example, the facts don't decide between *greenness* and *grueness* as the referent of GREEN. Reference magnets provide a further constraint on reference to eliminate all but one

[32] See e.g. McHugh and Way (2016), Howard (forthcoming).
[33] See section 7.3 for discussion of this point.
[34] Merrill (1980), Lewis, (1984), Sider (2011).
[35] Suikkanen (2017), van Roojen (2006), Dunaway and McPherson (2016).

property as the eligible referent for a given concept and set of reference-fixing facts. In this way, reference magnets are, as Ted Sider points out,[36] a further constraint on reference fixing: Reference magnetism is not a theory of reference itself, but a doctrine that can be (and perhaps must be) coupled with any theory of reference you have independently motivated.

Just how do reference magnets function to take in multiple eligible referents and output one eligible referent? As Lewis explains:

> This constraint looks not to the speech and thought of those who refer, and not to their causal connections to the world, but rather to the referents themselves. Among the countless things and classes that there are, most are miscellaneous, gerrymandered, ill-demarcated. Only an elite minority are carved at the joints, so that their boundaries are established by objective sameness and difference in nature. Only these elite things and classes are eligible to serve as referents.[37]

Lewis then weakens his claim a bit by allowing for matters of degree of eligibility of reference.[38] Others have followed this line of thought. What is important for our purposes is that, of the multiple eligible referents determined for a given concept in a given theory of reference, the doctrine of reference magnetism then kicks in to fix the concept onto the *most* fundamental, joint-carving referent that is eligible.

In order for reference magnets to assist the non-naturalist, we first need a theory of reference that latches onto at least *some* eligible non-natural properties. If the theory of reference designates no eligible referents, then reference magnets have no set of eligible referents to pare down. It's crucial, then, that the theory of reference the non-naturalist is supplementing with reference magnets already targets non-natural properties as eligible referents. This rules out the causal and teleological theories, since these theories don't target any non-natural properties as eligible referents. A more plausible candidate theory here may be a conceptual role semantics, in which we take the cluster of inferential relations between the normative concepts and Ramsify over them.[39] On such a theory, the set—or some relevant subset—of the normative concepts is interdefined, and then reference magnets do the work to attach them onto the relevantly isomorphic joint-carving non-natural properties.

I concede that such a view could provide a positive metasemantics for the non-naturalist. However, it comes at a cost which, other things equal, we

[36] Sider (2011: section 3.2).
[37] Lewis (1984: 227). Lewis credits the idea to Merrill (1980).
[38] Lewis (1984: 227–8).
[39] This sort of proposal was suggested to me by Billy Dunaway, Caleb Perl, and Mark Schroeder. It is also gestured at in passing by Jussi Suikkanen (2017: 20).

should want to avoid. For notice that, in order for such a view to work, we have to hope (or better: argue) that two things hold. First, we must hope that all normative concepts can be interdefined in an asymmetric Ramsified set. Ramsification requires asymmetry. Lewis himself implicitly concedes this in his famous paper on Ramseyan Humility:

> We have assumed that a true and complete final theory implicitly defines its theoretical terms. That means that it must have a unique actual realization. Should we worry about symmetries, for instance the symmetry between positive and negative charge? No: even if positive and negative charge were exactly alike in their nomological roles, it would still be true that negative charge is found in the outlying parts of atoms hereabouts, and positive charge is found in the central parts.[40]

As Lewis here assumes, symmetries will be problematic within a Ramseyan sentence if they can't be eliminated. But the worry is that we can't be confident that the distribution of the normative properties *will* be asymmetric in the relevant sense, at least without some independent (and presumably contentious) argument.[41] Suppose, for example, that all normative concepts are reduced to GOODNESS and BADNESS. What would make it the case that GOODNESS refers to *goodness* rather than *badness*? Presumably, nothing more *could* do the work—the work that reference magnets could do has run out. And the argument for an asymmetry ought to be independent from the metasemantic considerations raised here—or anyway, it seems to me that allowing metasemantic considerations to rule out certain views about the structure and relationship between different normative concepts is getting the inquiry backwards.

A second problem for the CRS + reference magnets view is that it assumes an isomorphism between the distribution and relationships between the normative concepts and the distribution and relationships between the normative properties. Let's say we generate a Ramsey sentence with normative concepts A, B, C, . . . N. Start first with distribution. Now suppose that one of the concepts, C, is defective, in the sense that it fails to refer to any normative property. Because all of the other concepts are, directly or indirectly, defined partly in terms of C, this defectiveness bleeds into the rest of the conceptual system, and so it's unclear whether *any* of the concepts continue to refer. Next turn to the relationships between the concepts. Suppose it is part of the inter-definition of A and B that [if x is A, then x

[40] Lewis (2009: 207). Thanks to Ryan Doody and Daniel Wodak for discussion here.

[41] As Zoë Johnson King suggested to me, thick concepts may be able to help, because they can attach the valences to the relevant descriptive properties as a matter of conceptual truth. I think this can't work for reasons that Parfit (2011: section 90) points out, but I can't explore this issue here.

is not B]. Further, suppose that there are no two properties that have all of the other relations that A and B have but are also such that [if x is A, then x is not B]. A and B are thus not isomorphic with any normative properties. So they don't clearly have referents.

