CHAPTER 3

NATURALIZING BELIEF FOR NATURALIZED EPISTEMOLOGY

In my first chapter, I outlined a new taxonomy of naturalized epistemologies, divided according to the methodological concepts exploited by varieties of "conceptual projects" used by various naturalisms. I argued that "optimistic" naturalized epistemologists might naturalize knowledge by analyzing the *concept* "knowledge" in accordance with different semantic theories (e.g., analytic or two-factor), or by exploring the metaphysics of the *fact* of knowledge by way of metaphysical identity criteria (e.g., supervenience). I noted in passing, however, that the naturalistic acceptability of many of these methodological concepts could easily be called into question. In this chapter, I want to explore and defend that claim in greater detail. But instead of showing how these methodological concepts would be applied to the concept "knowledge," I want instead to survey their application to one of the concepts in terms of which "knowledge" is usually analyzed: the concept of "belief."

There are several reasons for using "belief" as the case study for the naturalistic acceptability of these methodological concepts. First of all, it is not at all uncommon for one part of the *analysans* of "knowledge" to play proxy in debates over the naturalizability of knowledge. More typically, epistemologists will focus on the naturalizabiltiy of "justification," particularly its normative element. In my first chapter, I surveyed some of the debate concerning the naturalizability of the normative, and concluded that naturalists had a stronger case for naturalizing the normative than is usually conceded—provided that they are permitted the usual range of naturalization methodologies we are currently calling into question. So one reason to explore other elements of the "justified true belief" complex is that "justification" has already been examined in some detail. It had been examined because normativity was thought to be a most distinctively non-naturalistic property, and if naturalism could not countenance it, surely epistemology could not countenance naturalism. Another reason to examine "belief" is that it is directly connected to an even greater collection of properties thought to be distinctively non-naturalistic: both inten*t*ionality and inten*s*ionality (as well as a sizeable normative

element of its own). So "belief" is as much of a challenge to naturalizing knowledge as "justification," if not more so.

There are, of course, probably even more proposals for "naturalizing" belief and intentionality than there are for naturalizing normativity. It is, after all, the subject matter of great swaths of the philosophy of mind. The varieties of naturalization proposals for belief basically parallel those we examined for naturalizing knowledge, but in this case questions about the subject matter to be naturalized are in many cases *the same* as the questions about the methodological concepts that will do the naturalizing. Notice, for example, how many of the naturalization proposals we have examined revolve around different theories of *reference* (because we are concerned with the reference of the concept "knowledge"). But theories of reference are of interest to naturalizing belief not only because we want to know more about the reference of the *concept* "belief," but because the *fact* of belief itself seems to involve an intimate connection to the fact of reference: beliefs are thought to be individuated by their content, and *content* is often thought to be a function of reference (among other things). As we shall see at the end of the chapter, even if the more consistent naturalization proposals for "belief" can dispense with objections concerning putatively non-naturalistic methodological concepts, questions will linger concerning their treatment of their subject matter, particularly insofar as theory of reference is needed for understanding the content of scientific beliefs, which are of special concern to naturalized *epistemology*. This is the second reason it is useful to examine the "belief" component of the traditional analysis of "knowledge."

So if it should turn out that we cannot naturalize a concept of "belief" usable by the naturalized epistemologist, this inability may count as an objection to naturalized epistemology itself. Insofar as the naturalizability of belief is tied up with naturalizing *any* phenomenon by way of a "conceptual project," understanding why we cannot naturalize belief in this manner may also end up reflecting back on questions about the naturalization of normativity and of justification. Understanding

this may, in other words, feature a more fundamental objection to naturalized epistemology than some of the traditional critiques.

In this chapter, therefore, I will examine naturalization proposals using a taxonomy similar to the one developed in my first chapter. This time, however, I will develop in greater detail the naturalistic objections to various methodological concepts. I will first examine "analytic naturalism" about belief, and develop the objection that naturalization of this variety is unacceptable to the naturalist because of its reliance on the method of *a priori* conceptual analysis (in spite of some recent defenses of this method against traditional Quinean objections). I will then consider "conceptually regulated scientific naturalism" which relies in part on conceptual analysis, and in part on other factors to determine reference. I will argue that this second proposal is non-naturalistic, not only because of its appeal to conceptual analysis, but also because of its reliance on intensional concepts at odds with basic naturalistic precepts. Finally, I will examine "conceptually indifferent scientific naturalism", a naturalization proposal that does away with conceptual analysis entirely. I take it that the final version is the most naturalistically palatable, and for this reason my polemics against *it* will be the most important in the chapter: I will argue that this final version, while generally free of non-naturalistic methodological concepts, fails to furnish us with a concept of "belief" *usable for epistemological purposes*. Having surveyed these alternatives, I will conclude that no obvious candidate for naturalizing belief is available for what I have labeled "optimistic" naturalized epistemology.[14] As a result, the only consistent form of naturalism possible seems to be of the pessimistic variety.

Before I come to any conclusions about the viability of *any* belief-naturalization proposals, I need to explain why it is that any naturalists feel it is important for their epistemological purposes to naturalize belief in the first place. This is, of course, a point that the pessimists will disavow, but it is

[14] In my first chapter, I counted "supervenience" naturalism in a separate "metaphysical" category apart from proposals based on theories of reference. Between that discussion and the more detailed discussion of Kim's views on supervenience in the second chapter, it should be clear that supervenience is of little use for naturalization purposes on its own, that it usually must be conjoined with reference-theoretic concerns. For this reason we will not treat it separately this time, but in conjunction with conceptually-regulated scientific naturalism, a spot where Kim himself has most recently placed it (Kim 2005).

important to be precise about *what* they are disavowing. In the first section of this chapter, therefore, I will present the claims by "optimistic" naturalists such as Kim, Goldman, Kitcher and Kornblith regarding the need to naturalize belief

Why naturalized epistemology needs naturalized beliefs

Do naturalized epistemologists recognize that their project depends on a naturalistic account of belief? Consulting the relevant literature reveals that they do.

Writing in the critique of Quine that we have now examined in some detail, Jaegwon Kim argues that even if Quine's doctrine fails to support a concept of epistemic justification, it may share enough in common with traditional epistemology to warrant the title of "epistemology" if it shares a concern for the subject matter of the formation of beliefs. This is because, as Kim suggests (1988, 392), even a naturalized epistemology would need to:

> . . .identify, and individuate, the input and output of cognizers. The input, for Quine, consists of physical events . . . and the output is said to be a "theory" or "picture of the world"—that is, a set of "representations" of the cognizers environment. . . . In order to study the sensory input-cognitive output relations for the given cognizer, therefore, we must find out what "representations" he has formed as a result of the particular stimulations that have been applied to his sensory transducers. Setting aside the jargon, what we need to be able to do is to attribute *beliefs*, and other contentful intentional states, to the cognizer.

Kim raises the issue in order to argue that "belief" itself is an inherently normative concept, and that any difficulty naturalism has with normativity will translate into a difficulty with "belief." In my first chapter, I argued that naturalism's difficulty with normativity is not as obvious as might seem to some. Even if that is true, however, Kim's point is still significant: it suggests that if there is *some* difficulty naturalists have with "belief" (because of its normativity or some other concern), naturalists may have difficulty naturalizing epistemology. He does not, however, say much more here about why he thinks contentful intentional states are so important to epistemology.

Kim does have much to say, of course, about how beliefs might be naturalized. He has spent much time developing the notion of "supervenience," and argues in numerous places (especially Kim

2005) that intentional content must supervene on the physical, by way of a reductive explanation drawing on functional conceptual analysis. We will discuss his views of naturalization in detail under the second naturalization proposal below (conceptually-regulated scientific naturalism).

Other naturalized epistemologists say more than Kim about why naturalized epistemology needs to mention beliefs. One example is Alvin Goldman, whom we have already described as an "optimistic" naturalist of the analytic variety. Goldman (1986, 162) explains why his epistemology assumes "the existence of beliefs and other propositional attitudes":

> To say that something has content is to say that it has semantic properties: meaning, reference or truth-conditions, for example. Given my epistemological perspective, truth-conditions are especially important. Unless mental states have truth-conditions, there can be no true or false beliefs; *a fortiori* there can be no mental processes with epistemological properties that interest us, such as, power and reliability. My investigation of such properties of mental processes could not get started: it would be devoid of relevant subject matter.

Goldman goes on to say that he thinks any epistemology—naturalistic or otherwise—must find a place for mental content. If an epistemology wishes to take knowledge seriously, it must take belief seriously. Even a form of pragmatist epistemology unconcerned with knowledge would need to talk about agreement and disagreement, which would still presuppose semantic content (162–3). Goldman recognizes, of course, that Quine's naturalized epistemology or other "speech-act" epistemologies focus on explaining what people say, not on "internal ideas and beliefs" 163).[15] But since Goldman is an "optimistic" naturalist, he sets these latter naturalisms aside. Instead of considering epistemologies that do away with mental content, Goldman examines some of the popular naturalistic criticisms of mental content. I will examine a portion of his response here, as it serves as useful set-up for the naturalization proposals we are about to consider.

Goldman first considers the eliminativist arguments of Paul Churchland (1981; 1996), which stem from treating content attributions as expressed by a kind of theoretical folk psychology. If folk psychology has limited explanatory or predictive power, it may prove to be false, and the entities to

[15] We could easily think of Williams's deflationism as an example of a "speech-act" epistemology.

which it refers (presumably, defined by their theoretical role) would not exist. This seems particularly likely to Churchland, since he thinks the entities of folk psychology are irreducible to neuroscience—particularly entities constituted by some kind of "language of thought."

Goldman (1986, 164–7) responds to Churchland with three rejoinders. First, he wonders if folk psychology should be expected to be able to explain everything Churchland says it fails to explain (mental illness, creative imagination, intelligence differences, ball-catching abilities, or the functions of sleep). He also challenges Churchland's presupposition that contentful states would need to be composed of an internal mental language. I agree with both of Goldman's criticisms here, and will not pursue the issue further. Second, Goldman wonders whether folk psychology need be understood as a kind of theory, and mentions problems philosophers have had defining the functionalist program in philosophy of mind. In his later work, Goldman (1995) will of course articulate a positive alternative to the "theory-theory" approach to folk psychology: the so-called "simulation" approach. I will examine this proposal in chapter 4.[16] Finally, Goldman argues that even if folk psychology is theoretical, its ontology need not be rejected just because its predictive/explanatory power is not perfect. He concedes that we will probably never produce a "strong" reduction of folk psychological types to neuroscientific types, but suggests that other forms of reduction may be possible. In particular, he thinks a corrected version of folk psychology might reduce somehow to properties at higher levels of analysis, as other concepts in the special sciences might.

There is a tension between this response and his second point, of course, because usually philosophers who propose reductions to higher-level properties have *functional* properties in mind. Given his discontent with existing functional analyses of "belief," and also given the need to preserve folk psychological ontology in the face of its possible theoretical limitations, Goldman might appreciate our second naturalization proposal to be discussed below (conceptually regulated scientific

[16] It turns out that Goldman's version of simulation theory doesn't work to naturalize belief, even if it accomplishes other purposes (a point Goldman himself concedes). Other versions of simulation theory better suited for naturalizing belief end up to be incompatible with important scientific evidence, but we will settle this later.

naturalism), which distinguishes between reduction and "reductive explanation." It derives from a theory of reference designed to accommodate the fact that we often refer to more or different properties of a phenomenon than we can predict in advance.

Goldman then considers a second set of objections to mental content raised by Stephen Stich (1983). Rather than arguing that folk psychological properties are irreducible to neuroscientific properties, Stich focuses instead on irreducibility to cognitive science. He alleges that folk psychological attribution depends too much on the attributor's theory of the world, and risks illegitimately disenfranchising too many other subjects (who happen to have different theories) as genuine cognizers. (For example, "Mrs. T" who suffers from memory loss can tell us "McKinley was assassinated" but cannot tell us anything else about McKinley or assassination. We feel we cannot assign genuine content here if there is nothing else she can say.)

Goldman (1986, 167–9) responds to Stich by questioning whether we should expect sentences of English to capture the entire content of a subject's belief. He suggests that we might understand content as a function of ranges of possible worlds envisioned by the content-holder, and further mentions externalist theories of content (such as Burge (1979)) that would permit assigning content based on subjects' background beliefs, rather than simply their utterances. Now as it turns out, both the possible worlds and externalist views of content are given ample consideration in the second belief naturalization proposal below ("conceptually regulated scientific naturalism"). So the viability of Goldman's proposals will in effect be examined there.

Whether we consider his responses to Churchland or Stich, Goldman is clearly only leaving us promissory notes, entitling us to belief-naturalization once someone else's theory has been vindicated. Elsewhere (1986, 16), he writes that cognitive science may be able to do without a positive account of mental content. But his insistence on showing how such an account is at least *possible* (with the arguments above), shows that naturalized epistemology, unlike cognitive science, may not be able to sustain itself without this account. His reliabilist approach is clearly dependent on making sense of the

reliability of *belief*-formation processes. This suggests that his "optimistic" naturalized epistemology will default if his promissory note cannot be redeemed.

Another naturalized epistemologist who explicitly considers the question of the naturalizability of belief is Philip Kitcher (1992). Kitcher also considers objections to mental content from both Churchland and Stich, but focuses specifically on their significance for "the *current* practice of naturalistic epistemology and philosophy of science." He acknowledges that folk psychology may someday be called into question by advances in neuroscientific or cognitive psychological research, but that until that day comes, there seems to be little for the naturalized epistemologist to do except exploit the language of mental content rather than a previously unknown advanced scientific language:

> [I]n advance of developing this [scientific] language in sufficient detail to account for the sophisticated reasoning that appears to occur in human inquiry, there is no way of formulating naturalistic claims about cognitively optimal strategies. The very advantage on which eliminativists sometimes insist—to wit, the display of kinship between human beings and other cognizers—is also a bar to the adumbration of naturalistic epistemology along eliminativist lines. For the goal of naturalistic epistemology and philosophy of science is to understand and improve our most sophisticated performances, and about these eliminativists have presently very little to say.
>
> If we accepted the eliminativist indictment of traditional propositional approaches to cognition, then prospects for naturalism would be discouraging. In effect, we would be confronted with the choice between an inadequate idiom and one not yet developed (1992, 109).

