©2008 The Institute of Mind and Behavior, Inc.
The Journal of Mind and Behavior
Winter and Spring 2008, Volume 29, Numbers 1 and 2
Pages 99–138
ISSN 0271–0137

The Normativity Problem: Evolution and Naturalized Semantics

Mason Cash

University of Central Florida

Representation is a pivotal concept in cognitive science, yet there is a serious obstacle to a naturalistic account of representations' semantic content and intentionality. A representation having a determinate semantic content distinguishes correct from incorrect representation. But such correctness is a normative matter. Explaining how such norms can be part of a naturalistic cognitive science is what I call the normativity problem. Teleosemantics attempts to naturalize such norms by showing that evolution by natural selection establishes neural mechanisms' functions, and such functions provide the normativity requisite for a determinate semantic content. I argue that such attempts fail, because when specifying functions, and thus semantic contents, that are determinate enough to enable misrepresentation, they must tacitly appeal to human normative practices, especially the practice of giving intentional states as reasons for actions. I present a different tactic: using evolution by natural selection to avoid rather than solve the normativity problem. Representations' semantic contents and their intentional targets are irreducibly normative. Semantics and intentionality are constituted within human normative practices. However, evolution by natural selection can be used to naturalistically explain the transition from a world without human beings and human normative practices — and thus without any distinction between thoughts that may be called "correct" or "incorrect" — to a world in which such human practices and distinctions are commonplace.

Keywords: mental representation, normativity, naturalism

Representation is a pivotal concept in cognitive science, yet it suffers from a fundamental problem: there are serious obstacles to a naturalistic account of representations' semantic content and intentionality. The problem of specifying such content, as Ruth Millikan (1993c, p. 3) points out, is a normative problem. Of all of the different possible conditions that cause the representation to be activated, norms are needed to distinguish that subset of those

Requests for reprints should be sent to Mason Cash, Ph.D., Department of Philosophy, University of Central Florida, Orlando, Florida 32816-1352. E-mail: mcash@ucf.edu

events that are cases of correct representation from the cases of incorrect representation, or misrepresentation. A representation's content specifies what objects the representation should be used to represent, and what it should not be used to represent. Teleofunctionalists use evolution to attempt a naturalistic account of the norms involved, arguing that some mechanisms have evolved to perform the function of indicating or tracking particular external conditions, and that this evolutionary function provides the norms that specify each representation's content. I argue here that teleofunctionalists fail to reduce the normativity of representational content to non-normative properties of mechanisms and their evolutionary history, because attributions of function tacitly appeal to human normative practices; particularly the social linguistic practice of giving intentional states as reasons for actions. Nonetheless, I argue that evolution can be used in a different way, to show that explanations appealing to representations and their content can be respectable scientific explanations. While we cannot use evolution to naturalistically justify particular functional norms, we can use it to give a naturalistic explanation of how normative social practices evolved. This will not naturalistically *justify* any particular norm, as teleofunctional accounts attempt, but will give a naturalistic account of how things came to have a normatively constituted status, such as correctly representing an external object.

Representation is a Fundamental Concept in Cognitive Science

Prior to the last decade or so, it was rare to find an example of research in the cognitive sciences, especially philosophy, psychology, linguistics, and neuroscience, which did not assume that the brain contains representations that are computationally processed to produce decisions and actions. The concept of representation was so fundamental and so endemic that it was rare to find a book on cognitive science that did not stipulatively identify the discipline as united behind the assumption that all cognition involves represented information and computational brain processes operating on such representations. For instance, Dawson (1998) credits this shared assumption with enabling researchers from disparate cognitive science disciplines to speak a common language, and so to understand one another's research.[1] The recent recognition of the importance of seeing human beings as embodied creatures embedded in a physical and social world has somewhat muted the exclusive concentration on the brain in favor of the brain–body–world system, but apart from a few critical voices, most cognitive scientists still recognize a crucial role for neurological representations in explaining

[1] Von Eckardt (1993) similarly holds that the representational and computational assumptions define the discipline of cognitive science. Pfeifer and Scheier (1999, Ch. 2) detail the pervasiveness of these assumptions in what they call "classical" cognitive science.

many cognitive processes. Computational interactions between these representations and other neurological mechanisms are widely held to explain many of people's abilities and actions.