It's not clear how much of a problem this second problem is. Surely the conceptual role semanticist should allow for some flexibility and error in a conceptual system, compatible with its concepts having referents. But it is not a trivial question to ask how much is acceptable, and whether we can be confident that normative concepts rise above this bar. Perhaps this is where reference magnets are supposed to help. However, it isn't obvious that they can—for recall that the role of reference magnets is only to fix onto one reference from a set of eligible ones. So the conceptual role semanticist about non-natural properties must be sure that her theory allows for enough flexibility and error for our normative concepts to pick out the non-natural properties as eligible referents. This challenge may be surmountable, but it is not insignificant.

7.4. AN EPISTEMIC APPROACH TO METASEMANTICS

Non-naturalism runs into problems for each of the traditional metasemantic theories, as well as for the "connectedness" model. It seems that the non-naturalist metasemanticist is in serious trouble. Without a plausible explanation of how our moral thoughts could latch onto the non-natural, normative properties, non-naturalism faces a very serious metasemantic objection. I now defend a positive metasemantic theory for non-naturalism, closely related to Imogen Dickie's "justification-based" theory of reference fixing.[42] To make clear that, despite being influenced by Dickie's view, my view involves important nuances. I will call the view below the *epistemic theory of content*.

To motivate the epistemic theory, notice two central motivations for externalist theories. First, they capture the intuitive thought that, if we are able to think about some particular object or property, then we must bear some kind of special relationship to the object/property in question. This is important to avoid underdetermination problems. If I'm thinking about the gray mug in my cabinet, there must be some relationship between me and *that mug* that makes my thought fix on that mug rather than any of the thousands of qualitatively identical mugs in the world. Second, the

[42] Dickie (2015, 2016).

connection that externalists posit as required for reference fixing explains our intuitions about twin-earth cases. What explains the fact that Oscar's concept WATER refers to H2O, whereas Twin-Oscar's refers to XYZ? Externalist theories can provide this explanation easily.

As we saw above, it is these external-relational requirements that spell trouble for non-naturalism. Standard non-naturalists reject the claim that normative properties are causally efficacious, as well as the claim that representing non-natural properties fulfills some evolutionary or individualistic teleological function. But surely non-naturalists shouldn't deny that our moral thoughts bear *some* relationship to the non-natural normative properties. Even setting aside the rough motivations for some kind of externalist theory of content, non-naturalists should endorse the claim that there is some relationship between our moral thoughts and the non-natural normative properties in order to undergird justified moral beliefs or moral knowledge. This relation could come in any number of forms, depending on one's favored non-naturalist moral epistemology.[43]

These thoughts naturally suggest some kind of epistemic theory of reference fixing. The rough idea is that a concept *C* refers to some property *F* just in case the mode of justification for beliefs containing *C* non-luckily converges on *F*. Such a theory, if it could be made to work, would capture the intuitive motivations for externalist theories, but in a way that makes reference to non-causal properties such as non-natural normative properties in principle possible. This rough idea suggests two questions. First, can such a theory be made to work? Second, given the difficulties faced for non-naturalist moral epistemologists, is this really an improvement for the non-naturalist metasemanticist? I address each question in turn.

Helpfully, a book-length defense of an affirmative answer to the first question has been given by Imogen Dickie.[44] I cannot hope to recap the entire argument of the book here. But, briefly, Dickie motivates her core idea on the basis of two premises, which she calls *Principle connecting aboutness and truth* and *Principle connecting truth and justification*:

> *Aboutness and Truth:* "A thought about an object (a thought attributing a property to an object) is true iff the object has the property."[45]
>
> *Truth and Justification:* "Justification is truth-conducive: in general, and allowing exceptions, if a subject's belief is justified, he or she will be unlucky if the belief is not true and not merely lucky if it is."[46]

[43] Bengson (2015), Cuneo and Shafer-Landau (2014) provide two such examples.
[44] Dickie (2015). [45] Dickie (2015: 37).
[46] Dickie (2015: 38). This is Dickie's approximate formulation of the principle, but the precise details are outside the scope of this chapter.

Aboutness and Truth should be relatively uncontroversial. *Truth and Justification* is, perhaps, slightly more controversial, but not much. Even if justification doesn't *aim* at truth,[47] it is overwhelmingly plausible that, when things are going well, justified beliefs are, in virtue of their justificatory status, more likely to be true. I'll assume that something like *Truth and Justification* is correct in what follows.[48]

These two relatively mundane principles entail, for reasons explained in an appendix, a surprisingly interesting metasemantic thesis:

> *Aboutness and Justification:* S's <a is ϕ> beliefs are about an object iff their means of justification converges on the object, so that, given how the beliefs are justified, the subject will be unlucky if they do not match the object and not merely lucky if they do.[49]