Kitcher goes on to suggest that we may not need to make this choice, because the "eliminativist indictment" is often overstated. Clearly this is true, if Goldman's preceding objections are any indication. Kitcher mentions that while some cognitive research clearly shows that much scientific cognition is not propositional, this research still retains many of the "categories of traditional epistemology." Interestingly, one of the sources he cites is none other than Goldman's *Epistemology and Cognition,* which we have found not to contain any new insights on the subject. Kitcher (1992, 110) closes by suggesting that for the time being, we should use an approach that combines two options:

> (i) aim to develop the preferred rival idiom and defer projects of epistemic appraisal until they can be reformulated in these terms, and (ii) continue to use whatever

> resources from empirical studies of cognition can be used to formulate and address normative epistemological enterprises.

I take it that what Kitcher means by combining these approaches is to work on developing the "preferred idiom" *without* deferring "projects of epistemic appraisal," for otherwise one could not also continue to use existing resources to address the normative questions he mentions.

Like Goldman, then, Kitcher seems to be betting on the outcome of someone else's naturalization projects. Since we know that he owes no heavy debts to folk intuitions, it seems likely that rather than favoring the second approach below ("conceptually regulated scientific naturalism"), he may instead favor the third ("conceptually indifferent scientific naturalism"). This would make sense, particularly since Kitcher's own two-factor theory of reference is the general type of theory of reference that undergirds naturalization proposals of the third type. Not only would such an approach be useful for determining the reference of "belief," but insofar as theories of content themselves are intimately linked to theories of reference, Kitcher's theory may itself play a role in explaining the nature of content and thus, of belief. (At the end of our discussion of the third approach, we will, in fact, briefly mention how Kitcher's theory might be used to articulate a notion of content that is compatible with understanding the content of our "most sophisticated performances," i.e. our advanced scientific beliefs.)

A final naturalized epistemologist who helps to indicate the centrality of a naturalistic account of belief to his epistemology is Kornblith (2002). In Kornblith's view, knowledge is a natural kind, to be understood according to the causal homeostatis theory of reference of Boyd (1991), a two-factor theory similar in many regards to Kitcher's. As discussed in our introductory chapter, Kornblith argues that the category of *knowledge* has "theoretical unity," because of the way in which this particular capacity of organisms reliably produces true beliefs and permits successful action. One might think that Kornblith would, therefore, seek to show *belief* to be a natural kind, but this is not entirely clear. What is clear is that he does seek to offer some manner of naturalistic account of belief, drawing on investigations in cognitive ethology (the study of animal cognition). He argues that we can understand

beliefs as a particular sort of information-bearing state, where information-bearing states are understood as internal "representations" that enable everything from thermostats to ants to process information from the environment and respond to it in a particular fashion (2002, 35–7). What distinguishes belief from mere informational state is that belief is the kind of informational state that is available to connect to other informational states and, in doing so, "inform an extremely wide range of behavior" (42). Here Kornblith clearly has in mind the functionalist idea that we define beliefs by reference to their causal role in relation to inputs, outputs, and other mental states. We need to posit the existence of such functional states, he thinks, because we cannot understand the full complexity of animal behavior in the way we understand the actions of plants, strictly by reference to their biochemistry. He invokes Fodor's (1974) discussion of the explanatory value of higher-level properties of the special sciences. Just as *camshafts* may not reduce to a single physical type but explain automotive "behavior," the functional properties of *beliefs* may be multiply realizable in many media while still providing explanations of animal behavior. And Kornblith thinks we need to appeal to this higher level property of beliefs in order to explain complex animal (especially human) behavior.

Of course this suggests that "belief" is *not* a natural kind in the sense of a natural *physical* kind. What this means for the status of knowledge as a natural kind depends on whether it is possible for a natural kind to supervene on properties that are not themselves natural kinds. Perhaps there *is* a natural kind of *animal* belief, or *human* belief, however, and Kornblith has a way out. In any case, whether or not we can really regard belief or knowledge as natural kinds, Kornblith is still appealing to a broadly functionalist naturalization strategy. In fact his particular strategy seems particularly amenable to our third naturalization proposal below (conceptually indifferent scientific naturalism). Its view of content (as an internal state that is available for manipulation to deal with the environment) is strikingly similar to some views of content (especially Cummins (1989) and Waskan (2006)) that we

will examine below. Since Kornblith does not elaborate on his view beyond a few pages, we will need to look to the philosophers of mind discussed below to do the job for him.

This seems to be the case, in general. None of the major optimistic naturalistic epistemologists offer substantive account of their own of how belief is to be naturalized. Instead they rely on the proposals of others. Given the division of philosophic labor between epistemologists and philosophers of mind, this is understandable. However, as we shall now see, the philosophers of mind who endeavor to naturalize belief almost invariably have different purposes than those of the naturalized epistemologists.

Belief naturalization proposals

In the remainder of this chapter, I will outline three separate proposals for naturalizing belief, and argue that none achieves the task in the manner needed by the naturalized epistemologist. I borrow the taxonomy of naturalization proposals from Michael Tye (1992), who divides them into "analytic naturalism," "conceptually regulated scientific naturalism," and "conceptually indifferent scientific naturalism." In this section of the paper, I will draw on Tye's taxonomy to argue that none of these naturalizations proposals are satisfactory for the purposes of the naturalized epistemologist. Tye himself comes to the same conclusion, but I will produce my own reasons for thinking it, while at the same time updating and enriching the description of which more recent proposals fall under these categories.

The first category maps neatly onto the proposal of the same name for *knowledge* naturalization mentioned in the first chapter. The second and (part of the) third categories fall under what I've called "two factor semantical naturalism" about knowledge in the first chapter. So each of these proposals for how to naturalize *belief* derives from a theory of reference applied to the *concept* "belief." In the present taxonomy, what I called "supervenience" naturalism in the first chapter is

subsumed under conceptually regulated scientific naturalism. (As we discovered in chapters one and two, supervenience offers very little when taken by itself, and must be given conceptual guidance.)

Interestingly, not every advocate of a particular knowledge-naturalization proposal also advocates the parallel belief-naturalization proposal. A case in point is Alvin Goldman, who unabashedly defends the purely analytic approach to knowledge, but gestures toward conceptually-regulated naturalism in his discussion of belief. It is not entirely clear what explains this in Goldman's case, but it probably expresses the fact that contemporary adherents to analytic naturalism are hard to come by, for reasons we are about to discuss. It is probably just too obvious that analytic "naturalism" is too analytic for naturalists to stomach. The second and third proposals are more popular, in proportion to the extent to which they move away from traditional conceptual analysis and towards unfettered scientific investigation.

Analytic naturalism

Tye (1992, 424) describes analytic naturalism as "the thesis that our psychological concepts have necessary and sufficient conditions for their application—conditions that may be elicited by *a priori* examination." The rationale for considering it a kind of naturalism is presumably that the proposed *analysans* would involve only concepts that are themselves naturalistically acceptable, referring either to behavior or physical stimuli or some other scientifically respectable properties.

A typical example of a traditional form of analytical naturalism is Ryle's (1949) analytical or "logical" behaviorism. According to this view, to speak of mental states of any kind (whether intentional or phenomenal) is just to talk about dispositions to engage in particular types of behavior, given particular stimuli. "Gilbert believes it will rain" just means something like "Gilbert is disposed to bring an umbrella with him." Of course there are numerous well-recognized flaws in this analysis. Gilbert may not be disposed to bring the umbrella if he wants to get wet, even if he believes it will

rain. Specifying this condition of course requires reference to further mental states, which defeats the attempt to define particular mental-state attributions in neat, purely behavioristic ways.

Problems like this lent credence to the push for analytic *functionalism*, which followed the lead of analytic behaviorism by defining mental states in terms of the output (e.g., behavior) and input (e.g., stimulation) of systems, but included under each the possibility of *other mental states* (defined the same way). So, for example, to speak of Gilbert's belief that it will rain, we must refer not only to his disposition to bring an umbrella, but also his desire to stay dry.

A rigorous method of specifying functional definitions of mental states was presented by David Lewis (1972), who argued that *all* theoretical terms are defined by their causal role. A detective, for example, might define suspects X, Y and Z in terms of their hypothesized roles in a murder plot, i.e., by their individual interactions with the victim and/or their interactions with each other, where these interactions are defined in pre-theoretical or "observational" terms. When the detective asserts that a theory involving X, Y and Z is true (using a "Ramsey sentence"), he offers an implicit definition of the terms "X," "Y" and "Z" (which can be formalized with a conditional "Carnap sentence"). If the story ends up being false, the terms fail to refer (although they still have a meaning in virtue of picking out "alternative possible worlds") (252).

Lewis argues that the same style of definition can be offered for mentalistic terms. Lewis tells us to collect "all the platitudes" we know of describing the causal role of mental states in terms of other mental states, stimuli, and behavioral output (256). We then formulate our theory of the mental using Ramsey sentences, e.g. this rough functional definition of a belief that it is raining (here I reinterpret some of Lewis' formalizations in terms of the example of the functional definition of belief I have already been discussing):

$\exists x \exists y$(x is caused by rain & x is caused by the desire-for-dryness y & x causes umbrella-brining) or

$\exists x \exists y$ (Rx & Dxy & Ux)

for short. This Ramsey sentence is then turned into a meaning postulate for mental state terms using a Carnap sentence:

$$\exists x \exists y\ (Rx \ \& \ Dxy \ \& \ Ux) \supset (Ra \ \& \ Dab \ \& \ Ua)$$

Where "a" and "b" name the mental states of belief and desire. Strictly speaking, this must be paired with another conditional describing what happens if nothing fits the description,

$$\sim \exists x \exists y\ (Rx \ \& \ Dxy \ \& \ Ux) \supset a\&b = *$$

where this means that "a" and "b" fail to refer. Pairing these conditionals gives us the analytic truth that either these mental states do not exist *or* our platitudes are true ("most of" our platitudes) (257). Furthermore, we can use this analysis to *reduce* the mental to the naturalistic if we identify x and y with some other independently (naturalistically) specified entities, call them p and q, which realize the truth of the mental platitudes.

The view of reference Lewis relies on here is strongly descriptivist. As Stich has noted, it is also the view of reference that underpins the case for eliminativism offered by Churchland (Stich 1996, 29–34). Folk psychology, on his view, offers an implicit functional definition of mentalistic terms, and if folk psychology's basic tenets prove to be false or lack predictive/explanatory power, then these terms fail to refer. This is a possibility Lewis considers: what he takes to be an analytic truth is not that folk psychology is true, but that if it is true, then "belief" refers (and if not, it does not). So one reason that analytic naturalism is not conducive to naturalizing belief is simply the possibility that the debate over the power of folk psychology might have an unfavorable outcome, and the eliminativist could win too easily. Different, non-descriptivist accounts of reference (and their corresponding naturalization projects) would permit folk psychology to be false or unreliable, but still permit folk psychological terms to refer.

Going into the debate about the power of folk psychology would be too much of a digression. What I would like to do instead is to say more about the methodology of this entire naturalization proposal. One point is that a descriptivist theory of reference is not our only option. We may instead

opt for an externalist or causal theory of reference, following the Kripke-Putnam thought experiments, or a hybridized causal-descriptive theory (a two-factor theory). There is then the question of how we might naturalistically decide in favor of one theory of reference versus another. Drawing on problems similar to those noted by Quine's inscrutability of reference thesis (1969c), Stich (1996) has argued there is no naturalistic way to decide the matter. We shall return to this question later under the heading of the third proposal. For now, we shall work within this descriptivist framework and concern ourselves with whether the analytic method it relies on is consistent with naturalism.

The more immediate concern with this style of naturalization proposal is whether any proposal appealing to *analytic truth* is naturalistically acceptable, given the long-standing naturalistic animus—following Quine (1953b)—against that concept. Stich (1996, 79–80), in particular, doubts that Lewis' dependence on analyticity can be squared with a fully naturalistic outlook. This raises a question: should contemporary "optimistic" naturalists follow Quine's critique, or find some way to accommodate themselves to analyticity? In a critique of Stich, Tim Crane (1998) says that Stich moves too quickly. He notes that there are contemporary views of analyticity, in particular Boghossian's (1996), which needn't be committed to the eccentricities targeted by Quine. We should briefly examine Boghossian's theory to see if it will do the work the naturalist needs.

Boghossian distinguishes *metaphysical* from *epistemological* analyticity. *Metaphysical* analyticity is the kind sought by the logical positivists: a statement is analytic in this sense provided that it "owes its truth value completely to its meaning, and not at all to 'the facts'" (1996, 363). *Epistemological* analyticity—which Boghossian supports—concerns not the source of truth but the source of justification: a statement is analytic in this second sense provided "grasp of its meaning alone suffices for justified belief in its truth." Boghossian rejects the metaphysical concept of analyticity on the grounds the *mere* fact that a sentence S means that P could never *make* S true. Even simple identities like "Copper is copper" are true, in part, in virtue of the general fact about the world that everything is self-identical. It would be equally absurd to claim that prior to our meaning

something by the sentence "Either snow is white or it isn't," it wasn't *true* that either snow was white or it wasn't (364–5). But Boghossian thinks that the metaphysical concept of analyticity can be safely rejected without rejecting the epistemological concept.