A problem for this representationalist approach, one long lamented by philosophers of mind and long ignored by many other cognitive scientists, is that representations represent things — they are *about* things — and aboutness is very difficult to incorporate into a naturalistic science. Although many theorists fail to distinguish them, there are two distinct (but not entirely independent) components to representations' aboutness: (1) having *intentionality*, is being about a particular object or state of affairs, being directed towards a "target," as Cummins (1996) calls it; (2) having a determinate *semantic content*, is ascribing particular properties to any target to which it is applied, saying something particular about it (that may be true or false of the target). For instance, my believing that the person walking away from me across the street is my friend Deidre, is explained by my having a representation that (1) has intentionality; it is *directed at* the person across the street. This person is the representation's current target. It also (2) has a determinate *semantic content*, in saying something in particular about its current target; it says that this person is Deidre (not Abigail or Brian or a complete stranger). Separating the two issues, we can see that there are two (related but separate) questions here. The first, which I'll call the *intentionality question*, is "What makes one bit of the world (such as a state of my brain) the kind of thing that represents other bits of the world? What makes something the kind of thing that can have an intentional target?" The second, which I'll call the *semantic question* is, "What makes a thing that has intentionality, have a particular semantic content. What makes it correctly represent only particular kinds of targets and so misrepresent other targets?"

The difficulty is in giving naturalistic answers to each of these questions. A naturalistic explanation of a phenomenon brings the study of that phenomenon into the study of the natural world. In naturalistic explanations, no terms referring to entities outside of the natural, physical, realm can feature. Our fully naturalized explanations must ground out, referring only to entities and processes that can be given a natural, scientific, explanation. As Fodor (1987, p. 97) characteristically puts the problem of intentionality and semantics:

> I suppose that sooner or later the physicists will complete the catalogue they've been compiling of the ultimate and irreducible properties of things. When they do, the likes of *spin*, *charm*, and *charge* will perhaps appear upon their list. But *aboutness* surely won't; intentionality simply doesn't go that deep. It's hard to see, in the face of this consideration, how one can be a Realist about intentionality without also being, to some extent or other, a Reductionist. If the semantic and the intentional are real properties of things, it must be in virtue of their identity with (or maybe of their supervenience on?) properties that are themselves *neither* intentional *nor* semantic. If aboutness is real, it must be really something else.

Most cognitive scientists and philosophers of mind concerned about this problem take Fodor's challenge seriously, and hold that representations' semantic content and intentionality are to be explained by reducing them to "something else"; to more physically respectable properties of neurological states and processes. Fodor continues (1987, p. 98), outlining the constraints on any attempt at such a reduction:

> Here are the ground rules. I want a naturalized theory of meaning: a theory that articulates in non-semantic non-intentional terms, sufficient conditions for one bit of the world to be about (to express, represent, or be true of) another bit.

The stipulation that a naturalized account of representation must be given in "non-semantic non-intentional" terms is due to the obligation to avoid circularity: these two relations are the phenomena we are trying to explain, and so cannot feature in the explanation. Thus any explanation of a representation's content or intentional target that itself appeals to or depends upon the intentional states of people is supposed to be ruled out, as not a naturalistic explanation. Intentional states must be shown to reduce to or supervene upon the kind of basic and irreducible properties of things that physicists will have in their complete catalogue.