First, a word about the notion of a "means of justification." For Dickie, each concept has a "proprietary means of justification."[50] To see what this comes to, imagine a case where there is conflict between two potential sources of justification. You're looking in the fridge and you see that there is a package of tofu sitting on the shelf. You form the belief *we have tofu in the house*. Your roommate then shouts from the other room "We are out of tofu!" You now have testimonial evidence that conflicts with your perceptual evidence.[51] For Dickie, a proprietary means of justification is the means of justification that you take—other things being equal—to trump in cases of conflict. In a case where you are directly perceptually linked up with some object, you will tend to take that to trump your testimonial evidence. But in other cases, such as cases in which an object is very far away and you know that your interlocutor is knowledgeable, testimonial evidence will trump perceptual evidence. Whichever piece of evidence tends to trump for some concept *C* will be *C*'s proprietary means of justification. And because of its potential relevance to the moral case, notice that a means of justification being proprietary does not entail that all or even that the majority of our beliefs containing the concept make use that means of justification.

Here I want to briefly flag a worry about the notion of a proprietary means of justification. It's unclear to me, *contra* Dickie, that there are any

[47] Contra Cruz and Pollock (2004).
[48] Dickie (2015: section 2.1) has a much more sophisticated defense of this principle.
[49] Dickie (2015: 37). Again, this is Dickie's approximate version of the principle.
[50] Dickie (2015: 50–7).
[51] Notice too that beliefs using the concept in question will still have the same reference even when they are formed using some non-proprietary means of justification. The proprietary means fix the reference of a concept, but the concept can be constituents of all sorts of beliefs that are not proprietary. Thanks to an anonymous referee for pointing out the ambiguity here.

such means of justification that will always serve the role of a proprietary means of justification. That is, it seems that, with respect to the very same concept, there will be contexts in which perceptual evidence will trump, and contexts in which testimonial evidence will trump. So what are we to say about such a case?

There are a number of options here. On the one I prefer, proprietary means are fixed by what the fundamental mechanism of justification for beliefs about *C* is. This means that, if testimonial evidence ultimately traces back to the perceptual evidence of someone else, perception retains its status as the proprietary means of justification.[52] But rather than get bogged down exploring this and the other options, I think it is worth noting that this is arguably less of a worry for the moral case. Insofar as there is some source of justification for moral beliefs (intuitions, rational insight, reflective equilibrium, etc.), it seems clear that this source will trump other sources, such as testimony.[53] While I grant that this is a contentious claim, it would take quite some time to defend. So instead, let me just grant that, if this idea of a proprietary means of justification cannot be defended or revised, the epistemic theory of content is in some trouble.

Return to *Aboutness and Justification*. This principle is not quite what the non-naturalist metasemanticist needs, for two reasons. First, as stated, the principle is about fixing reference for objects, not properties. But altering the principle into one about properties is straightforward:

> S's <a is ϕ> beliefs are about a property iff the (proprietary) means of justification converge on the property so that, given how the beliefs are justified, the subject will be unlucky if they don't track the property and not merely lucky if they do.

The second problem is more substantive. *Aboutness and Justification* provides a biconditional relationship between a means of justification and the referent of beliefs that the means of justification generates. It doesn't yet give us a complete metasemantic theory, since it doesn't tell us whether it is the means of justification that *determines* the referent or vice versa. And in fact, Dickie argues that there is no priority here.[54]

If the non-naturalist metasemanticist is going to use *Aboutness and Justification* (or something like it) to make theoretical progress, it must be that a particular means of justification fixes reference. Dickie rejects this

[52] So e.g. this would entail that the proprietary means of justification for any mathematical belief is going to be a priori, whether or not I am disposed to defer to professional mathematicians' testimony over my own a priori reasoning.

[53] As skepticism about forming beliefs based on moral testimony may help to show.

[54] Dickie (2015: 3.5).

position, on the grounds that "[o]ur grip of the kind of factor that justifies beliefs seems to rest on our grip of the kind of factor that—in *most,* or *nearby,* or, *optimal* circumstances—will result in formation of beliefs that are true."[55] In other words, it appears that our conception of justification will ineliminably refer to modes of belief-formation that tend toward truth, in which case we need a grip on the truth conditions, in which case we need to know which objects and properties figure in the beliefs being formed before we can determine the conditions on justification. Using a mode of justification to fix reference won't work, then: it helps itself to a concept's referent, the very thing it is attempting to explain.

This argument must fail: It proves too much. It generalizes far beyond an epistemic theory of content determination. To see this, consider first a *prima-facie* problem for the causal theory of reference. Any plausible theory of reference better entail that your concept DOG refers to dogs. But, of course, there are a number of circumstances in which your DOG concept will be tokened by non-dog things: large cats, small horses in the distance, stuffed dogs, and so on will often token your DOG concept. So the causal theorist is going to have to say something about why DOG doesn't pick out *dog-or-largecat-or-smallhorse-or-stuffeddog*, but just *dogs*.[56] Causal theorists have had a number of things to say about how to do this.[57] But what is most important is that these ways do not, and need not, refer to the fact that DOG refers to *dogs*. So, at least in principle, there is no circularity in solving this problem. For similar reasons, we shouldn't—at least without further argument—assume that a story about a mode of justification must assume the content of the justified beliefs that it feeds out.