Boghossian characterizes epistemological analyticity in terms that are strikingly similar to Lewis' view of the meaning of theoretical terms. He adopts a "conceptual role semantics," according to which some expressions "mean what they do by virtue of figuring in certain inferences and sentences" (382). This is more expansive than Lewis' view, because it characterizes the meaning not only of individual terms via the sentences in which they are used, but also entire sentences via the inferences in which they figure. Boghossian thinks that theory of meaning is unavoidable for expressions such as logical constants like "not," "and," and "or": by themselves they have no distinctive "flash" meaning. As Frege emphasized, it is only in the context of entire sentences that they have any meaning at all. This conception of meaning, Boghossian says, points directly to the desired kind of epistemological analyticity: because, for example, the meaning of logical terms derives from their use in sentences and inferences that we take to be true, *if* those terms mean what they do, then those sentences/inferences have to be true/valid. Knowing, then, that the terms *do* mean what they mean, we then acquire *a priori* justification for believing in the truth/validity of the relevant sentences/inferences (and likewise in the truth or validity of any other sentences/inferences that determine the meaning of any other expressions we know).

Of course any given term or expression can be used in *many* different sentences or inferences. Boghossian realizes that the Quinean objection to conceptual role semantics is that the particular sentences or inferences that constitute the meaning cannot be isolated, but he convincingly argues that unless we assume some of them constitute the meaning, the very notion of meaning itself must be indeterminate. He notes that while most philosophers seem to agree with Quine's critique of analyticity, few agree with his arguments for the indeterminacy of meaning (Quine 1960). He urges

philosophers to go with their intuitions against indeterminacy, and therefore embrace the possibility of meaning-constitution, and with it, analyticity.

But that's it. Boghossian does not dissect Quine's arguments for the indeterminacy thesis. Perhaps this approach is legitimate to philosophers who place a high premium on their intuitions. There is, after all, a plausible objection that the indeterminacy thesis is a *reductio ad absurdum* of Quine's philosophy, rather than a surprising conclusion derived from mundane premises. But to Quine, at least, the premises from which he derives his indeterminacy thesis are central to his philosophic naturalism. Meanings, or propositional objects, are intensional, and lack the precise identity conditions naturalists would like. Indeterminacy also follows from meaning holism, which follows on one level from his confirmation holism (the Quine-Duhem thesis) (Quine 1969a, 80–1), and on another from his inscrutability thesis (Quine 1970, 182). In chapter 5, I will examine his arguments for indeterminacy in some detail, and suggest that both levels of argumentation are rooted in the hypothetico-deductive view of confirmation, which has long been cherished by naturalists. If naturalists want to reject Quine's indeterminacy thesis, they will need to reject these cherished naturalistic views.

But of course, if naturalists do this, we may wonder why they would need to be naturalized *epistemologists* anymore. As we learned in chapter two, *Quine's* reasons for wanting to naturalize epistemology are inextricably connected to his indeterminacy thesis. More broadly, Quine's reasons are connected to his rejection of the possibility of *a priori* epistemology. If, however, Boghossian's view gives us a conception of analyticity that permits us a source of *a priori* justification, it is less obvious why we need to naturalize epistemology. With *a priori* justification, we ought to be able to analyze our concept of knowledge and seek first principles of its justification. So even if Boghossian's argument vindicates analyticity—and it is not clear that, to the serious naturalist, it does—it then runs the risk of proving too much. Analyticity in the service of naturalizing belief runs the risk of obviating the very need to naturalize knowledge.

As a sidebar, it is interesting to note that there have been a few attempts to naturalize the *a priori* itself, that is, to show that some natural system apart from the senses could be responsible for the justification of some of our beliefs. Georges Rey (1988) argues for a view like this by supposing that we might have a subsystem in our brain capable of "grinding out the theorems of first-order logic," one that causes its possessor to be able to believe truths of logic (33). This would count as knowledge insofar as truths of logic are true if anything is, and *a priori* because beliefs in these truths would be accepted independently of sensory experience. Rey thinks that the model of justification here is simply that of the reliabilist: beliefs caused in this way would simply be "absolutely reliable," i.e. "the result of a process that reliably issues in true beliefs in *all possible circumstances"* (34). In answer to opposition from critics, Rey clarifies that beliefs caused in this way would be more than accidentally true. He claims that if logical truths can be specified through Tarskian recursion, using logical operators and referential devices, as true-in-a-language, then the logical truths are "those sentences that are true by virtue of the pattern of operators alone, independently of the assignment of the referential devices, i.e., they are true 'merely by virtue of their logico-syntactic form'" (35). The logical-synactic properties of the subsystem that cause belief in the truths of logic, then, may be the very properties that also make them true.

In criticizing this view, I will leave aside, for the moment, the fact that it presupposes the very notion we are attempting to naturalize: the notion of belief (though this is a major problem). I want, instead, to make methodological points. The core of this proposal is, of course, to account for the *a prioricity* of logical truth—not yet for a wider concept of truth in virtue of meaning. But an immediate worry is that even if truths of logic are true at least in part because of their form, Rey's contention that they are true by virtue of this form *alone*, and not in virtue of anything in the world, is quite tendentious. It seems to involve all of the problems of metaphysical analyticity that Boghossian rejects. But even if this did provide an acceptable account of the *a prioricity* of logical truth, it is difficult to see how this account would serve the naturalist's purpose by clarifying the methodology of

naturalization. The analytic naturalist, seeking to naturalize belief, needs more than an account of logical truth. Rey does offer to extend his account to include truths in virtue of meaning, by suggesting that we might also have beliefs that result from the application of our logical subsystem to certain meaning rules. Meaning rules, on this conception, would be "slots" in the "file" that functions as a concept, the slots that constitute the concept's identity by specifying rules for determining an extension (37). Now Rey acknowledges that it is very difficult to specify just which "slots" constitute the concept's identity. The problem here, presumably, is the same problem as Boghossian faces: which sentences or inferences constitute an expression's meaning, as opposed to other truths in which it figures? Rey says he is not concerned with specifying the meaning rules, just with saying that *if* we can specify them, then there is an available notion of naturalistic *a priori*. Perhaps, but my criticism of Boghossian applies here just as well. The rejection of any principled account of the meaning-specifying sentences or inferences is at the core of Quine's account of naturalized epistemology. Finding such an account is where the real work is needed, and if we had one, we may very well not need naturalized epistemology in the first place. The record of attempts to naturalize the *a priori* is not encouraging here. Others who have attempted the same strategy, such as Kitcher (2000), have also done so by characterizing the *a priori* via some kind of reliabilistic warrant, have concluded that even if there is a coherent naturalistic concept of the *a priori*, it's unlikely that there *is* any *a priori* knowledge of interest to speak of. Kitcher, in particular, argues that even mathematics may not be *a priori* under this concept of the *a priori*.

In the preceding discussion of both Boghossian and Rey, I have relied heavily on the idea that it is difficult to provide a naturalistic account of the meaning-constituting inferences or sentences of a particular term, particularly in light of Quine's indeterminacy thesis about meaning. I've said that this does not mean that Quine is right, or that there is no workable account (naturalistic or otherwise) of meaning. An interesting case is the view of Michael Devitt (1996), a serious naturalist who nonetheless espouses a "semantic localism" about meaning (as opposed to Quine's semantic holism,

which leads to his indeterminacy thesis). Devitt thinks meaning can be understood naturalistically via theory of reference, and that different theories of reference can help explain different kinds of semantic behavior (descriptivism works for some concepts, but he thinks terms of descriptions must ultimately acquire reference through causal connections to the world (160–1)). So perhaps a theory of meaning can be naturalized by a theory of reference. What is important about this theory is that even if it does provide a naturalistic account of meaning, Devitt insists that it lends no quarter to a theory of analyticity. Just because meanings are real does not mean we can know them through *a priori* conceptual analysis (which makes sense if some meanings are constituted by causal connections) (18-38). So this naturalistic theory lends no comfort to either metaphysical or epistemological accounts of analyticity.

This leaves open the question of whether naturalists like Devitt could still use an acceptable theory of reference to determine, *a posteriori*, whether or not there are theoretically important properties picked out by "belief." Perhaps a functionalist-descriptivist theory can be adopted without its baggage about analytic truth. Or perhaps another theory of reference entirely will do the job. We will indirectly explore some of these other theories while looking at subsequent naturalization proposals.[17] In the meantime, we should ask: if there is some account of meaning or reference that could be exploited without reliance on analytic truth, will it help the analytic naturalist? Do theorists have *a priori* access to the *meaning* of their concepts, even if they don't have *a priori* access to the truth? Devitt has suggested that they do not, and some important evidence seems to support him.

[17] But as we have already noted, Quine and Stich raise pressing problems about naturalizing theories of reference. Quine's thesis of the inscrutability of reference, in particular, holds that numerous incompatible reference schemes can explain the same behavior. Devitt, in particular, does little to address Quine's in-principle critiques of meaning and reference. He quickly dismisses Quine's argument from confirmation holism, on the grounds that it depends on a kind of verificationism that he takes most philosophers to find unacceptable. This is a mistake, however, because a closer examination of Quine's corpus suggests that his "verificationism" is not the crude sort so easy to critique (Raatikainen, 2003). He also offers no independent critique of Quine's indeterminacy of translation argument, which in my view is a reformulated version of the argument from confirmation holism (see "Quine's acquiescence in skepticism"). In any case, in our discussion of conceptually indifferent scientific naturalism, we will look to see if any available accounts of reference do the job the naturalized epistemologist needs, in particular the job of accounting for the reference of advanced scientific theory.

There is a substantial body of research in cognitive psychology, usually embraced by naturalists, that is widely thought to show that we do not have access to any necessary and sufficient conditions encoding the meaning of our concepts. This tradition of research is usually thought of as beginning with the work of Eleanor Rosch, who was inspired by Wittgenstein's "family resemblance" view of concepts. Rosch uncovered "typicality effects" in subjects' application of various concepts, evidence often cited as establishing that concepts are encoded by "prototypes", long lists of properties, *most* of which must be satisfied in an instance for a concept to apply to it, rather than an exhaustive list of necessary and jointly sufficient properties. So, for example, even if it is thought that one essential characteristic of being a bird is that an organism be capable of flying, an ostrich may still count as a bird if it has *enough* of the other prototypical features of birds. Analysts could allow that our definitions may stand in need of revision, naturalists say this means that we can never predict, *a priori*, how new discoveries might necessitate new methods of categorization. This is a point that seems to stand even if the full-fledged prototype *theory* of concepts does not stand: conceptual analysis simply does not account for the propotypical aspects of our concepts. Stich (1988; 1992), Tye (1992) and Ramsey (1992) all invoke these findings from empirical cognitive psychology to dismiss the possibility of analytic naturalism in regards to concepts of the mental.

Frank Jackson (1998) offers a response to naturalists who invoke this psychological research about the difficulty of *a priori* access to necessary and sufficient conditions. He says that even if *we* cannot list all of the necessary and sufficient conditions of a concept's application, this certainly does not mean that there isn't anything that it is to *be* whatever the concept refers to. There may still be some infinitely long disjunction of properties that determines what it is to be some phenomenon, for example grooming behavior. What the conceptual analyst has to do, says Jackson, is simply to do "enough by way of conceptual analysis to make it plausible that the purely physical account of our world makes true the grooming-behavior account of our world" (62). What exactly it is to do "enough" is of course an interesting question. He mentions Lewis' functional analysis of the mental as

an example, though concedes that it does not complete the naturalization project. A second stage is needed, in which the physical realizers of these functions are identified. This proposal is, in fact, at the heart of the second kind of naturalization proposal which we are about to examine.

It is just as well that we should move on to examine the next proposal, because this first (analytic naturalist) proposal is not very popular among naturalists. We have already noted that Goldman, although an advocate of analytic naturalism about knowledge, seems to favor a different proposal for belief. Even Lewis (1995), whose view of the meaning of theoretical terms is the most conducive to analytic naturalism, admits more recently that it offers only a "recipe" for analysis. Later on he seems to side more with Jackson on how that recipe is to be completed.

Conceptually-regulated scientific naturalism

Tye (1992) has described "analytic naturalism" as searching for *a priori* necessary conditions of the mental. In the preceding section, we have found that this proposal faces severe methodological problems as a naturalistic approach to belief and the mental, more generally. Not only does it seem to be difficult to naturalize *meaning* in the way a functional analysis of the meaning of "belief" would require, but it seems equally difficult to naturalize the kind of *a priori* access needed to exploit knowledge of that meaning for philosophic purposes. Because of considerations like this, Tye describes a second category of naturalization proposals that is intended to overcome the problems of the first. He calls this second category "conceptually regulated scientific naturalism," which he describes as follows:

> Scientific investigation, together with philosophical reflection regulated by our pretheoretical conception of mental states, is needed to come to a full understanding of their essences. (424)
>
> According to [this view], mental state types have non-mental essences. The task of the philosopher of mind is to specify what *sorts* of [non-mental] essences these are and correlatively to say which sciences will discover them, the primary constraint on any acceptable proposal being that it must be compatible with our ordinary, pretheoretical views about where the boundaries of the mental lie. (426)

Immediately we can think of some early naturalization proposals that might have fit this description. J.J.C. Smart's (1959) identity theory is one example. Smart thought that scientific investigation had revealed that the referents of our concept "pain" could be (contingently) identified with the stimulation of C-fibers. Likewise a Nagel-style reduction of the mental to the physical, which would connect primitively understood mental predicates to physical predicates through "bridge laws," would probably also count as this variety of naturalism. Both of these proposals would qualify as "type-type" physicalism, according to which a pre-theoretical mental type (e.g., pain) might be identified with or reduced to a physical type (e.g., C-fiber stimulation).