The principal obstacle for a naturalistic account of semantics and intentionality is misrepresentation. It is a common feature of human mental lives that we often have false beliefs. I can see a person across the street as my friend Deidre, when it (embarrassingly) turns out to be a stranger with a similar hairstyle who happens to dress and walk like Deidre. I can believe that there is a cold beer in the fridge, but be disappointed to find the fridge empty of beer. Consider also the well-worn example of a frog that snaps at and catches flies, but which will also snap at beebee pellets that are lobbed past it. Cases of misrepresentation like these occur when there is a mismatch between what a representation's semantic content says of its current intentional target and the actual properties of that target. Most theorists of mind agree that the frog misrepresents the beebee pellet as a fly. The difficulty is in naturalistically specifying the content of the representation that controls the snapping action in a way that allows for the possibility of misrepresentation. This, according to the prevailing wisdom, requires us to naturalistically justify the claim that the representation *correctly* represents when its target is a fly and *misrepresents* when its target is something other than a fly, such as a beebee pellet. Attempts to naturalize semantic content must therefore use naturalistic relations to distinguish what Fodor (1990, p. 60) calls "type one" cases where the representation is caused to be activated by a target that it *correctly* represents from "type two" cases where it is caused to be activated by a target that it *misrepresents*. Such an account cannot beg the question by assuming that a mechanism whose activation is caused by both flies and beebee pellets correctly represents flies and not beebee pellets. Similarly, we cannot simply stipulate or assume that the representation caused to be activated by seeing Deidre and by seeing the stranger has a content that *correctly* represents Deidre, or that it *refers* to Deidre or that it is *true* when applied to Deidre. Terms like "correct," "refers" and "true" refer to semantic or intentional relations, and so stand in need of naturalistic explanation; they cannot ground an explanation of intentional targets and semantic content.

This is what Millikan (1993c, p. 3) refers to as the "normativity problem"; the problem of giving a naturalistic explanation of what turns out to be a normative distinction between what a representation *correctly* represents and what it misrepresents; between what targets it *should* be used to represent and what targets it should not be used to represent. This is, at its heart, the same rule-following problem that Kripke (1982) attributes to Wittgenstein; a problem that applies both to the meanings of words and to the content of concepts. The problem is that of precisely distinguishing correct from incorrect application. Past applications of the representation (word, concept) can be described as conforming to a potentially infinite number of distinct but overlapping norms. Which of these norms should we use in judging the correctness of future applications? Current dispositions to apply the concept in question similarly cannot specify precisely how the concept ought to be applied. Then there would be no distinction between how the user *takes* the concept to apply from how it *correctly* applies, and thus there would be no misrepresentation. If we are to make sense of the fact that people can misrepresent then we need to look further than current dispositions for a principled naturalistic, non-circular explanation of the distinction between correct and incorrect application of a representation to a target.

Fodor's (e.g., 1990) *disjunction problem* illuminates the difficulty in giving a principled naturalistic specification of a representation's content. Causal relations seem to be a good candidate for purely naturalistic relations. However, accounts that appeal purely to causal relations to specify the content, Fodor argues, fail to non-circularly specify why the representation should not be understood to have a more broad-ranging content, such that the alleged type two case of misrepresentation in fact counts as a type one case of correct representation. In the case of the frog, for instance, theorists who argue that the representation's content is the froggy equivalent of THAT'S A FLY, need to allow for the fact that beebee pellets can also cause the representation's activation.[2] They therefore need to defend against the claim that, since both beebees and

[2] Here I adopt the convention of using SMALL CAPITALS to indicate the content of a representation. Roughly, my BASEBALL representation is a representation with the same content as the word "baseball." The frog's FLY representation being activated by a target is the froggy equivalent of the frog thinking "that's a fly" of the target.

flies can cause the representation's activation, the representation's content is the disjunctive THAT'S EITHER A FLY OR A BEEBEE, or perhaps is something like the descriptive THAT'S A SMALL MOVING DARK THING. Either of these contents would correctly represent both flies and beebee pellets, and so the representation would be incapable of misrepresenting the beebee pellet. In the human case, purely causal relations cannot naturalistically justify the claim that the content of the representation activated when I see Deidre is THAT'S DIEDRE; a content *incorrectly* applied to strangers who happen to look and dress and walk like Deidre. Since this stranger also can cause the representation's activation, why is it not the case that its content is THAT'S SOMEONE WHO WEARS HER HAIR, DRESSES, AND WALKS IN SUCH AND SO WAYS? This content would *correctly* apply both to Deidre and to the stranger. We need to naturalistically justify the claim that this is a representation with the semantic content THAT'S DIEDRE, which incorrectly represents the stranger, and not one with the content THAT'S SOMEONE WHO WEARS HER HAIR, DRESSES, AND WALKS IN SUCH AND SO WAYS, which would correctly represent the stranger.