It is true that the path here is harder to hoe, because justification is itself a normative notion.[58] It will be hard to determine success conditions without knowing what facts the means of justification attempt to track accurately. But there is no reason to think this can't be done. Notice that the situation is structurally similar to the causal theorist's. The causal theorist must give an account to separate the causal conditions that determine content from those that do not. And this account must not help itself to the content that is being fixed. Similarly, the epistemic theorist must give an account to separate the beliefs formed on the basis of the mode of justification in

[55] Dickie (2015: 111).

[56] This is a problem originally flagged by Fodor (1984).

[57] See Adams and Aizawa (2017: section 3) for an overview.

[58] However, note that there is no circularity in having a normative notion in the metasemantic story. Unlike in the internalist case, we aren't providing subjects with the justification-facts in order to fix their reference for a concept like JUSTIFICATION. (Thanks to Aaron Elliott for discussion here.)

question that determine content from those that do not. She needs a theory of which beliefs are epistemically unlucky, in the sense that they are justified beliefs, but not ones that fix content. Once we set aside the epistemically unlucky beliefs, we can determine the content of the belief in question by seeing which object/property the non-lucky beliefs converge on.[59] How can the epistemic theorist distinguish the reference-fixing beliefs from those that are merely unlucky? After all, it may seem that the most natural ways of doing this will involve appealing to the content of the beliefs in question, which the epistemic theorist can't help herself to on pain of circularity.[60]

While a full solution of this problem lies outside the scope of this chapter, let me gesture at one that seems promising. (If this solution is ultimately untenable, that's ok—what is important is that this problem mirrors a problem for any externalist metasemantics. So even a lack of a solution is not in itself a reason to reject the epistemic theory proposed here.) Take some proprietary means of justification, such as visual experience. Such a mode plausibly has (or would on a fully worked out theory) a canonical list of good conditions for justification-conferring uses of the mode—for example, good lighting, being awake, sober, and so on. Notice that, if we are careful, nothing on this list will itself appeal to the content of what is seen. When and only when a visual experience meets the conditions on the list, it will count as a reference-fixing instance for the concepts being deployed in the downstream belief(s). Now this won't be *quite* enough, because even under good conditions, visual experiences can get things wrong. So there will remain some *prima-facie* reference-fixing beliefs that are nonetheless intuitively false. This requires, perhaps, some further modal condition: The token belief will count as reference-fixing just in case (a) it was formed in good conditions for its proprietary means of justification, and (b) it isn't the result of an "epistemically" deviant chain.[61] A similar story could be told for any proprietary means of justification.

Let's walk through how our normative concepts could fix onto non-natural properties, according to the epistemic theory of content. Imagine a toy non-naturalist epistemology, rational insight. (It won't matter how this gets fleshed out, so the reader should imagine this mechanism however she so desires.) On such a view, agents form their (pure) moral beliefs on the

[59] If they converge on nothing, then the concept is defective in some way. What to say about defective concepts is a vexed matter that I can't address here. But notice that the possibility of a concept's being defective is a positive feature of the epistemic theory. Surely any theory of reference fixing should allow for and explain the possibility of defective concepts.

[60] I thank Joshua Schechter for pressing me to say more here.

[61] I take it that whatever gets said about what makes a causal chain deviant could also be said about epistemically deviant chains (some of which are surely causal).

basis of rational insight. When things go well, and the canonical conditions for this means of justification are met, rational insight converges on the facts about normative properties. An agent can't—at least non-luckily—form justified beliefs on the basis of other means of justification, such as perception or testimony. These are not sources of fundamental non-natural information, and so they will not count as the reference-fixing modes of justification. Only the outputs of rational insight count as fixing the content, and so the non-naturalist who holds such a view doesn't need to worry that beliefs will converge on some natural property.[62] There is also no risk that the means of justification will converge on the wrong normative property: Part of what makes it the case that WRONGNESS targets *wrongness,* as opposed to, say, *goodness,* is that the means of justification make beliefs containing WRONGNESS non-luckily true, which can only be explained by WRONGNESS referring to *wrongness*.[63]

7.5. HOW THE EPISTEMIC THEORY HELPS THE NON-NATURALIST

My proposal is that the non-naturalist metasemanticist endorse *Aboutness and Justification*, or something like it, read as a theory about content fixing for (at least) our moral concepts. I think there are three important reasons why the non-naturalist should find the epistemic theory of content attractive. First, and most importantly, the epistemic theory explains how reference to non-natural normative properties is possible. This alone makes the epistemic theory better suited for non-naturalist metasemantics than the other views discussed above. Second, the epistemic theory helps to unify two independent objections to non-naturalism—on the one hand, that non-naturalists cannot explain how beliefs in non-natural properties are possible, and on the other, how non-naturalist moral knowledge is possible. The epistemic theory of content, if correct, would entail that a wholly adequate non-naturalist epistemology would simultaneously rebut both metasemantic and epistemological concerns for non-naturalism. And finally, the

[62] As Gunnar Björnsson points out, it *could* turn out that the normative concepts refer to natural properties on such a view, just in case there is some equally or more fundamental natural property that rational insight converges on. But that seems unlikely, at least on a standard non-naturalist view according to which the normative properties are fundamental (or close to it).