Of course towards the end of the 20th century, type-type physicalism lost much of its popularity. "Multiple realizability" arguments, in particular, suggesting that the mental could never be identified with or reduced to a single physical type, because there could be beings in other possible worlds—or even unknown beings in this world—which realized mental properties without having the same neurophysiological basis as ours (Putnam 1975; Fodor 1974). The question then became how to naturalize the mental in lieu of finding a distinctive physical property containing within it the key to all mentality. Of course the multiple realizability problem contained the seeds of its own solution: there must be some way in which we would be able to *recognize* these multiply-realized properties as mental, and the usual candidate is a *functional* criterion. So this much the newer approach shares with the analytic naturalist. But explaining how we identify properties as mental is not enough to *naturalize* them. More is needed, as Jackson observed at the end of the last section, to show how these mental tokens are *realized* in the physical. So perhaps mental types cannot be identified with or reduced to physical types, but they may *supervene* on the physical, as realized in functional types.

Tye gives a few somewhat unconvincing examples of naturalists who seem to fit the mold of his "conceptually regulated scientific naturalism. But clearly Jackson's idea of *beginning* with conceptual analysis (of the functionalist variety), and proceeding to find the physical realizers (of whatever type) of those functional types is a paradigm. Writing in 1992, Tye might not have been able

to predict the rise of an entire school of philosophic methodology, later in the 1990s, that would supply a semantics just for this view of naturalization. I am speaking of the "two-dimensionalist" semantics of Frank Jackson (1998) and David Chalmers (1996). Chalmers, in particular, applied the semantics to questions in the philosophy of mind. His approach seems to have been endorsed more recently by Jaegwon Kim (2005).

Each of these thinkers makes special use of the concept of "supervenience." Both Chalmers and Kim argue that while supervenience does not *reduce* the mental to the physical, it does offer a "reductive explanation" of the mental (Chalmers 1996, 43; Kim 2005, 93–120). According to Chalmers, a phenomenon such as belief is reductively explained by lower-level natural properties when it *logically supervenes* on those properties (1996, 47–8). Logical supervenience, in turn, is understood as follows: "[A]-properties supervene *logically* on [B]-properties if no two *logically possible* situations are identical with respect to their [B]-properties but distinct with respect to their [A]-properties" (35).[18] Less formally, B properties *determine* A properties.

Kim (2005, 101–2) offers a useful schematization of the steps taken in a process of reductive explanation:

> STEP 1 [FUNCTIONALIZATION OF THE TARGET PROPERTY]
> Property M to be reduced is given a *functional definition* of the following form:
>
> Having M = def. having some property or other P (in the reduction base domain) such that P performs causal task C.
>
> STEP 2 [IDENTIFICATION OF THE REALIZERS OF M]
> Find the properties (or mechanisms) in the reduction base that perform the causal task C.
>
> STEP 3 [DEVELOPING AN EXPLANATORY THEORY]
> Construct a theory that explains how the realizers of M perform task C.

I would now like to describe these steps in more detail, drawing in particular on Chalmers and his semantics.

[18] I have inverted Chalmers' "A" and "B" here to bring the formulation in line with earlier discussions of supervenience in this dissertation.

Chalmers says that the first step, "functional analysis," requires only a "rough and ready" analysis, and that it is common—and necessary—to begin a variety of reductive explanations in science in this manner. He gives the example of "reproduction": without an analysis of "reproduction" as some kind of "ability of an organism to produce another organism in a certain sort of way," science could never ascend from descriptions of relationships between complex entities and explain how these entities *reproduce* (1996, 43–4).

Now one might claim at this point that the same objections raised against the conceptual analysis of the analytic naturalist apply in equal measure to the first stage of this conceptually regulated reductive explanation. Problems included an unaccounted-for concept of meaning, uncritical reliance on a descriptivist theory of reference, and difficulty of *a priori* access to the necessary and sufficient conditions of concept application. But Chalmers has a more sophisticated view of conceptual analysis which offers answers to each of these objections. The sophistication comes from his view of semantics, which supports each of his responses. Notice that Chalmers makes liberal use of the concept of "logical possibility." In the earlier chapter on Kim, we explored how any notion of supervenience relies on *some* conception of possibility or necessity, and found that Kim's attempt to couch it in terms of nomological necessity was largely unsatisfactory. Chalmers' invocation of strict logical possibility bypasses that problem, but as a result he owes us an account of logical possibility and necessity. The account he will present is the same that fills out his distinctive view of conceptual analysis: it is a traditional possible worlds account (57).

When Chalmers describes the condition of supervenience as that in which no two logically possible situations are identical with respect to B-properties but distinct with respect to A-properties, he means there is no possible world with the same B-properties as ours but with different A properties (70). His account of the analysis of the *concept* of an "A" also invokes possible worlds. Chalmers divides the meaning of a concept into two "dimensions" (hence this is a "two-dimensionalist" semantics). The first dimension, which he calls "primary intension," is "a function from worlds to

extensions reflecting the way actual-world reference is fixed" (57). He makes use of Putnam's example of the concept "water" to illustrate this function. Our primary intension of water picks out the extension the "dominant clear, drinkable liquid in oceans and lakes" in each possible world: we say that water is H_20 in our world, that it is XYZ in another possible world, and even that water is *both* H20 and XYZ in a possible world if one occupies our lakes and the other occupies our oceans (57–8). This, then, is how Chalmers' primary account of the notion of meaning. Its only substantive philosophic assumption is the notion of possible worlds as primitives (66).

We can now appreciate Chalmers' responses to the second concern about analytic naturalism, regarding its descriptivism. Although he uses the example of "dominant clear, drinkable liquid in oceans and lakes" to illustrate how primary intension works, it is not the description so much as our dispositions to call things water that matter. "It is the function itself, rather than any summarizing description, that is truly central," Chalmers tells us (59). He says this is compatible with a causal theory of reference if our reflections on our dispositions leads us to think that reference is secured by a causal connection. For example, reflecting on our disposition to call watery stuff "H_20" on Earth, but "XYZ" on Twin Earth might lead us to formulate a causal theory of reference, given that we are in causal contact with the former substance on Earth, but not on Twin Earth. This is why I say that Chalmers' view is possibly viewed as a "two-factor" theory of reference. Usually causal theories of reference need supplementation with descriptive aspects, so if Chalmers permits a causal element, it is likely he would admit that both causal and descriptive factors function to achieve reference.

Finally, the contrast between primary and *secondary* intension allows him to account for the particular results of the Twin Earth thought experiments that are thought to undermine *a priori* access to important necessities. Kripke (1972) argues that sentences like "Water is H_20" are necessary but *a posteriori*: given the empirical discovery that water is H_20 in the actual world, it is H_20 in all possible worlds. Chalmers' response is that primary intension, being a function from worlds to extensions, is held in essentially conditional form: *if* watery stuff in our world is H_20, then it is water; if watery stuff

in another world is XYZ, then it is water, etc. This much, he says, is *a priori*, as it is determined merely by reflection on our speech dispositions. What Kripke is correct about is the *secondary* intension of water, which depends on the primary intension. Using the primary intension, when we learn that watery stuff is H_20 in the actual world, we fix the reference of "water" in the actual world, but then "rigidify" it and grasp that water is H_20 in all counterfactual possible worlds. Because if water is *this stuff* in our actual world, and this stuff is H_20, nothing that is not H_20 in other possible worlds can count as water, even if it is watery stuff. The trick, according to Chalmers, is that this *a posteriori* necessity depends on the *a priori* analysis of the primary intension. So while "Water is H_20" is not *a priori*, "Water is watery stuff" is (62). This is all Chalmers needs for his reductive explanation of the mental, because he thinks that the functionalist definition of belief is comparable to "Water is watery stuff" (79).

Understanding primary intension as a function from worlds to extensions also enables Chalmers to answer concerns raised by naturalist psychology about definitions and our *a priori* access to them. Chalmers acknowledges, of course, that crisp definitions in terms of necessary and sufficient conditions are not always available. But he argues that verbal definitions in terms of necessary and sufficient conditions are only often useful summaries of the meaning of our concepts, not the meanings themselves (78). The kind of meaning relevant to reductive explanation is primary intension, which is not a description but a function, expressed by our speech dispositions. The prototypicality effects noted by psychologists reveal what our speech dispositions are; as such, they reveal something about our primary intensions. For this reason they are no problem for conceptual analysis.

Having presented Chalmers' answers to objections to conceptual analysis, we have completed our description of Kim's first stage of reductive explanation, i.e., functionalization of the target property. Surely for theorists attempting to formulate a reductive explanation of belief, some verbal statement of a functional definition of "belief" is necessary. But Chalmers' point is that whatever the limitations to the process of definition, they are not significant, given that the theorist has more

immediate access to his primary intension. Having access to that primary intension, he can now move to the second and third steps of reductive explanation: identification of the realizers of the functional property.

Identity and reductionist theories have stumbled at this second step. Chalmers alleges that reductive explanation does not stumble, for two reasons. First, reductive explanation does not need to explain by reference to *types*: we need only offer reductive explanations of *tokens* of higher-level phenomena like belief (43). Supervenience provides the apparatus to offer this kind of explanation: a *particular* higher-level functional property can be said to supervene on a lower-level property if, already being in possession of the concept of the higher-level property, we can infer it from knowledge of the lower level property (76), or if we simply cannot conceive of a world with the lower level property *without* the higher-level property (73). Second, Chalmers thinks that the lower-level property need not be physical, strictly speaking. In our second step, we can descend to the level of neurophysiology, positing neurophysiological states as the realizer of functional properties. If we then explain how human neurophysiological states perform the functions in question, we will have completed our reductive explanation (without having to commit to the idea that all possible beliefs are realized in the same way) (46). But we also do not *need* to descend as far as the neurophysiological: he suggests that cognitive science could offer more "abstract" models of mechanisms giving "how-possible" explanations in terms of the known causal organization of organisms (46).

Chalmers thinks that such explanations are in principle available for psychological concepts like learning and belief—though notoriously, he thinks they are not possible for phenomenal concepts (because we can imagine zombies). As a result, Chalmers thinks that intentional psychological concepts like belief are "straightforwardly logically supervenient on the physical": whatever lower-level properties we identify as the realizers of belief, we cannot imagine beliefs differing where these realizers do not (82–3). Of course Chalmers thinks that because we can imagine conscious properties (phenomenal qualia) differing without physical differences (because of zombies), these do not

supervene. And he recognizes the possibility that intentional content may itself depend on phenomenal qualia, in which case intentional properties themselves might not supervene. But he thinks there is at least a third-person version of intentionality available that does not depend in this way. This version of intentionality, therefore, he takes to be supervenient. Supervenience, according to Chalmers, is the guide to judging the place of a phenomenon in the natural world (32). In this sense, we can say that Chalmers' theory of reductive explanation offers a *naturalization* of belief, or at least a proposal for how a naturalization might be achieved (and thus a plausibility argument for how it *will* be achieved). Insofar as this style of naturalization begins with *a priori* analysis of a primary intension, and ends with an identification of a non-mental essence, it looks like a good example of what Tye calls conceptually-regulated scientific naturalism.

Having outlined the most compelling proposal for a conceptually-regulated scientific naturalism (Chalmers'), we are now in a position to question its naturalistic credentials. First we need to question whether the concept of meaning (primary intension) exploited by this view is naturalistically respectable. Then we need to determine whether primary intension, even if naturalistically respectable, can be readily accessed in the *a priori* manner that Chalmers insists. Finally, we will look at likely candidates for the supervenience base, and whether or not they themselves can be naturalized.

As we have mentioned, Chalmers' reliance on both the notion of supervenience and the notion of primary intension depends on claims about logical possibilities and necessities, which claims are interpreted by reference to talk of logically possible worlds. In our earlier chapter on Kim's critique of Quine, we mentioned his reliance on the concept of epistemic supervenience and the naturalistic problems it faced. Having dismissed possible-worlds based accounts of possibility and necessity as hopelessly non-naturalistic, we noted how Kim attempted to characterize supervenience in terms of a *nomological* notion of possibility and necessity, but this proved cumbersome and implausible. Why, then, were we so convinced that possible worlds semantics was incompatible with naturalism? Much

of the reason is the presumption that talk of necessity and possibility generally appears to be irreducibly intensional (with an "s"). Quine noted as early as "Two Dogmas of Empiricism" that terms within the scope of modal operators fail to be intersubstitable *salve veritate*, and lack scientifically respectable extensional identity conditions (Quine 1953b; Quine 2004a). Since Quine is a paradigm naturalist, there is a certain presumption that skepticism about intensionality is a hallmark of the naturalist outlook. Talk of possible worlds would seem to feature the same problems, and probably more—given that it also seemingly relies on *a priori* access to these worlds, which naturalists are also likely to doubt (Brandom 2001, 598; Moreland 1998).

But Chalmers presents his reliance on possible worlds in a manner that some naturalists might find compelling. While he says that the notion of a logically possible world is to be treated as "something of a primitive," he also says that we should treat them "as a tool, in the same way one takes mathematics for granted" (Chalmers, 66). This is a provocative answer to the Quinean critique of modality, because Quine himself was quite content to treat mathematics in just the manner Chalmers suggests, as naturalistically acceptable if only because of its pragmatic indispensability for doing science. Now Robert Brandom (2001, 599) responds specifically to this practical indispensability argument. He says that it may very well be that actual scientific practice relies on modal notions, and for this reason, they must be seen as pragmatically indispensable. But he denies that this is sufficient reason to generate naturalistic respectability for use of modal terminology in *semantics*. Philosophical semantics needs to be more self-conscious and critical, in order to adjudicate the legitimacy of modal notions.