Attempts at naturalizing semantic content assume that it is possible to give a naturalistic explanation of the representation's determinate semantic content. It has proven very difficult, however, to satisfy the demand for a naturalistic account of such a determinate content; an account that provides a naturalistic justification for the normative distinction between targets a representation correctly represents and targets it misrepresents.

Teleofunctional Semantics: Using Evolution to Naturalize Semantics

Teleofunctional semantics, which explains the normativity of semantic content by appeal to norms of proper biological functioning, and which explains these by appeal to evolution by natural selection, is held by many to provide the most promising attempts to overcome the normativity problem and naturalize semantic content.[3] Ruth Millikan's (1984, 1989, 1993c) teleofunctionalism, for instance, stresses the way a history of natural selection can bestow a "proper function" on a biological structure, and uses the normativity of this function to distinguish correct uses from misuses. An item has a proper function, for Millikan, if it belongs to a type of items whose tokens are reproduced, one as a copy of the other, because of the function such items

[3] For instance, Colin McGinn (1989, p. 168) argues that although teleofunctional accounts have a long way to go, they "get more things right than any other theory I know of." He concludes that "we should just adopt it as our best working hypothesis. It accounts for a good deal, and I have yet to see it refuted." Robert Brandom (2001, p. 593) also applauds Millikan's approach in particular as "by far the most sophisticated and well worked-out exemplar of this kind of explanatory project" (it should be noted, however, that Brandom disagrees with this approach).

serve. This function distinguishes how the item *should* be used from how it is in fact used. A heart's proper function, for instance, is to pump the blood. Hearts also make noise and cause a rhythmic pulse, but their proper function is to pump the blood. Each heart has this proper function because of natural selection in favor of creatures with this type of organ that pumped blood more efficiently than creatures with alternative pumping mechanisms. Our long distant ancestors having an organ with four chambers, for instance, that pumped blood more efficiently than alternatives, brought it about that this type of creature reproduced more efficiently than creatures with alternatives. Thus the item was reproduced more than alternatives because it pumped blood better than alternatives. It was not reproduced because it makes a different rhythmic noise from alternatives. This, Millikan argues, specifies a functional normativity for that type of item: its function is to perform the task that led to its reproduction. This functional normativity allows for the occasional token of the type to fail to perform its function, without undermining the normative distinction between correctly and incorrectly performing the function. It simply has to be the case that the type is reproduced because its tokens perform that function often enough that they continue to be reproduced. Thus even particular hearts that do not pump very efficiently (or at all, for that matter) still have the function of pumping blood because performing this function is the reason hearts have been and continue to be reproduced. This is how such a heart can qualify as "defective"; by failing to do what hearts should do.

Millikan extends this analysis to explain how structures in a creature's brain can have the proper function of representing states of affairs.[4] A history of natural selection explains how entities in the brain can have the proper function of correlating with external states of affairs. They have been selected and reproduced by evolutionary processes because they correlate with environmental conditions. Such a neurological structure is used to effect and guide behavior that confers a selective advantage when such conditions hold. This correlation between the internal mechanism and environmental conditions has the consequence that another structure (which Millikan [1989, p. 285 ff.] calls the "representation consumer") can use this representation to guide behavioral responses that convey a selective advantage (over alternative behavior producing mechanisms) when those environmental conditions obtain. That such neural structures enable representation consumers to play this behavior-guiding role, explains their existence and continued reproduction. Such a structure

[4] When I use the term "representing" here, I do not mean to imply correct representing. Something that has the function of representing Xs has the content THAT'S AN X. This representation can have, on occasion, a Y as its intentional target; on such occasions, it misrepresents the Y because it has the function of representing Xs.

has thus acquired through natural selection the proper function of detecting such conditions. It thus can be described as a representation with a semantic content specified by the environmental conditions it is its proper function to detect. When it is activated, its intentional target is the object or condition that the representation consumer orients the creature's behavior towards.