[63] Importantly, the theory of epistemic luck we endorse can't itself make indispensable use of the referent in question, on pain of circularity. Thanks to Aaron Elliott and Bar Luzon for discussion here.

epistemic theory is ideally placed to capture Autonomy and "just too different" intuitions, the central motivations for non-naturalism in the first place. I'll briefly discuss each of these advantages in turn.

7.5.1. Non-Naturalism and the Epistemic Theory of Content

Virtually no non-naturalists are skeptics. So they're committed to:

> *Moral Knowledge*. Human beings have some non-accidentally true justified moral beliefs.

Moral Knowledge entails that there is a source of justification for at least some of our moral beliefs. There is some belief-forming method we have—in Dickie's phrase, a "means of justification"—that, if we're not unlucky, converges on truths about non-natural moral properties. Just what is this means of justification?

Dickie, who is concerned with means of justification for ordinary objects, discusses two potential means of justification: perception and testimony.[64] It's possible for the non-naturalist to argue that these are the means of justification for non-natural properties as well, but this has not been the traditional approach of non-naturalists. What's important is that the anti-skeptical non-naturalist is committed to *some* proprietary means of justification. If their theory is plausible, it should turn out that in good cases the means of justification results in non-luckily true beliefs, and thus that the means of justification will, over time, converge on the non-natural properties.

The fact that means of justification are only constrained by their tendency toward truth is central to avoiding the problems of the previous externalist metasemantic theories. Causal and teleological metasemantic theories are ill-placed for the non-naturalist metasemanticist precisely because they impose constraints on content-fixing that non-naturalist properties can't meet. Even if we have direct epistemic access to certain properties or objects, on such views, so long as we aren't causally or teleologically related to these properties, we can't form beliefs about them. An epistemic theory of content can capture what's good about these theories while remaining agnostic about the possibility of referring to other properties. Maybe we can't non-accidentally track non-causal properties. But if we can, only the epistemic theory of content can move us from this non-accidental tracking to content fixing. This is just what the non-naturalist needs. If the non-naturalist has an otherwise compelling theory of intuition,

[64] Dickie (2015).

rational insight, or moral perception, then they have a mode of justification that the epistemic theory of content can exploit to provide a metasemantics for the non-natural properties.

7.5.2. Metasemantic and Epistemic Objections to Non-Naturalism: One and the Same?

There may seem to be an obvious but deep problem with what's been just said. The epistemic theory of content is all well and good for domains where we have a clear epistemology. But the moral domain is not one in which we have a clear picture of how knowledge is possible. And the situation is especially dire for non-naturalists, given Benacerraf-influenced arguments and evolutionary debunking arguments, which try to show that non-naturalism makes moral knowledge impossible. Isn't appealing to the epistemic theory of content trying to solve a difficult problem by reference to an epistemological theory with an even more devastating problem?

I admit, there's something ironic here in appealing to non-naturalist epistemology to solve a problem for non-naturalism, given that epistemology is one of the biggest sources of concern for such a view. However, as noted above, the non-naturalist is already committed to there being *some* solution to these problems. The epistemic theory of content only aims to piggyback on whatever solution this might be, and use it to fix content onto the non-natural properties. A wholly adequate non-naturalist epistemology will already be committed to there being some means of justification that non-luckily converges on the non-natural normative facts.[65] The epistemic theory of content says that this is all that is needed to ensure that our moral beliefs pick out the non-natural properties. So it follows from commitments that the non-naturalist already has—controversial as they might be—that the epistemic theory of content can provide a proper metasemantics for non-natural normative beliefs.

Of course, none of this goes any of the way toward actually providing a theory of justification for non-naturalism. Such a project is non-trivial, but it is a task for the non-naturalist qua epistemologist, not qua metasemanticist. And this is tentatively good news for the non-naturalist, for two related reasons. First, it reduces two families of objections to non-naturalism—epistemic and metasemantic—to one. This means that, if it can be done, a positive epistemological theory for non-naturalism could rebut two potentially devastating sets of objections to non-naturalism in one fell

[65] This mode of justification also cannot presuppose that the metasemantic explanans is already met. Thanks to Bar Luzon for pointing this out to me.

swoop. Second, this is contingently good news for the non-naturalist given the fact that much more work has been done on addressing non-naturalist epistemology than on non-naturalist metasemantics. So the reduction of the two families of objections into one has the fortuitous upshot of rendering the success of non-naturalist metasemantics dependent on a problem that has already received an overwhelming amount of attention and work. Obviously, no consensus has emerged, even among those sympathetic to non-naturalism, about how to solve these epistemological problems. But there is hope.

7.5.3. Autonomy and the "Just Too Different" Intuition

Finally, the epistemic theory of content fits perfectly with the autonomy claims, core motivations for non-naturalism. Recall *Epistemological* and *Metaphysical Autonomy:*

> *Epistemological Autonomy*. An agent could be wholly procedurally rational and fully informed about all the natural facts and yet ignorant or mistaken about the normative.
>
> *Metaphysical Autonomy*. The normative facts are irreducible to the natural facts, in the sense that there is no conceptual entailment or complete metaphysical explanation from the natural facts to the normative facts.