I presume that Brandom's reason for thinking this is that if we can be instrumentalists about the modal, we might as well be instrumentalists about "belief" or even about "knowledge," or any other philosophically controversial concepts. The point is that modal notions are at least as philosophically controversial as these mentalistic concepts—*and for exactly the same reasons* (both modal and mentalistic concepts are intensional). Given the equal amount of controversy, an

instrumentalist about modality would need to present a special reason for which that controversial concept needed no realistic naturalization, while the others did. I take it that optimistic naturalized epistemologists, who are realists about more than just belief, do want more than instrumentalism about belief, and certainly more than instrumentalism about knowledge. So it seems to follow that naturalists will need to produce some account of modality which is, if not a modal realist account, then at least an account that provides a special reason for which modality can be treated instrumentalistically, while at the same time situating it in a scientific context in a way that is consistent with our being realist about other things.

Genuine modal realism runs up against a variety of traditional naturalist and empiricist objections tracing back not just to Quine, but to Hume. This leaves the possibility of modal fictionalism, which I briefly examined already in chapter 2. If naturalists can sketch a naturalistic account of modal fictionalism, perhaps modality itself can be "naturalized," in effect providing us with our special reason for treating it in a non-realistic manner. Indeed there are theories of modal fictionalism available in the literature (see, e.g., Rosen (1990)). According to these views, literally speaking, existence claims about possible worlds are false; there is only the actual world. But possible worlds talk is really shorthand for literally true statements about certain convenient fictions. Translations for the shorthand would look something like the following (courtesy of Nolan (2002)):

Possibly P iff according to the fiction of possible worlds, P is true at some possible world.

Necessarily P iff according to the fiction of possible worlds, P is true at all possible worlds.

Presently, however, I will argue that these fictionalist accounts do not succeed in a manner favorable to the goals of the naturalized epistemologist.

One concern noted by critics of modal fictionalism is that the translations above count crucially on understanding "according to the fiction of possible worlds…." Rosen offers several possible translations for this construction: "If PW were true, then P would be true; If we suppose PW, P follows; It would be impossible for PW to be true without P being true as well" (1990, 344). But as

Nolan (2002) observes, these translations invoke modal concepts themselves, and would render modal fictionalism a circular explanation for the literal truth of modal claims. Attempts to reduce these modalities to a primitive modality also do nothing to advance the explanatory value of this account. Modality is still modality.

Even if a version of modal fictionalism could eliminate any concerns about circularity, an even more pressing concern looms for the usability of such an account by the naturalized epistemologist. What sense is to be made of "according to the fiction"? As Nolan (2002) notes, most presentations of modal fictionalism proceed on the assumption that modal fictions operate just like ordinary fictions (say, about Sherlock Holmes). But this of course raises questions about the ontology of fiction itself. Normally we would think about fiction as being a kind of counterfactual *representation*, a portrayal of how things might be but are not. Fiction even seems to have intentional content: stories about Sherlock Holmes are *about* a detective. But by supplementing theory of supervenience of the mental on the physical with an account of the language of possible worlds, it is representational content that we are trying to explain.[19] Thus it seems that appealing to modal fictionalism to account for possible worlds talk in a naturalistically respectable manner is a non-starter for the naturalized epistemologist, even if it could serve other purposes outside of naturalizing epistemology.[20]

---

[19] Of course Chalmers is interested in explaining *doxastic* representational content, which is obviously not the same as fictional representational content. Presumably fictional representational content would not have the same world-word relations as the doxastic kind (which is why it is fictional). But it almost seems that fictional content itself presupposes doxastic content: witness the extent to which fiction is often understood as involving "suspension of disbelief."

[20] It is perhaps worth mentioning that another naturalization proposal would be disqualified if the modality of possible worlds is truly unacceptable to naturalism. Although J.J.C. Smart's identity theory was widely rejected by philosophers of mind later in the 20th century, new versions of type-type physicalism eventually arose, exploiting Kripke's idea of *a posteriori* necessity. According to Block and Stalnaker (1999), "Pain = C-fiber stimulation" and "Consciousness = pyramidal cell activity" function much like Kripke's "Water = H_20." Since these are *a posteriori* necessities, they do not require the kind of conceptual analysis described in Chalmers' account. So technically speaking they are not conceptually regulated scientific naturalism, but conceptually indifferent. However, I do not plan to address this proposal under the next section. I mention it here because, by relying on Kripkean identities, the view obviously also relies on possible worlds semantics, and should be questioned as a legitimate form of naturalism for this reason.

None of this is to say that the style of reductive explanation described by Chalmers, one that relies on supervenience and possible worlds-semantics, is unworkable *per se*. Likewise for the workability of the concepts of supervenience or possible worlds-semantics on their own. It is only to say that they do not seem to be ideal candidates for naturalization methodologies.

Although the status of Chalmers' methodology seems to disqualify conceptually-regulated naturalism on its own, perhaps there are other brands of conceptually-regulated naturalism available. There are numerous advocates of versions of functionalism that presuppose that the functional properties of belief somehow supervene on naturalistically acceptable properties. Supposing the possibility that there are other views available, I need to raise a second objection to these proposals. Even assuming they find a methodology aside from Chalmers' to determine the nature of the supervenience relation, there are still important naturalistic objections to raise about the alleged supervenience *base*.

Recall, of course, that according to Chalmers—and many agree with him—we are to search for the realizers of belief by searching for neurophysiological or other cognitive systems that have a certain functional *causal* role. A number of different philosophers propose different types of causal role that could realize these functional properties. But what is the naturalistic status of "causality"? Chalmers is aware of this problem. He notes that by his own account of supervenience, laws of nature and facts about causal connections do not supervene on physical facts (86). That is, for the typical Humean reasons, we can imagine the course of nature departing from regularities we have observed in the past. Now a popular response to the Humean problem is to explain causal concepts in terms of counterfactuals and other modal notions. While this response may be legitimate to a non-naturalist, it seems highly dubious to the naturalist, given that counterfactuals are usually then expressed in terms of possible worlds. Chalmers says that he is willing to including physical laws in his supervenience base, on the assumption that there is "something irreducible in the existence of laws and causation," but admits that this "steps over the metaphysical puzzle rather than answering it" (86). Once again,

this may be a legitimate move for the non-naturalist, a naturalist who is willing to treat so many notions as irreducible may start to wonder why he couldn't just treat belief, or even *knowledge* as irreducible. This would of course obviate the project of naturalized epistemology.

At this point it might be objected that surely causal and nomological notions could be treated instrumentalistically, insofar as scientists surely rely on them constantly. Likewise, scientists surely seem to rely upon counterfactual conditionals in the very practice of setting up experiments, viz. "If this and this were set up, such and such would occur." Both of these points are true, but the question concerns their significance for the *metaphysical* status of these concepts. For these concepts to be effective in the description of the supervenience base for intentional concepts, they must of a metaphysical status that is clearer and less controversial than intentional concepts themselves, and the possibility of giving pragmatic equivalents of them does nothing to establish their metaphysical clarity. As Hume would have claimed, understanding causal and nomological concepts as reporting mere regularities is consistent with their pragmatic indispensability. Hume's own "skeptical solution" to the problem of induction—an early effort at naturalized epistemology, if ever there was one—was to say that our understanding of constant conjunction was merely a matter of "custom and habit," but not a reflection of any metaphysical relationships in the world. Yet Chalmers needs causal and nomological relations to be metaphysical if he is to exploit them as a supervenience base. Likewise, Quine himself (1994) recognized that a universally quantified truth-functional conditional could help explicate scientific experimental language, without resort to counterfactuals or other modal notions. If counterfactuals are really needed for anything (perhaps in the description of scientific laws), he also thought some nuanced version of a truth-functional conditional ("with a complex antecedent some of whose clauses are left tacit, to be divined from context and circumstances" (149)) could do the job. This pragmatic explication of certain scientific concepts does nothing to help secure the respectability of a metaphysical supervenience base, however. Whether it is realistic to fully expunge intensional concepts from scientific practice is, of course, a controversial question (see Hookway (1988)). If it is

not, this may well serve as an effective critique of Quine and of naturalism. Before such a critique succeeds, however, we have to recognize the indebtedness of naturalistic philosophy of mind to these putatively non-naturalistic concepts.

The problem of the modal status of causal concepts cascades into the various theories of the *content* of belief that have been proposed by avowed naturalists. According to functionalism, "belief" is defined implicitly by reference to its causal role, which includes its input, output, and relation to other beliefs. Particularly because of widespread dissatisfaction with descriptivist theories of reference, naturalistic theories of the content of belief usually appeal to a causal theory of reference to account for the "input" end of belief's causal role. Consider, for example, Jerry Fodor's causal covariationist theory of content (1987). According to this theory, cognitive content is determined by the reliable causation of mental tokens by the properties they are about. The problem for the naturalist with Fodor's theory is not simply that it appeals to the notion of causality. The deeper problem is that to make his theory plausible, to show that not everything that causes a mental token counts as a case of successful reference, he must also account for the possibility of misrepresentation or error. And to do this, he must find a way to "idealize" the causal covariation: it counts as successfully referential only under certain circumstances. For example, suppose "mouse" is reliably tokened by mice. A subject sees a mouse and tokens "mouse." This counts as successful reference. But if a subject sees a *shrew* and mistakes it for a mouse, he still tokens "mouse." This is not successful reference, according to Fodor, because if mice didn't reliably cause "mouse" tokens, shrews wouldn't either. But as Brandom (2001, 591) rightly observes, this appeal to counterfactuals once again requires robust modal resources.[21]

The problem Fodor encounters with misrepresentation is a problem concerning the normativity of intentionality. Other putatively naturalistic theories of content have proposed dealing with the problem through other means, which can also be described as broadly functionalist. Rather

[21] For problems that Fodor's theory faces on its own terms, see Cummins (1989, 55-66).

than focusing merely on proximate causes as a source of referential "input," these theories consider the broader historical influences on an organism's representational content, particularly evolutionary influences that determine *teleological* facts about the organism. As Sober reminds us, it can be useful to put the "function" back in functionalism (Sober 1985). Prominent teleosemantic theories have included Dretske's (1986) and Millikan's (1984). In a memorable example from Dretske, we learn about marine bacteria called with internal magnets (magnetoseomes) that work like compass needles to cause the bacteria to move deeper in the water in the Northern Hemisphere, because of the direction of the Earth's magnetic field. This behavior is explained by the evolutionary advantage of seeking deeper, oxygen-free waters, which the bacteria need to survive (Dretske 1986, 26). Dretske concludes that the function of the magnetosomes is to indicate the presence of oxygen-free water. There are problems about how best to specify the function of the magnetosomes here, some that place its representational content in more distal features of the environment, some that place it in more proximal features.[22] Regardless of how the theory is formulated to specify the function, *any* specification of function requires an account of teleology which, it turns out, depends on crucially modal concepts. Recent proposals for naturalizing teleological functions (such as Wright (1976)) look like the following "etiological" account (courtesy of Neander (2004), quoting Wright (1976, 81)):

> The function of *X* is *Z* if and only if,
> 1. *Z* is a consequence (result) of *X*s being there,
> 2. *X* is there because it does (results in) *Z*.

To understand "consequence" and "because," however, philosophers exploiting this theory of teleology will resort to counterfactuals. Brandom (2001, 594) formulates the typical gloss as applied to Dretske's example: "if it *had not* been the case that the mechanism…*successfully* led to the ancestors of contemporary bacteria to less toxic waters, then they *would not* now respond as they do to magnetosomes." So the usual naturalistic objections may be raised again.

[22] Even though the teleosemantic view was formulated in part because of the problem of misrepresentation faced by the causal covariance theory, this problem of how to specify the function leads to a misrepresentation problem of its own. See Cummins (1989, 73-75).

There are naturalistic epistemologists who embrace naturalistic accounts of normativity, even when they do not follow the teleosemantics view of content. They are, instead, more interested in naturalizing normativity for the sake of understanding the normative concept of justification, especially via the concept of reliability. Often accounts of cognitive evolution are used to account for this source of normativity (Kornblith 2002, 68). For this reason, I would like to mention some more general problems for understanding teleological functions in purely naturalistic terms, problems which arise apart from concerns about modality.

Marc Bedau (1991) grants the overall appeal of etiological theories of teleogy, but argues that they contain flaws which cannot be eliminated without bringing in non-naturalistic considerations. Bedau points to the example of certain clay-based crystals that seem to fulfill all of the criteria for natural selection-based teleology that Wright addresses. These crystals are produced through as chemical processes cause molecular structures to copy themselves. When small bits are broken off, these act as "seeds" to grow again into bigger crystals. Occasionally small defects in crystals occur due to external interference, but when new crystals are created from portions containing these defects, the reproduced crystal contains the same new structure. Furthermore, some of these structures are more stable than others, meaning that some endure better than others. Yet we do not want to say that these new structures function *in order* to permit the crystal to endure longer. It seems wrong to apply biological teleology to crystals, even though they exhibit many of the superficial traits of evolution through natural selection. I believe that there is a theory of natural teleology available, developed by Harry Binswanger (1990), that adequately explains why the action of cellular respiration exhibits teleological functioning, while the growth of crystals does not. Interestingly, however, it is of no use to a naturalist interested in using teleology to naturalize the normativity of intentionality, because it presupposes this intentionality.