Fred Dretske (1988, 1994) also uses a history of natural selection to naturalize semantic content in a similar way to Millikan, though Dretske's account focuses less on the benefit to the consumer of the representation in virtue of its consumption of the representation.[5] Dretske's account is more externalist, focusing on the particular type of environmental feature that causes the representation's activation. Dretske (1994, pp. 470–471) argues that internal structures whose properties depend in a lawlike way (or at least a reliable way) on particular external conditions are "natural indicators" for these external conditions. These structures are "recruited" by natural selection processes for controlling behaviors that confer a selective advantage when such external conditions hold. Because these indicators are used to produce behavior that is selectively advantageous, these structures acquire the *function* of carrying information about (or indicating) the kinds of objects or conditions for which that behavior confers a selective advantage. Structures with this function are representations, with contents determined by those objects or conditions that they have been selected for indicating.

Thus for both Millikan and Dretske, the function of the snap-initiating and guiding part of the frog's visual system is to represent or indicate or carry information about the presence of edible bugs, enabling a representation consumer to initiate and guide snaps at edible bugs, enabling the frog to eat them. It has this function because of natural selection in favor of creatures with this type of neurological mechanism that enabled frog ancestors to detect and respond in a selectively advantageous way to the presence of edible bugs (by snapping at and eating them). Having something that enabled frogs to snap at edible bugs and to not waste energy snapping at things that are not edible bugs, led to the creatures' reproduction, and thus to the reproduction of the type of mechanism that performs this function. Millikan (1993a) argues that unless we assume that the firing of the detector corresponds to the presence of an edible bug "we cannot account, *with any single explanation that covers historical instances of consumer success generally*, for *why* the consumer produces the effect that is its function" (p. 127, Millikan's emphasis). This mechanism, therefore, has

[5] Although I concentrate here on Millikan's and Dretske's accounts — and on their similarities rather than their differences — a similar critique can be made of other teleofunctional accounts, such as those of Neander (1991, 1995) and Papineau (1987). For some of the important differences between Millikan and Dretske, see Millikan (1993a). For an explicit contrast of all these accounts that applies a set of critiques roughly similar to the one I make in this paper, see Perlman (2002).

acquired the proper function of representing edible bugs. It is *for* representing edible bugs. It is not for representing beebee pellets or small moving dark spots, Millikan argues, because "none of the other correspondences . . . is relevant to *this* kind of explanation of the consumer's performance" (Millikan, 1993a, p. 127, Millikan's emphasis).[6] Only a correspondence with an edible bug can account for the reproduction of the creature and thus of the representation producer and consumer. The representation's proper function, therefore, is to be the bearer of a particular content: the content THAT'S AN EDIBLE BUG.

Millikan and Dretske both also explain learned associations in similar teleological terms. Dretske (1994, p. 475) treats learning as analogous to natural selection, arguing that learning can also recruit neurological systems that are natural indicators to the service of guiding behavior. Learning, he argues, happens in cases in which a somewhat intelligent agent needs to perform action A in conditions C and has a natural indicator of conditions C. This indicator is recruited for guiding A because such an action is rewarding in some way. For instance, I was not born with any innate behavioral response to cats. But as I learned to recognize cats and to interact with cats, a neurological mechanism whose activation reliably covaried with the presence of cats was recruited to guide cat-directed behavior because of this covariation. Because this behavior was rewarding (for instance by successfully helping me meet goals), this process of recruiting such "natural" indicators to behavior-guiding roles bestowed the function of carrying the semantic content THAT'S A CAT on this mechanism. Each activation of the representation carries the information THAT'S A CAT because it's a token of a type that has acquired the function of detecting the presence of cats and guiding cat-directed behavior. Its tokens thus correctly represent targets that are cats, and misrepresent targets that are not cats (such as a skunk on a dark night).

Millikan also keeps learning mechanisms grounded in evolutionary selection, arguing that natural selection has resulted in mechanisms that have the proper function of enabling creatures to reproduce rewarded or otherwise successful behaviors. Millikan adds to Dretske's account the notion — she claims (Millikan, 1993a, p. 133) it's possibly one Dretske would endorse — that "inner" rewards such as the confirmation of one's beliefs and avoiding contradiction also enable associative learning to recruit mechanisms as representation producers. She also adds the notion of "derived proper functions" to this account (Millikan, 1984, Ch.

[6] Millikan avoids a source of indeterminacy in the content by arguing that proper functions, including semantic content of representations, should be "described according to the most general principles available." Thus "a proper function of my heart is to help me to wiggle my toes, but only as falling under the much more general description of supplying my organs with oxygen and nutrients so that they may do whatever their individual jobs may happen to be" (Millikan, 2002, p. 124). This is why, she argues, the content should be described using the most general term "edible bug" rather than the more narrow "fly."