We saw above that these claims cause problems for certain attempts to provide a non-naturalist metasemantics. Non-naturalists claim that the metaphysical gap, sometimes expressed in terms of the "just too different" intuition, between natural facts and normative facts, is too wide to be crossed epistemologically. An epistemology for the purely normative facts won't be assisted by gathering more non-normative facts. Our means of justification (in terms of the source of input) for normative facts is going to have to be fundamentally different from our means of justification for run of the mill natural facts, given their metaphysical status as independent from them.

The epistemic theory of content nicely accommodates this line of thinking. To see why this is so, just recall from above that, for the non-naturalist, the proprietary means of justification for normative beliefs will be unlike the means of justification for their natural counterparts.[66] This fits perfectly

[66] Not all non-naturalists accept this claim (e.g. Seitya 2012), myself included, ironically (Werner 2016, 2018). But most do, and I think even those who do not accept that the sentence in the text is strictly true have reason to accept a variant of it so long as they're committed to Metaphysical Autonomy.

with the line of thinking behind the Autonomy Theses. The Autonomy Theses are preserved, and thus the core motivation for non-naturalism is preserved—even explained in part—by the epistemic theory of content. A similar explanation will hold for any non-naturalist theory that aims to capture Epistemological Autonomy.

7.6. CONCLUSION

Non-naturalists have done a great amount of work on the metaphysics and epistemology of irreducibly normative properties. Considerably less work has been done on their metasemantics. This is surprising, because many of the traditional metasemantic views rule out the possibility of referring to and having beliefs about non-natural properties. Thus an underexplored objection to non-naturalism remains unsolved. Non-naturalists may have believed that they could help themselves to other realist-friendly metasemantics for normative terms. The first goal of this chapter was to argue that that is mistaken: None of the traditional metasemantic theories, even those explicitly given to be realist-friendly, fit with non-naturalism, especially given considerations surrounding the Autonomy Theses. Thus, the pessimistic half of this chapter argued that non-naturalists really do face a metasemantic challenge.

My second goal in this chapter defended a sketch of a positive metasemantic view, indebted to recent work by Imogen Dickie. On this view, what makes a given normative concept refer to a non-natural property is that its means of justification converge onto the facts that the property figures in. The view has powerful independent motivation. Furthermore, it avoids the problems that other metasemantic theories cause for non-naturalists, as well as according nicely with some of the central motivations for non-naturalism.

The success of this theory of content at explaining how our normative beliefs pick out the non-natural properties depends on providing an adequate epistemology for non-naturalism. Since this task is notoriously difficult, non-naturalists are not out of the woods. If the epistemic theory defended above is correct, then, defending a positive epistemology for non-naturalism is even more urgent. If it can be done, the non-naturalists will have made a significant amount of progress, not just epistemologically, but metasemantically as well.[67]

[67] Thanks to an anonymous referee for pressing me to unpack the reasoning in the following Appendix.

APPENDIX

Aboutness and Justification: Dickie's Arguments

In the main body of the chapter, I claimed that

> *Aboutness and Justification:* S's <a is ϕ> beliefs are about an object *o* iff their means of justification converges on the object, so that, given how the beliefs are justified, the subject will be unlucky if they do not match the object and not merely lucky if they do.

is entailed by the following two principles:

> *Aboutness and Truth:* "A thought about an object (a thought attributing a property to an object) is true iff the object has the property."[68]
>
> *Truth and Justification:* "Justification is truth-conducive: in general, and allowing exceptions, if a subject's belief is justified, he or she will be unlucky if the belief is not true and not merely lucky if it is."[69]

But the proof here is not at all obvious. I here briefly walk through the proof of the biconditional; my discussion in this appendix is heavily indebted to Dickie's own discussion.[70]

Begin with the left-to-right conditional. The proof here is straightforward. Suppose

> 1. S's belief that <a is ϕ> is about *o*.

From 1 and Aboutness and Truth, we get:

> 2. S's belief that <a is ϕ> is true iff *o* is ϕ.

From 2 and *Truth and Justification*, it follows that:

> 3. Justification that renders S's belief that <a is ϕ> unlucky if false and not merely lucky if true will make it the case that S's belief is unlucky if *o* is not ϕ.

Which thereby gives us the left-to-right conditional:

> 4. If S's <a is ϕ> belief is about *o*, justification that renders S's belief that <a is ϕ> unlucky if false and not merely lucky if true will make it the case that S's belief is unlucky if *o* is not ϕ.

[68] Dickie (2015: 37).

[69] Dickie (2015: 38). This is Dickie's approximate formulation of the principle, but the precise details are outside the scope of this chapter.

[70] Dickie (2015: ch. 2).

Let's turn, then, to the right-to-left conditional. Here things get trickier. Here is the conditional to be proven:

> *RtL Aboutness and Justification:* If the proprietary means of justification for S's <a is ϕ> beliefs converge on some object *o*, such that S will be unlucky if they do not match *o* and not merely lucky if they do, then S's <a is ϕ> beliefs are about *o*.