According to Binswanger, we first come to grasp teleological concepts by grasping our own purposive behavior. We then project purposes onto other animals (children and primitives go too far,

and see purposes in plants and insentient nature), and explain their behavior in those terms. So teleology has its origins in our grasp of conscious purposes, but since only living beings are known to have these conscious purposes, it is intimately connected to the functioning of living beings. Eventually scientists find that they can formulate a concept of teleology ("goal-causation") that applies to non-conscious beings (such as plants): past instances of an action (e.g. cellular respiration) contribute to the survival of an organism, which in turn causes furtherance of the action (e.g. cellular respiration). In fact they understand this mode of teleology by analogy to purposive teleology.[23] This much of the theory resembles Wright's etiological account. But because our concept of teleology originates in the grasp of our own conscious purposes, this helps to keep even the non-conscious concept of teleology anchored to the biological. Crystals are not biological; therefore they do not count as teleological. There is a bigger story to be told here about why the concept of the teleological may be extended to the non-conscious but not to the non-living. According to Binswanger, it has to do with the explanatory power of goal-causation applied only to living beings (it explains not only superficial actions, but every aspect of their structure, down to the constant need for action at the cellular level). The main point, however, is that while this theory solves the problem of Bedau's crystals, and satisfies the judgments of biologists, it does so through an account of teleology that presupposes an understanding of an intentional concept ("purpose"), which is not available to the naturalist seeking to understand intentionality via teleology.

It looks, then, that there are serious difficulties for the naturalistic respectability of both the supervenience relation and important elements of the supervenience base, including everything from facts about causality to facts about teleology. But perhaps there is a naturalized account of modality that I do not know about. In that case, perhaps, naturalists could defend Chalmer's use of

[23] There are two aspects to the analogy. First, there is a commonsense analogy between the ontogeny of purposive action and etiological teleology: past instances of desire-satisfication also explain forward-looking desires in purposive teleology. Second, there is the history of the discovery of the theory of natural selection applied to phylogenetic teleology: Darwin himself understood natural selection in part because of an analogy to artificial selection, i.e. the purposive behavior of animal-breeders.

supervenience and primary intension. Even in that event, however, a final question to ask is: is there good reason to think we really have *a priori* access to primary intension, such that we can determine whether or not belief logically supervenes on a suitably naturalistic base?

Schroeter (2004) observes that there are a number of contemporary attempts to vindicate the methodology of conceptual analysis besides Chalmers' and Jackson's two-dimensionalist semantics, all of which allege to overcome difficulties with traditional forms of conceptual analysis. Most of these views take seriously the challenge posed by the Kripke/Putnam thought experiments, which they take to show that reference depends in some way on external factors, and hence that meaning "ain't in the head." The task for the new versions of conceptual analysis is to show that even if unpredictable externalist factors determine reference in some respect, there are other respects where this is not the case. In Schroeter's terminology, contemporary conceptual analysts concede that we do not have *a priori* access to the complete *applicability conditions* of our concepts, the complete truth about what it takes for something to fall under a concept in a given possible world, or the "semantically basic features" for a concept. But they do insist that we have *a priori* access to what Schroeter calls a concept's *reference-fixing conditions*. Rather than specifying the semantically basic features, reference-fixing conditions instead offer a generic "recipe" for determining the applicability conditions, usually via implicit metaphysical and epistemological assumptions. For example, these analysts tell us that the externalist thought experiments reveal that we have an *a priori* commitment to a *sortal* for various concepts, e.g. that water is a natural kind, one that is predominantly found in a certain state of matter having in certain locations, with a certain color, etc. Also the thought experiments reveal an *a priori* commitment to the idea that reference is to be fixed through our causal interaction with that natural kind. All of these commitments are said to be grasped *a priori* because we find that prior to empirical investigation, our strongest disposition is to call any substance "water" that meets these conditions. We may later discover that water in our world is actually H_20, and in that case we will call only those substances in counterfactual worlds that are H_20 "water," even if they do not fit

meet the same sortal conditions. But this is possible only in virtue of exploiting our *a priori* grasp of the reference-fixing conditions, first in relation to the actual world. Clearly Schroeter's "reference-fixing conditions" are functioning in much the same way here as Chalmers' "primary intension."

But Schroeter argues that this new picture of the *a priori* abandons infallibility about applicability conditions for an equally difficult infallibility about reference-fixing conditions. Let me illustrate what I take to be her objection by modifying her example about water to fit something Chalmers says about its reference-fixing conditions. Although Chalmers verbally summarizes the primary intension of "water," as the "dominant clear, drinkable liquid in oceans and lakes," he sees this as consistent with our *a priori* judgment that the ice and water vapor are also made of water. Presumably this is because he thinks that the sortal for water not only includes that it is a natural kind, but a specific type of natural kind that retains its identity through state changes. But imagine that we are Empedocles, and think that water is one of the four elements. It is hard to say if thinking of water as one of these elements is the same as judging it to be a natural kind in the way we do, but it is clearly the sortal under which Empedocles classifies it. For this reason Empedocles (or some other less-sophisticated ancient Greek) might not be inclined to say that water vapor is *water*. He might think water vapor is what we find when water transforms into *air*, or perhaps some mixture of water and air. What's more, ancient Greeks uninfluenced by modern psychophysics might not be inclined to say that what counts for successful reference is a causal connection to its referents. Aristotle, for example, seemed to think that mind could not be blended with the body at all, for fear "admixture" might hinder its potential to take in the intelligible form of every possible object of thought (Aristotle 1941, III:4).

Or imagine a case in which both metaphysical and epistemological assumptions combine to render judgments about "gold" very different from ours. Imagine a philosopher like John Locke, who is explicitly skeptical about the possibility of real essences. Even if he were told a story about the atomic number of gold, he might never agree that the substance we call gold could be a vapor (of the kind we now claim to be used in certain lasers), because of his conviction that the reference of the

term is fixed by a nominal essence specifying a yellow, malleable metal. If anything, it seems like assumptions about metaphysics and epistemology are even *more* variable than assumptions about gold or water. Recent work in "experimental philosophy" suggests that intuitions about reference are culturally idiosyncratic (Machery et al. 2004), which should be expected, given that not even philosophers agree on theory of reference. These intuitions seem, therefore, to be little help in specifying a useful method of *a priori* analysis. As Gary Ebbs (2003, 252) notes in a similar critique of contemporary conceptual analysis:

> The main problem with this proposal is that what we *actually* say when we find ourselves in a previously imagined situation almost always trumps our earlier speculations about what we *would* say if we *were to* find ourselves in that situation.

But Schroeter is not being entirely fair to the conceptual analysts. Perhaps they need only appeal to an *a priori* account of *justification* here. To say that these intuitions are *a priori* is not necessarily to say that they are infallible, but simply to say that they are independent of experience. Perhaps this is the view of some of the conceptual analysts. But I have two responses. First, this view may beg the question. The cultural idiosyncrasy of folk semantic intuitions suggests that they are *not* independent of experience, but learned, instead, from a predominant cultural-philosophical milieu. So there is even less reason to think that folk semantics is innate than there is to think that folk theories of gold or water are innate.

Second, even if our intuitions about reference are *a priori* in the sense of being independent of experience, this does not yet mean that they offer *a priori justification*, which is what the present group of naturalists wants—they want a conceptually-*regulated* scientific naturalism. It may be that we always need to start with some view about reference-fixing conditions to engage in any inquiry at all, but this may be only because we need to start with some view or other before we can acquire a justified view (after engaging in "reflective equilibrium"). This proposal, in fact, sounds very much like the final naturalization proposal, to which we shall now turn.

Conceptually indifferent scientific naturalism

Just because naturalists tire of looking for forms of *a priori* approaches to naturalization does not mean there is no other approach. We will consider one last approach, which Tye calls "conceptually indifferent scientific naturalism." According to this approach,

> mental states may well turn out not to have most of the properties we ordinarily attribute to them. Moreover, even if they don't turn out this way, it is certainly not metaphysically or conceptually necessary that they have such properties; and neither is it sufficient. So, any conclusions we draw from thought experiments which rest on intuitions we might have about the mental states of non-actual creatures in the light only of our ordinary, everyday conception of those states may well be in error. In matters of the mental, science, together with philosophic theorizing based on it, can be our only guides. (Tye 1992, 427)

Now Tye himself goes on to criticize this approach on the grounds that indifference towards our intuitions causes our naturalization proposals to go "out of control." "Why, on earth, should we accept a view that goes so directly against what we pre-theoretically suppose?," he asks. But Tye's criticism here is question-begging. By presuming that we need conceptual regulation to constrain our naturalizations, he assumes that conceptually-regulated naturalism is viable. We have reason to think it is not. More importantly, however, the advocate of the conceptually-indifferent approach may have ready answer to the question of why we should accept a view going against our intuitions: the theory may be empirically useful, allowing us to predict and explain important phenomena. Conceptually-indifferent scientific naturalism is really *pragmatic* naturalism.

Stich (1992) suggests that this is just the kind of answer that an advocate of a fully naturalistic naturalism about the mind should give. He points out that it is parallel to a great many other philosopher's endeavors in relation to existing sciences. Often scientists will use poorly defined or undefined concepts to yield empirical success (he gives examples of "fitness," "grammaticality," and "space-time.") The job of philosophers of science is to examine the scientific use of the concept and make its meaning more explicit, perhaps even to propose improvements. This is exactly the approach a naturalization of the mental might take, by looking to existing notions of "representation" in use by the best cognitive science, describing them and perhaps "patching them up." Stich even points to the work

of Cummins (1989) as exemplifying this approach in the philosophy of mind. Cummins admits that he is not trying to analyze any folk psychological concept. Which concept of "representation" we should to explicate, says Cummins, is a question of choosing a theoretical framework and finding the concept of "representation" that plays an explanatory role in it. Cummins says he wishes to explicate the notion of "representation" used in "'orthodox' computational theories of cognition," which "assumes that cognitive systems are automatic interpreted formal systems" (1989, 13). According to Stich (1992, 253), the upshot of this attitude is a pluralism about concepts of "representation":

> [I]f different paradigms within cognitive science use different notions of representation, then there isn't going to be *a* theory of mental representation of the sort we have been discussing. There will be *lots* of theories. Moreover, it makes no sense to ask which of these theories is the right one, since they are not in competition with one another. Each theory aims to characterize a notion of representation exploited in some branch of cognitive science. If different branches of cognitive science use different notions of representation, then there will be a variety of correct accounts of mental representation….I see no reason to suppose that there is a unique correct framework for theories in cognitive science.

Apart from the apparent plurality of concepts of "representation" to be found in science, there is another reason that motivates Stich's pluralism here. When discussing the first two naturalization proposals, we have seen how each is in effect underpinned by a theory of reference. Analytic naturalists generally determine the reference of "belief" by looking exclusively to descriptions associated with the concept, while conceptually-regulated scientific naturalists generally rely on two-factor theories of reference that are compatible with a causal-historical account of reference. But Stich (1996, 37–54) wonders what it would even mean to have a naturalistic theory of reference to begin with. (He asks this question with attention to determining the reference of "belief," not the reference of particular *beliefs,* but we will later see that the questions are interrelated.) He considers that a theory of reference might be an account of our folk semantic intuitions, or a "proto-science" identifying some scientifically useful "word-world relation." The folk semantic account, of course, would involve the same types of problems we have already seen in attempts to naturalize folk psychological intuitions: these intuitions, even when pitched at a generic level, are variable and fallible, and there is evidence

suggesting that they are culturally idiosyncratic. The proto-scientific account of reference would be as useful (or not) as the account of representation or belief that we are currently considering. Presumably different scientific research programs could exploit different word-world relations to explain different phenomena. So there would be no single correct theory of "reference": on some accounts, "belief" might *refer,* while on other accounts, the same concept "belief" might fail to *refer** (where "refer*" exploits a concept of "refer" useful for some different purpose). From all of this, Stich concludes that there are no determinate facts to adjudicate between competing theories of reference and determine a single, correct notion of "reference." For this reason, he thinks that looking to reference to determine ontology (what Quine called "semantic ascent" and what he and Bishop call "the flight to reference" (Bishop and Stich 1998)) is a hopeless pursuit. So just as there can be no single correct notion of "reference," there can likewise be no single correct notion of "belief" furnished to us by a single correct notion of "reference."

This pluralism about concepts of "belief" or "representation" is not necessarily a problem for the conceptually indifferent naturalist. If the lesson is that we choose scientific concepts of "representation" for the sake of their explanatory power in a particular domain of research, then provided that we have such a domain of research we are interested in explaining, we should be able to find a relevant notion of "belief" or "representation." Fortunately, we did enter this discussion with a research program: we entered from the domain of epistemology. So the crucial question we must now address is: is there a scientific concept of "belief" or "representation" that will suit the purposes of a naturalized epistemology, one that will yield an understanding of beliefs that can be true or false, beliefs that can be produced by reliable or unreliable processes? And most importantly, will this meet one of the goals mentioned by Kitcher, will it enable us "to understand and improve our most sophisticated performances," i.e., our most advanced scientific beliefs?

Because of the plurality of "representation" concepts that are possible under a conceptually indifferent approach, we obviously cannot examine every one for its conduciveness to epistemological

purposes. At best we can examine a few representative cases. A good place to begin is Cummins himself. Stich's imprimatur suggests, at minimum, that his outlook is purely naturalistic. So, does it help us with epistemology?

According to Cummins, representation is a different issue from intentionality. Whereas intentionality concerns the content of thoughts (conceived in terms of belief-desire folk psychology) Cummins is primarily concerned with the sort of representations involved in computational systems (1989, 88). This does not yet mean that computational representation will have no explanatory value in explaining mental content; it is merely a well-understood starting point that may shed light on issues beyond computation. According to this computational view, for a system to function as a representation is simply for its elements to be isomorphic to the objects represented. That is, if there exists an interpretation mapping elements of one system on a one-to-one basis onto elements of another, the first system *represents* the second system. A simple example is an adding machine whose buttons, internal states, and display serve to represent the argument, function, and output of various mathematical operations. Representation, then, is simply a kind of simulation by one system of another. This concept of "representation" is said to explain a important facts about practices that use representations, like the operation of calculators.