2) to explain the process whereby teleofunctions that have been built into an animal by natural selection can produce new teleofunctions by interacting with external conditions. Millikan (1993b, p. 225) gives the example of a rat encountering a new food. The rat will nibble a small amount and wait to see if it becomes sick. It learns not to eat foods that taste the same as the food that made it feel ill. No further selection is needed to effect this change in the rat's nervous system, but a mechanism in its nervous system comes to have a new proper function, of representing tastes of foods that should not be ingested. The proper function of the rat's representation (with the content FOOD TO AVOID) is derived from a mechanism that has the proper function of bringing about this kind of learning.

Millikan also uses the notion of derived proper functions to account for language (Millikan, 1986, Ch. 9, 1998). People have as their biological inheritance a set of mechanisms that have the proper function of bringing it about that the person learns conventions for the production and interpretation of utterances. Linguistic expressions, argues Millikan (1998, p. 36 ff.), have "cooperative proper functions," derived from the cooperative purposes to which speakers conventionally employ the expression and the purposes towards which hearers recognize the expressions are conventionally employed. The expression is reproduced (that is, continues to be used) because it serves this cooperative function. For example, a particular expression can have the proper function of drawing the hearer's attention to a particular type of circumstance or object. The expression "the door," for instance, has acquired the function of directing people's attention to doors through a long history of people using tokens of this expression (and its etymological ancestors) for that kind of purpose. It continues to be reproduced because both speakers and interpreters recognize that it serves this cooperative proper function. These teleofunctions of public language expressions, Millikan (1993a, p. 133) argues, "become translated into teleofunctions attaching to items in individual language learners' heads," such that a representation of a concept shares the proper function of the public expression with which the concept is associated.

On these accounts (especially Millikan's), my DEIDRE representation has a similar function to that of the word "Deidre" when I use it or hear it used. Through interacting with Deidre and engaging in conversations about Deidre, a representation producer has come to produce representations that have the proper function of representing Deidre. Representation consumers use this representation to produce successful Deidre-directed actions (including linguistic actions). These actions are directed at targets that are Deidre often enough that the representation producer can continue to be relied upon in guiding successful Deidre-directed actions. Rare occasions when a stranger is misrepresented as my friend Deidre may result in embarrassment, and perhaps in a refinement of the representation's content or of the conditions under which the representation is activated. However, these still count as applications of a DEIDRE representation because of the fact that we could not account for all my past successful Deidre-directed actions without assuming that the representation corresponds with Deidre, and so has the content DEIDRE. Assuming that it corresponds with a more proximal condition, and so has the content, SOMEONE WHO WEARS HER HAIR, DRESSES, AND WALKS IN SUCH AND SO WAYS, would not enable us to explain the Deidre-specific behavior it has helped produce. Nor would it account for the fact that tokens of a representation with this precise "aspectual shape" continue to be produced and used to guide actions.

Evolution by natural selection, it is plain to see, is the keystone of the above accounts of biological functions, which are used to attempt to naturalistically reduce away the normativity involved in specifying representational content. Millikan and Dretske appeal to evolutionary processes to explain how a mechanism can come to have the function of representing a particular type of object or situation, and thus can misrepresent when applied to a target that is different from those it has the function of representing. Downplaying the minor differences between their accounts, we could say that they both argue that this can happen in three ways: (1) through selection of a mechanism because of its covariance with the to-be-represented condition, (2) through a history of natural selection of a mechanism that guides learning and confers proper functions on mechanisms that represent entities the creature learns about, or (3) through a history of natural selection for the ability to learn a language community's conventions for producing and interpreting utterances. Thus the normativity problem of semantic content, for Millikan and Dretske, is to be answered by appeal either to the normativity of biological functions "designed" by natural selection processes, or to that of functions derived from such biological functions.