One reason the proof of this side of the conditional is trickier is that we need to rule out two alternatives to establish that S's <a is ϕ> beliefs are about *o*. First, it may be that S's beliefs are about some other object, *o**. Second, it may be that S's beliefs are about nothing at all—they may have no referent. Let's consider each in turn.

How can we rule out that S's beliefs are about some other object, *o**? In order for this to be so, we would need two objects, *o* and *o**, such that all of their intrinsic and relational properties—or at least all of them accessible via the proprietary means of justification—are the same. In the case of ordinary objects, at least, this simply won't happen, because some of the relational properties of *o* and *o** have to do with their relations to the believing agent herself. And barring spatiotemporally overlapping intrinsic duplicates, this won't happen.[71] Now this reasoning only works for ordinary, physical objects, whereas in the chapter, I am concerned with non-natural properties. So a question can be raised about why one's normative beliefs may not be about some other *property*, *F**, such that *F* (the genuinely normative property) is not identical to *F**. But in the case of properties, I submit that it is just impossible for there to be two distinct properties which share all of the same features, and so the problem just doesn't arise. It is true that there may be some natural property *F** which is extensionally equivalent to *F*, but *F* would still, assuming non-naturalists are right, have some second-order features that *F** does not.

The second possibility is that, while S's proprietary means of justification converges on *o*, her beliefs are nonetheless about *nothing*. She has failed to secure reference because, presumably, there is some other condition on fixing reference that she has not met. This alternative is less plausible on its face. We would need some powerful argument to the effect that, even though S's beliefs *consistently* and *non-coincidentally* track facts about *o*, nonetheless her beliefs are not about *o*. And it is hard to imagine how such an argument would go.

I wholly realize that the above arguments leave some space for disagreement and rebuttal. But I hope to have done a good job of motivating the view without getting too far afield from the present project, which is to assume that Dickie's theory is, broadly speaking, correct, and to show how it can be extended to non-naturalist metasemantics.[72]

[71] See Dickie (2015: ch. 2) for detailed reasoning along these lines.

[72] For extremely helpful feedback and discussion on earlier drafts, I'm thankful to John Bengson, Gunnar Björnsson, Teresa Bruno Niño, Janice Dowell, Billy Dunaway, Kevan Edwards, Aaron Elliott, David Enoch, Nikki Fortier, Zoë Johnson King, Avi Kenan, W. Scott Looney, Bar Luzon, Hille Paakkunainen, Caleb Perl, Jared Riggs, Mark Schroeder, and Byron Simmons.

References

Adams, Fred, and Ken Aizawa (2017). "Causal Theories of Mental Content" in *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/content-causal/

Benacerraf, Paul (1973). "Mathematical Truth," *Journal of Philosophy* 70(19): 661–79.

Bengson, John (2015). "Grasping the Third Realm," *Oxford Studies in Epistemology* 5: 1–38.

Blackburn, Simon (1984). "Supervenience Revisited" in Ian Hacking (ed.), *Exercises in Analysis: Essays by Students of Casimir Lewy*. Cambridge: Cambridge University Press.

Boyd, Richard (1988). "How to Be a Moral Realist" in G. Sayre-McCord (ed.), *Essays on Moral Realism*. Ithaca, NY: Cornell University Press.

Boyd, Richard (2003). "Finite Beings, Finite Goods: The Semantics, Metaphysics and Ethics of Naturalist Consequentialism, Part 1," *Philosophy and Phenomenological Research* 66(3): 505–53.

Chalmers, David (2012). *Constructing the World*. Oxford: Oxford University Press.

Cheyne, Colin (2001). *Knowledge, Cause, and Abstract Objects*. Dordrecht: Kluwer Academic Publishers.

Clarke-Doane, Justin (2017). "What is the Benacerraf Problem?" in Fabrice Pataut (ed.), *New Perspectives on the Philosophy of Paul Benacerraf: Truth*, Objects, Infinity. Cham: Springer.

Cruz, Joe, and JohnPollock (2004). "The Chimerical Appeal of Epistemic Externalism" in Richard Schantz (ed.), *The Externalist Challenge*. Berlin: De Gruyter.

Cuneo, Terence, and Russ Shafer-Landau (2014). "The Moral Fixed Points: New Directions for Moral Nonnaturalism," *Philosophical Studies* 171(3): 399–443.

Dickie, Imogen (2015). *Fixing Reference*. Oxford: Oxford University Press.

Dickie, Imogen (2016). "The Essential Connection Between Epistemology and the Theory of Reference," *Philosophical Issues* 26(1): 99–129.

Dunaway, Billy, and Tristram McPherson (2016). "Reference Magnetism as a Solution to the Moral Twin Earth Problem," *Ergo: An Open Access Journal of Philosophy* 3.

Elliott, Aaron (2014). "Can Moral Principles Explain Supervenience?" *Res Philosophica* 91(4): 629–59.

Enoch, David (2011). *Taking Morality Seriously*. Oxford: Oxford University Press.

Field, Hartry (1989). *Realism, Mathematics and Modality*. Oxford: Blackwell.