Explaining the operations of calculators is one thing, but what about the theorizing of scientists? At first, it might seem like Cummins' notion of representation is an ideal fit for explaining scientific reasoning. He says that computationalism "embraces a rationalist conception of cognition": a representational system counts as *cognitive* when it serves to facilitate behavior that is "cogent, or warranted, or rational relative to its inputs" (1989, 108). A cognitive system is an "inference engine" that relates propositional contents to other propositional contents. The objects of this kind of computation are symbols, which represent conclusions and premises (109).

There is quickly some trouble, though. If cognition occurs only where there is inference, and if inference is governed by the laws of the special sciences, then where there are no special sciences that

postulate inference-facilitating "natural kinds," it seems that there can be no cognition. If there are no special laws of clothing, for example, there might be no cognition about clothing (112). This computationalist theory of cognition, then, will work "only for the cognition of autonomously law-governed domains" (113). Cummins thinks this is unsatisfactory, and a more satisfying account of cognition will require a functionalist specification of various modes of cognition: "cognition will [need to be] what humans do when they solve problems, find their way home, etc." (113). We might stop at this point and wonder how exactly this functionalist specification of cognition is supposed to work. It is not clear if it would involve any of the non-naturalist commitments for which we have attacked other functionalists. (The chapter that describes the proposal does not make clear what kind of functionalism it is—only that it is to be distinguished somehow from causal and conceptual role functionalisms (115–9).)

But this problem of specifying cognition functionally already assumes that the computationalist view of cognition is more advanced than we actually have reason to believe, because it assumes that an isomorphic account of representation can furnish inferences among propositional representations. The problem about "clothing" cognition takes for granted that law-assisted inferences would be unproblematic, but really explaining the possibility of representation of and by scientific law is the greatest challenge for an isomorphic view. The closest Cummins comes to discussing law-based representation is an example concerning Galileo's use of geometrical diagrams to calculate the distance traveled by an accelerating body. There is a clear sense in which the structure of the diagram is isomorphic to the magnitudes of motion it is used to represent (94). But no where is any indication given how to generalize from this example.

Clearly much scientific reasoning does not proceed by the use of diagram, but through the use of mathematical and conceptual representations. If mathematical equations are isomorphic to the systems they model, then this must be so in virtue of our conceptual interpretation of the marks on the page. Consider a simple physical equation such as $f = ma$. Is the equal sign in this equation

isomorphic to some aspect of a body under force? No one would say it is. The equation derives from our knowledge of a proportionality relationship: given a constant applied force, the amount of acceleration is inversely proportional to the mass, hence the product of an arbitrary mass and an arbitrary acceleration is constant. The equal sign here represents no single aspect of a body under force, but indicates the fact that any given constant force creates a constant proportionality relationship between a range of values of mass and acceleration. If the marks on paper bear any isomorphism to these facts, it is only in virtue of our understanding of the concepts of varying force, mass and acceleration. Now of course once the equation is understood this way, it can be *used* in a way that is dynamically isomorphic to certain systems: as we change the acceleration variable, the mass variable will change in the same proportion as *acceleration* itself changes with respect to *mass* itself.

Of course, in a later work, Cummins (1996) himself admits and even insists that neither concepts, language, nor knowledge structures in general serve any representational role (131–46). So perhaps a physical equation was never meant to count as an example of a representation in the first place. This makes sense, since the idea that language is isomorphic with reality has gone the way of the *Tractatus*. Perhaps the full story of scientific cognition requires a story about how non-representational devices like equations get connected with other more clearly representational devices, or perhaps certain domains of scientific cognition do not require representation at all (recall that Cummins has distinguished representation from intentionality). Yet it does seem odd to count a geometrical diagram as a scientific representation but not a physical equation. Certainly many scientists *think* that equations represent important physical relationships without being isomorphic to them. And while *understanding* physical equations presupposes many other concepts and abilities, and that given this understanding, equations can be used in an isomorphic way, it still seems that the equation itself represents an abstract fact about physics that is not reducible to any isomorphism. All of this seems disingenuous to a naturalized epistemology that wants to explain the most "sophisticated performances" of scientists.

But perhaps this line of objection defers too much to our pre-theoretic concept of "representation," and does not yet take seriously using that concept as a term of art for a particular explanatory purpose. This is not the only gap between the isomorphic conception of representation and our pre-theoretic concept. By Cummins' own admission, the isomorphic conception of representation furnishes representations on the cheap. That is because any given structure will be isomorphic with a multiplicity of other systems. One consequence of this view is that there is a serious gap between representation and intentionality. Whereas isomorphic representations are "radically non-unique," intentional content is supposed to pick out exactly one state of affairs. To use his examples, intentional content is the kind that makes a thought about Central America just about Central America, and not also about Gödel numbers (1989, 137–8). One implication of this is that isomorphic representations will certainly not support any anti-individualistic notion of "wide content": the states of a representational system, on his view, are individualized entirely by the computational states of the system (116–7). He also doubts that any functional specification of cognition of the kind described above could ever help to produce intentional representations out of non-intentional ones (142–3).

The incompatibility of isomorphic representations with intentional content, particularly with "wide" intentional content, is especially threatening to the endeavors of naturalized epistemologists. In my first chapter, I already mentioned how some naturalists like Kitcher make significant use of causal theories of reference (closely related to theories of wide content), in order to explain the continuity of reference of trans-theoretic terms, in order to answer "pessimistic meta-induction" arguments against the reliability of science. But there are, perhaps, even more ways in which traditional projects in naturalized epistemology depend on wide content. One recent, provocative argument by Majors and Sawyer (2005) even suggests that the notion of wide content is crucial to reliabilist theories of justification, by way of serving to answer one of the most daunting objections faced by the theory. According to their argument, only an externalist reliabilist conception of justification keeps justification truth-linked. Yet an infamous objection to reliabilism brings about reformulations of

reliabilism that de-link justification and truth. A twin of ours living in an evil demon possible world might engage in the same reasoning practices as ours, yet be radically mistaken. It seems that since he is reasoning responsibly, he is justified. But then reliability is not necessary for justification. Reliabilism can be saved, say Majors and Sawyer, if true justification is reliability in one's "home world," where the home world is the one in which a subject actually develops, and his contents are individuated in a wide fashion. This allows us to explain how our twin, the victim of the demon, fails to be justified because he fails to have reliable beliefs: even though it may seem that his mental life is the same as ours, in fact it is not—because his mental states are individuated by a different environment. This may be a provocative and controversial use of wide content for a reliabilist epistemology, but it is, no doubt, consistent with a long tradition in naturalized epistemology of making use of externalist or causal theories. (Whether causal theories themselves can be naturalized themselves is, of course, a question for the previous section.)

Waskan (2006) proposes a conceptually indifferent naturalist theory of representation which is similar in many ways to Cummins', but which permits room for wide, intentional content. Waskan assumes that the main purpose of cognitive science is to explain how "we humans are able to behave in a such an unhesitating…and effective manner in the face of even highly novel environmental conditions" (90). In barest outline, this is to be explained by positing a capacity of forethought that permits us to represent the way the world is, and manipulate this representation in order to represent the way we would like it to be (90–1). Another way to think about this is: the purpose of the concept of "representation" is to explain and make predictions about our very ability to explain and predict. Like Cummins, he believes that the kind of representation needed to explain this ability is an isomorphism. But since isomorphism is "cheap," he also tries to specify the kinds of isomorphisms that are relevant to explaining behavior: these are isomorphisms between a subject and a system, which connect to the subject's behavior-guiding mechanisms and permit a subject appropriately related to that system to function successfully in it (96).

Now Waskan argues convincingly that even on this isomorphic conception of representation, the content of representations is still wide. But this does not imply that mental states *themselves* are anti-individualistic. Even if mental content is determined in part by external aspects of the environment, mental content is thereby a relational property of mental states, not constitutive of the identity of the states themselves (80–83). So wide mental content does not threaten *weirdness* of the mental; the mind itself does not "reach out and touch" the world. So mental states themselves can still be described in an individualistic manner, relevant to psychological explanation. What's more, simply because wide mental contents have no immediate causal bearing on subjects in the way that the psychological properties of mental states do does not mean that they have no explanatory value in their own right. Waskan argues that causal impotence does not imply explanatory impotence if *knowledge* of a causally impotent relational property can still furnish predictive inferences. Drawing on an example from Cummins (1996), Waskan tells the story of the Autobot, a small car guided by a slotted card that successfully navigates a maze, even though it does not come into contact with the walls of the maze. Even though the isomorphism between the slotted card and the walls of the maze is merely a relational property, and there is no immediate causal relation between the two of them, *knowing* about this isomorphism still helps us understand why the Autobot is successful (104). It seems that our knowledge of wide content could serve the same explanatory purpose: by using isomorphic representations of our own, we are somehow able to act successfully in the world (105).

Now much of this seems fine to me. I'm sure there is a sense in which isomorphisms can help explain certain kinds of successful behavior. The big question is whether the same conception of representation can account for what Kitcher calls "our most sophisticated performances," our scientific reasoning. Cummins says very little in the attempt to explain the relevance of isomorphic representations to scientific reasoning, but Waskan says a great deal more. In fact I think much of what he says is plausible and compelling. But as I will argue below, I think this plausibility comes from

presupposing, from time to time, the possibility of representations that are not themselves capable of being understood as isomorphic.

One domain of advanced thinking that Waskan relates to isomorphic representations is the domain of *non-concrete* representations. It seems difficult, he says, to understand our representations of properties such as being a war criminal, ownership, economic inflation, or electricity, in terms of any kind of pictorial isomorphism. Of course non-concrete value-laden concepts like "war criminal" and "ownership" will be difficult on anyone's theory (hence the entire discipline of value theory). But Waskan thinks that the concepts of "economic inflation" and "electricity"—along with many other non-concrete concepts, might be understood by analogy or metaphor to representations we can depict through pictorial isomorphism. Economic inflation, for instance, is presumably understood by direct analogy to actual, physical inflation. While understanding the causes and effects of electricity does not rely on analogy, understanding "the 'thing' itself" does require analogies to the flow of water through pipes, etc. (139).

Now Waskan mentions an objection to the reliance on analogies and metaphors by Prinz, on the grounds that "metaphors leave remainders" (Prinz 2002, 171–2). That is, to think of two things as alike in one respect is also to think of them as different in others. Flowing electricity is not literally flowing water. The remainder is what makes electricity *electricity*, rather than water. Waskan concedes this, and says that the remainder for "electricity" is to be handled by knowledge of the special causes and effects of electricity, which presumably are not shared by water. Where analogy and metaphor are useful is in understanding electrical phenomena themselves, as apart from their causes and effects. I think this position is entirely appropriate (indeed I make use of the cognitive power of analogy in a later chapter). And Prinz wouldn't seem to disagree, since he says that there is nothing wrong with researching the role of analogies and metaphors in cognition. But Prinz's point is that they "neither fully exhaust our understanding of abstract concepts nor explain how they get linked to their referents in the world" (172). This, presumably, is a point Waskan would agree with, but then

the question is: even if knowledge of the causes and effects of electricity is sufficient to distinguish our concept of electricity from our concept of water, where does our knowledge of causes and effects of electricity come from, and how does it get represented? Can it be represented through isomorphs of any kind?[24]

The knowledge of cause and effect that leads us to posit the existence of electricity is knowledge about static charges, about the transfer of charges, about the effects of transferring charges on magnets, etc. How might knowledge about charged physical objects be represented with isomorphs? Understanding the different types of charges involves understanding the results of a complex set of experiments (undertaken by Gilbert, Gray and Dufay) involving different degrees of attraction and repulsion of different types of materials. Now a single isomorphic representation in a physical system is presumably very good at representing another *single* physical system. But how does a single isomorphic representation come to represent the diversity of considerations involved in a concept like "charge"? Presumably we shouldn't be afraid to accept help from other representations that help condense some of this diversity for us, like concepts of materials and actions that help us to report the results of these experiments. But the question about isomorphic representation can arise again for some of these concepts. A single representation may be isomorphic with a piece of amber, for example. But pieces of amber come in a variety of different shapes and sizes. When Waskan said he could explain non-concrete domains of cognition using analogies and metaphors, it seemed that we were about to get an answer. But it turns out that many of the remainders of the analogies and metaphors turn out to be non-concrete themselves. This is not necessarily a problem, provided that these non-concrete remainders can themselves be understood in terms of some further combination of analogies or metaphors and other isomorphic representations. But this implies that *at some point* we must have some isomorphic representations that acquire the ability to represent generalized content

[24] In what follows I limit my answer to how I presume this knowledge is acquired for experts, because I assume that the layman's knowledge and beliefs about scientific topics are either incomplete or parasitic in some way on the experts.

(since analogies and metaphors will not create non-concrete representations without the aid of other representations). The question is: *how* is generality of representation created in the first place?

Probably because of this question, Waskan devotes a separate section to the question of genera, or "universals." To show how genera might be represented using his resources, Waskan gives the example of pre-algebraic, spatial proofs of the Pythagorean theorem (Waskan, 142–6). These involve first constructing literal squares on the sides of any given right triangle, and then, by a series of manipulations (as if using construction paper), showing that quite literally, the area of the square on the hypotenuse is equal to the area of the sum of the area of the other two squares. Now I have no doubt that many such proofs are possible in geometry, and no doubt that each of the manipulations involved in synthetic proofs could be modeled, and predictions about these manipulations made, using isomorphic representations. My concern is whether such a proof really helps establish any truths about *triangles in general*.