Teleosemantics Fails to Solve the Normativity Problem

I disagree that these teleofunctional accounts have naturalistically explained how a neurological mechanism can have a specific semantic content. In this section, I argue that they fail to reduce away the normativity, because judgments regarding which events count as actions for which representational explanations should be given, and regarding which actions count as mistakes and which count as correct actions, all depend on norms of human practices. These judgments depend both upon our general linguistic practice of giving intentional states as reasons for action, and upon the normative scientific practice of giving natural selection explanations of biological functions in which our interests drive ascription of functions to mechanisms. In the subsequent section, I explain why this intentional psychology is itself a normative practice, and that this is not reducible to non-normative properties of representational states. In the final section, I explain an alternative, non-reductionist, use of natural selection to naturalize intentionality and semantics by giving a naturalistic evolutionary account of the origin of such practices.

The problem with assuming that evolution by natural selection can determine a mechanism's function, and so its content, independently of our assessments can be seen by attending closely to the example of the frog that snaps at flies and at beebee pellets. Consider a variation on Dennett's (1995, p. 408) example of a frog enclosure at a zoo where the edible bugs that this species of frog has been selected for snapping at are not present.[7] In this enclosure, however, there are nutritious flying insects of a species that this type of frog has never encountered before. The question here is this: Does a frog that snaps at this new type of insect make a mistake? Judging it to be a mistake, in which the frog misrepresents the new fly as an OLD FLY gives us a troubling fixedness of content. This would not even allow the historical exaptations that happened, for instance, as the populations of flies changed over evolutionary time and the frogs' snap-guiding mechanisms were exapted for snapping at new kinds of fly. With this kind of fixedness, snapping at a new fly would count as a mistake even though it seems to be in the frog's interests to eat these new insects. Millikan's (2002, p. 124) "descriptive generality" principle, however, recommends that we use the most general description available, and thus would we not judge the case as a mistake, or even as an exaptation, but as a case of the frog correctly representing this new fly just as its ancestors represented the insects they snapped at, as an EDIBLE BUG.

We get onto a very slippery slope, however, in trying to decide just how general or specific we should be in our descriptions of the content. Millikan advocates the descriptive generality principle to avoid being what she considers too specific. She also wants to avoid being too general, however, such that the frog correctly represents even a beebee pellet as a SMALL MOVING DARK SPOT.

This desideratum of a just-determinate-enough description of content brings problems, however. Disagreement over the appropriate balance between specificity and generality is perhaps the principal difference between Millikan and Neander (1995). Neander distinguishes Millikan's "High Church" teleofunctionalism from her own "Low Church" version by advocating more proximal and general descriptions than Millikan's. For instance, Neander argues that the appropriate description of the frog's content is the lower level description SMALL MOVING DARK SPOT, such that the frog does not

[7] Dennett uses his account, in which there are no flies, and only food pellets lobbed past the frogs, to advocate a change in the meaning of the internal structure, with perhaps an indeterminate period in which it is not clear what to say about the structure's function. Like Millikan, Dennett sees the content as arising from the function of the internal state (see Dennett, 1995, p. 403). But unlike Millikan (1989, p. 284), Dennett sees the functions not as the systems *own* functions, but as functions interpreters have reason to attribute to the system.

misrepresent when it snaps at a beebee pellet. She argues that we should not use the most general, lowest level description possible (this would be at the neuronal components activated, which would not allow for any misrepresentation). Rather, we should describe the content "at the lowest level at which the trait in question is an unanalyzed component of the functional analysis" (Neander, 1995, p. 129). Neander allows for misrepresentation when the mechanism malfunctions, for instance when it causes the frog to snap at a lily pad. However, Neander's reason for staying at this slightly higher level seems to be that this level is the most informative. Although Neander sees this as information "in the objective sense in which one statement is more informative than another if it excludes more possibilities" (p. 139, n. 8), this still begs the question because even this allegedly "objective" sense of information depends upon our interests, judgments and explanatory activities.[8]