Fodor, Jerry A. (1984). "Semantics, Wisconsin Style," *Synthese* 59(3): 231–50.

Harman, Gilbert (1977). *The Nature of Morality: An Introduction to Ethics*. Oxford: Oxford University Press.

Heathwood, Chris (2015). "Irreducibly Normative Properties," *Oxford Studies in Metaethics* 10: 216–44.

Hume, David (1975). *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*. Oxford: Oxford University Press.

Jackson, Frank, and Philip Pettit (1995). "Moral Functionalism and Moral Motivation," *Philosophical Quarterly* 45(178): 20–40.
Joyce, Richard (2001). "Moral Realism and Teleosemantics," *Biology and Philosophy* 16(5): 723–31.
Joyce, Richard (2006). *The Evolution of Morality*. Cambridge, MA: MIT Press.
Lenman, James (2010). "Uggles and Muggles: Wedgwood on Normative Thought and Justification," *Philosophical Studies* 151(3): 469–77.
Lewis, David (1984). "Putnam's Paradox," *Australasian Journal of Philosophy* 62(3): 221–36.
Lewis, David (2009). "Ramseyan Humility" in David Braddon-Mitchell and Robert Nola (eds), *Conceptual Analysis and Philosophical Naturalism*. Cambridge, MA: MIT Press.
Liggins, David (2010). "Epistemological Objections to Platonism," *Philosophy Compass* 5(1): 67–77.
McConnell, Neil (2015). "The Deviance in Deviant Causal Chains," *Thought: A Journal of Philosophy* 4(2): 162–70.
McGrath, Sarah (2014). "Relax? Don't Do it! Why Moral Realism won't Come Cheap," Oxford Studies in Metaethics 9: 186–214.
McHugh, Conor, and Jonathan Way (2016). "Fittingness First," *Ethics* 126(3): 575–606.
McPherson, Tristram (2012). "Ethical Non-Naturalism and the Metaphysics of Supervenience," *Oxford Studies in Metaethics* 7: 205–34.
Merrill, G. H. (1980). "The Model-Theoretic Argument Against Realism," *Philosophy of Science* 47(1): 69–81.
Millikan, Ruth (1984). *Language, Thought and Other Biological Categories*. Cambridge, MA: MIT Press.
Millikan, Ruth (1995). "Pushmi-Pullyu Representations," *Philosophical Perspectives* 9: 185–200.
Moore, G. E. (1903). *Principia Ethica*. Mineola, NY: Dover.
Neander, Karen (2012). "Teleological Theories of Mental Content," in Stanford Encyclopedia of Philosophy. <https://plato.stanford.edu/entries/content-teleological>
Oddie, Graham (2005). *Value, Reality, and Desire*. Oxford: Clarendon Press.
Parfit, Derek (2011). *On What Matters*, vols 1 and 2. Oxford: Oxford University Press.
Ridge, Michael (2014). "Moral Non-Naturalism," in Stanford Encyclopedia of Philosophy. <https://plato.stanford.edu/entries/moral-non-naturalism>
Ruse, Michael, and Edward O. Wilson (1986). "Moral Philosophy as Applied Science," *Philosophy* 61(236): 173–92.
Scanlon, Thomas (2014). *Being Realistic about Reasons*. Oxford: Oxford University Press.
Schroeter, Laura, and Francois Schroeter (2003). "A Slim Semantics for Thin Moral Terms?" *Australasian Journal of Philosophy* 81(2): 191–207.
Schroeter, Laura, and Francois Schroeter (2014). "Normative Concepts: A Connectedness Model," *Philosophers' Imprint* 14(25): 1–26.

Schroeter, Laura, and Francois Schroeter (forthcoming). "The Generalized Integration Challenge in Metaethics," *Nous*.

Setiya, Kieran (2012). *Knowing Right from Wrong*. Oxford: Oxford University Press.

Sider, Theodore (2011). *Writing the Book of the World*. Oxford: Oxford University Press.

Sinclair, Neil (2012). "Metaethics, Teleosemantics and the Function of Moral Judgments," *Biology and Philosophy* 27(5): 639–62.

Street, Sharon (2006). "A Darwinian Dilemma for Realist Theories of Value," *Philosophical Studies* 127(1): 109–66.

Sturgeon, Nicholas (1998). "Moral Explanations" in James Rachels (ed.), *Ethical Theory 1: The Question of Objectivity*. Oxford: Oxford University Press.

Suikkanen, Jussi (2017). "Non-Naturalism and Reference," *Journal of Ethics and Social Philosophy* 11(2): 1–24.

Van Roojen, Mark (2006). "Knowing Enough to Disagree: A New Response to the Moral Twin Earth Argument," Oxford Studies in Metaethics 1.

Wedgwood, Ralph (2007). *The Nature of Normativity*. Oxford: Oxford University Press.

Werner, Preston J. (2016). "Moral Perception and the Contents of Experience," *Journal of Moral Philosophy* 13(3): 294–317.

Werner, Preston J. (2018). "Moral Perception without (Prior) Moral Knowledge," *Journal of Moral Philosophy* 15(2): 164–81.