Of course there is an ancient explanation for how such proofs *can* establish a universal principle: as long as we recognize that the size or ratio of the sides has no bearing on the outcome of the proof, we recognize that the outcome is true of all right triangles, no matter their size or shape—the same point made by Berkeley against Locke in defense of nominalism. Waskan notes, of course, the likely objection that isomorphic representations would not necessarily account for this recognition that size and shape do not matter. He concedes this, but notes that it does not call into question the fact that the synthetic spatial proof can prove something about every triangle. It can if we have the recognition that size and shape do not matter, wherever that recognition comes from. I will grant this, but note that it simply raises a further question: where then do we get that recognition that size and shape do not matter, if not from representations? Waskan says it is the "combined effect of knowing that each individual manipulation made over the course of the proof would have had the qualitatively identical outcome no matter the initial lengths of the right triangle's three sides" (145). I will grant that there are some fairly primitive cognitive mechanisms that enable us to understand each step of the

proof without resorting to representations. At certain stages, for example, we need to grasp a triangle as staying the same shape through a rotation, or to grasp how two triangles, if lined up the right way, will form a straight line along one of their common edges. Each of these could be accounted for by perceptual-level, non-conceptual abilities. Even still, it seems that in order to know that these same manipulations would yield the same results for any triangle, one would need to be able to imagine any *triangle* going through the same operations. But to recognize these as *triangles*, even though triangles can be very different, would still presuppose an abstract, nonisomorphic representation of triangles.

I think there is a way an advocate of the isomorphism view could respond to this challenge, but let me take a brief digression by mentioning that Waskan does have another recourse available. He can use the appeal to metaphors and analogies to explain the very idea of *generality*. In one section, he makes the provocative suggestion that the very idea of category membership involve analogies to actual physical "containers," or "joints" along which nature is to be "divided" (150). Like all other analogies, however, this one too will have what Prinz calls "remainders." There is obviously an important difference between a particular object's relationship to a potentially infinite number of other particular objects (in the case of category membership) and a particular object's relationship to particular a container. Given these differences, fruitful use of the analogy will require answers to this question: In category membership, in what way does a potentially infinite number of objects have a "boundary" that separates these objects from other sets of objects, and such that some objects are "contained" within this boundary, and others are not? Obviously the boundary and the containment are not literal.

It may be tempting to answer this question by saying that it is not the job of a theory of representation to explain how this analogy is possible. It is enough that we *do* analogize things this way; other cognitive mechanisms besides representation may explain how we do. It may also be tempting to appeal to research in psychology that suggests a psychological mechanism accounting for our ability to regard things as members of categories that from a very early age, children have a

tendency to be "psychological essentialists," who categorize superficially different objects as if they shared "hidden mechanisms" in common. But is "as if" really enough to explain the ability to regard things as members of a category, or simply another way of restating the ability that is to be explained? Indeed, saying that we regard superficially different things as members of the same category simply because we think they share some common hidden mechanism is simply to push the problem to a deeper level, since the question must then be answered what makes it possible for us to regard the mechanisms as the same, particularly when they themselves are likely to be different in subtle and important ways.

Finally, there may be something fundamentally misleading about saying that we can understand category membership by an *analogy* to containers, because to say that we analogize two things is to say that we regard one as similar to the other in some explicitly considered respect. Yet it is precisely the ability to regard two things as similar in a respect that we are in effect trying to explain when we attempt to explain the possibility of thoughts about generality, so we cannot appeal to it in order to explain generality. Only if there is some mechanism that might account for an awareness of similarity that does not involve explicit appeal to the *respect* of the similarity might there be a way out.

I believe that the advocate of isomorphic representations may have a way of explaining the implicit grasp of similarity, at least perhaps for similarities with respect to simple attributes, like shape, for the sake of developing representations like "triangle." One answer is suggested by Prinz (2002), and I actually make use of this suggestion in a later chapter. Prinz describes how our most basic-level concepts might be grounded in experience through the use of representations he calls "proxytypes." A proxytype is the product of a "long-term memory network," a system that dynamically represents a *range* of possibilities through the transformation of a single image into another. We can, he says, "zoom in" on our representation of an object in a continuous manner, or even continuously transform an image of one object into an image of another. This kind of

representation allows us to see different objects as similar to each other, as the representation of one is easily reachable by transformation of the representation of another (141–44). This is a model which could, presumably, be applied to the concept "triangle." The differences between right triangle, scalene and isosceles are such that we can easily imagine transforming one into the other, into the next, just by growing and shrinking lines and angles. So perhaps there is a source of generality for the isomorphic representationalist after all, provided that isomorphs are understood *dynamically*. Presuming that proxytypes could form the basis for our first source of generality, we could eventually build on this source and acquire further concepts through analogies and metaphors, synthetic proofs, and other resources.

However, there are only so many things that can be represented using proxytypes, as Prinz is the first to admit (166). Strictly speaking, it is a theory for "basic level concepts," concepts of *objects* like chairs and tables, men and dogs, the kind of concepts children first learn and which research shows involve the maximum level of generality without sacrificing a high level of informational richness (Rosch 1978). (We can think of "triangle" as a basic level concept for shape concepts, though it is doubtful that shape concepts could be learned prior to object concepts.) It is very unlikely that proxytypes could be transformed to account for even one level greater of abstraction. For example, while it is comparatively easy to imagine differently shaped chairs transforming into each other, it is harder to imagine chairs turning into tables or beds or shelves, in order to account for the level of generality comprised by "furniture." This would not be a problem if only the concept "furniture" could be defined in terms of the more basic level concepts, but the only available basic level concepts would be "chair," "table," "bed," "shelf," etc., and a disjunctive definition of "furniture" as "chairs or tables or beds or shelves" would beg all of the relevant questions. The very problem to be solved is how these items of furniture come to be associated with each other, given that it is not as easy to imaginatively transform each into the other.

At this point I want to begin to wrap up, by suggesting that the problem faced by these isomorphic theories of representation is a problem faced by naturalistic theories of representation, in general. Consider that a naturalist might, at this point, abandon epistemology for metaphysics, and say that representations succeed in picking out more general properties in the world just in case they bear some appropriate causal relationship to them, perhaps via a reliable covariance between the property represented and the tokening of the representation, *a la* Fodor. Prinz, for example, seems to consider this solution for the higher-level concepts he can't account for with proxytypes (173, 241–51). Prinz considers a problem associated with causal covariance theories which, I think, represents a number of problems with *naturalistic* theories of intentionality, in general, including the isomorphism theory. This is what Devitt and Sterelny (1999, 79–81) have called the "*qua* problem." Suppose that I am in causal contact with a particular wildebeest (or have a representation that is isomorphic with it). But a wildebeest is also a mammal, an animal, a subspecies of gnu, and the prey of lions. Given that our representation is in contact with (or isomorphic with) this particular wildebeest, what then do we say this representation is a representation of? Which of these categories?

Now Prinz thinks he can solve the *qua* problem using the right kind of nomological covariance theory of content. We can say that our concept refers to *wildebeests* rather than to animals because whereas wildebeests reliably cause tokens of *wildebeest*, animals do only under special occasions. The usual problem with causal covariance theories of this variety is the disjunction problem mentioned in the previous section: what is to prevent the content of "wildebeest" from being "wildebeest *or* bush pig," given that the two can sometimes be mistaken for each other? Prinz dismisses Fodor's asymmetric dependence answer to this problem. Instead he invokes a Drestke-style answer that presupposes that there is a well-defined learning-period for a concept. He avoids the modal problems that usually attend these kind of idealization stories by saying that the content is fixed not by what *would* cause a concept during a learning period, but by "the actual class of things to which the object(s) that caused the original creation of the concept belong" (250). The problem is that this

solution to the disjunction problem itself raises its own version of the *qua* problem. Which class is picked out by that "incipient cause"? If it is the class of things that look like wildebeests, then it does nothing to solve the disjunction problem, and only deepens the *qua* problem.

The usual solution to the *qua* problem offered by naturalists is to offer two-factor theories of reference. We have already seen this at work in the semantics of David Chalmers, who appeals to the intuitive dispositions of the "primary intension" to fix a sortal that is affixed to a causal source of reference. But other even more palpably naturalistic theories rely on the same strategy. Stanford and Kitcher (2000), for example, show quite convincingly how the *qua* problem can be overcome provided the proper kinds of background beliefs or knowledge. The present point is that this solution only works if we can take the reference of that background theory—*its* representational content—for granted. And yet it is precisely that kind of content we are attempting to account for: if the *qua* problem applies even to a concept as close to perception as "furniture," the generality of great portions of our background theory will need to be accounted for.

There are, in fact, interesting parallels to be drawn between the problem of cheap isomorphism and the *qua* problem. The first of these problems was, in effect, that an isomorphic representation represented too much: very many things are isomorphic to a single representation. The second is that a particular isomorphic representation does not represent enough: it cannot, by itself, represent general categories, because the particular members of general categories differ too much. Cummins thought that the first problem could not be solved, not even with the addition of further cognitive resources: if something represents in virtue of its isomorphism, nothing about how we use it or the further cognitive resources we bring to bear will *stop* it from being isomorphic to too many things. A parallel point exists for the *qua* problem: if isomorphism between a single representation and a single object fails to account for *any* amount of generality of representation, it is unclear how the addition of further cognitive resources would increase the amount of generality without themselves relying on additional

general representations of their own. Indeed, as I have argued, cognitive resources like analogy only seem to do the trick when they presuppose other conceptual representations.

It is interesting, incidentally, that Prinz uses the example of "wildebeest" to illustrate his causal covariance theory. It is plausible that there could be a causal covariance between wildebeests and "wildebeest" representations, but it is plausible because something like the proxytype theory can account for *how* there could be such a covariance: the nearly effortless psychological ability to see wildebeests as similar would lead to this tokening. But when we are talking about higher-level concepts, the explanation for the causal covariance is not as obvious. Different pieces of furniture are very different in shape and size. The psychological channel that might have enabled the covariance, via proxytypes, is not available in this case because of the difficulty of imaginative transformation. So the usual solution is to find some property intrinsic to pieces of furniture—like a function—which is common to all, amidst their many differences. Interestingly, a similar approach is at work in the supervenience views we considered in the earlier section. Supervenience is just a kind of *vertical* covariation of properties, rather than the horizontal kind. Yet properties, treated as causal agents—in either vertical or horizontal covariation theories—are what naturalists like Quine would call metaphysical "creatures of darkness." This is no surprise, since Quine (1953c) thought that reified attributes were just as intensional as meanings and modalities, having the same difficult identity conditions (e.g., the attribute of exceeding 9 = the attribute of exceeding 9, but the attribute of exceeding the number of planets ≠ the attribute of exceeding 9).

If the *qua* problem is as serious as I suggest, it may not just be a challenge, but an insuperable barrier. For there are those who think that higher-level intentionality can never be naturalized. The *qua* problem points to the fact that at some level of abstraction, understanding representation may require appeal to an irreducible element of intentionality. Indeed the problem may be present even at the beginning of abstraction. I have suggested the proxytypes could account for the most basic recognitions of similarities, but of course we *do not* imagine every possible triangle, even if we *can*

imagine every possible triangle. Triangle proxytypes seem to offer at best the potential to represent triangles, but only if we add some kind of wordless order to regard all possible transformations within a range as triangles.

One might respond to the *qua* problem as Cummins suggests some might deal with the problem of "cheap" isomorphism, by indicating that content needs to be individuated functionally. So, for instance, one might say that it is not just the isomorphism to a wildebeest that makes for a representation of a wildebeest *qua* wildebeest, but the particular way in which wildebeests interact with us, how our representation allows us to deal with them, etc. Suppose, for example, that we always defend ourselves against wildebeests in a certain way, but not against other mammals. But this solution is more of an abandonment of theory of representation than it is a improved theory of representation. If it is possible to account for general types of responses to objects just in virtue of properties of the objects, we do not need isomorphic representational middlemen to account for our behavior. This would be a Wittgensteinian use theory of meaning, rather than a representational theory. Naturalistic or not, it is difficult to see how this kind of theory would deliver an account of cognitive content of the sort needed by the naturalized epistemologist.

Conceptually indifferent naturalists may well offer accounts of representation that offer genuine explanatory value without resorting to non-naturalistic assumptions. They may also offer theories of representation that plausibly show how the reference of scientific beliefs might be fixed. The problem is that the useful naturalistic concepts of representation do not seem to account for either the uniqueness or generality of scientific reference. Yet this is what is needed from a conceptually indifferent naturalism. We do not require of it that it satisfy our folk intuitions about representation, but we do require that it serve a useful purpose. Since we came to this discussion looking for a notion of representation that would serve our purposes as naturalized epistemologists, concerned with explaining the "most advanced performances" of scientists, it appears that conceptually indifferent naturalists face quite a challenge.

Conclusion

In the above, I have argued that naturalization proposals for the concept of "belief" (or "representation" or "intentionality") fail to deliver the goods needed by a fully naturalized epistemology. Analytic naturalism fails because of its reliance on conceptual analysis and on the substantive notions of meaning that go along with it. Conceptually regulated naturalism fails not only because of its even more implausible reliance on analysis, but because of its reliance on numerous substantive intensional concepts required to make sense of supervenience, none of which pass traditional naturalist muster. Finally, conceptually indifferent naturalism fails, not because it contradicts naturalist methodology in the way that the first two proposals do, but because it fails to deliver the kind of naturalized belief that naturalized epistemologists, studying the origin and justification of advanced scientific beliefs, need.

At the end of my first section, I noted that the naturalized epistemologists who maintained epistemology's need for a naturalized notion of belief were rarely in the business of doing that naturalization for themselves. Instead they decided to let the philosophers of mind do the job for them. But now we can see that this is one particular division of labor that proved inefficient. Even when naturalization proposals seemed more successful on their own terms (as in the case of conceptually indifferent naturalisms) they do not deliver goods in the proper form needed by the epistemologists. Sometimes division of labor is perilous. Sometimes, if you want to do the job right, you'd better do it yourself. The failure of the optimistic naturalists to do the job themselves, in the end, opens the door for pessimists, like Quine, who certainly *did* do the job for themselves—but with a much different outcome than the optimists had hoped for.