This kind of dependency on our explanatory interests is the general problem underlying the difficulty of defending (in a non-arbitrary, non question-begging way) a particular level of determinacy of content over other levels of determinacy. It will undermine all such attempts to use the history of natural selection to fix semantic content. Accounts like Millikan's must have more generality than the particular conditions that have historically caused the representation's activation, so that the frog can correctly snap at a new species of edible bug its ancestors never encountered. But they also want to have enough specificity that there are at least some possible cases of misrepresentation (if there are none, then this would undermine the claim that the system's function grounds a normative distinction between correct and incorrect cases). The problem is that any history of past selection conforms with a potentially infinite number of descriptions that agree about the cases so far, but disagree only about future cases. Consider, for example, a situation where there are very few nutritious bugs, and to supplement the food supply, zookeepers lob small nutritious food pellets through the air past the frogs, which the frogs snap at and eat. Does a frog that snaps at the pellet make a mistake here? Is this a case of misrepresenting the food pellet as an EDIBLE BUG? Or is it a case of correctly representing it as FOOD?

The crucial point here is that any principled answer to such questions of content will depend somewhat upon the history of natural selection (or in other cases on the history of learning), but it will also depend upon *our judgments and interests* — as theorists interested in explaining the snapping in representational terms and trying to specify the content of a mechanism. Answers depend upon our judgments about whether this case counts as *relevantly* similar to the historical conditions of natural selection. Importantly, these judgments will depend upon observers' decisions about what *matters* about the his-

[8] For a more detailed critique of Neander on this point, see Perlman (2002, pp. 284–285).

tory of natural selection. These decisions are ultimately based either on what we theorists think *the frog's goals ought to be* and whether these actions count as intelligent, appropriate or rational, or on the particular aspects of the process of evolution by natural selection that we think are most relevant or important for our explanatory activities.

To see why this is so, let's further explore the grey area between snapping at a fly and snapping at a beebee pellet. Imagine, for instance, that the food pellets that keepers lob past frogs to feed them are only half as nutritious as a fly. Is this still the same kind of thing the frog's ancestors snapped at? Would a frog snapping at such a food pellet make a mistake? Evaluations of "same kind" here would seem to depend on whether we judge that frogs in this situation should be catching flies instead and ignoring the pellets. Alternatively, imagine that the pellets are as nutritious as flies, but contain a hormone that sterilizes frogs that eat too many of them. What matters more here, eating enough pellets to survive, or eating enough flies and few enough pellets that it's possible to reproduce? Is snapping and catching the pellet still a correct response to the pellets? Is a frog that snaps at and catches a sterilizing pellet acting "successfully" here? Whether a frog's action counts as "intelligent" "appropriate" or "successful" is determined in part by the goals we ascribe to the frog. There are many long-term and short-term goals it would be reasonable to ascribe to the frog, but no fact about the history of evolutionary selection can determine whether we should describe the frog's goal as simply eating bugs, eating food, eating nutritious food, surviving to reproductive age, actually reproducing, or any number of other goals. Thus, it is not immediately clear how we should answer the above questions. Since the frog's ancestors were not presented with these sorts of environmental challenges, no fact about the history of natural selection determines whether we should describe current frogs as continuing a pattern of ancestral behavior or deviating from that pattern.

The original normativity problem remains. It is possible to give a potentially infinite number of descriptions of this type of representation's "successful" targets so far (going back into evolutionary history). The content could equally well be FLY, EDIBLE MORSEL, EDIBLE NON-TOXIC MORSEL, EDIBLE MORSEL THAT WON'T DECREASE MY CHANCES OF REPRODUCTION, or any number of other candidates. Any one of these descriptions could accurately describe all the "successful" frog snaps so far. The important point here is that even the thesis that the semantic content of this mechanism is SMALL MOVING DARK SPOT cannot be ruled out. Let us make the fairly safe assumption that, after an initial "tuning" period millions of years ago, flies and other edible insects have been the only small moving dark spots that members of a particular species of frog and their ancestors have encountered and snapped at (nobody has yet lobbed pieces of meat or beebee pellets past these frogs or their ancestors). Millikan (1993a, p. 127), you will recall, argues that "we cannot account, *with any single explanation that covers historical instances of consumer success generally*, for *why* the consumer produces the effect that is its function unless we assume that the relevant correspondence is with an edible bug." However, an explanation of the content that describes all these historical instances of representation consumer success as the representation's corresponding with SMALL MOVING DARK SPOTS would also correctly describe all the "successful" snaps. This also would explain, as Millikan demands, "*why* the consumer produces the effect that is its function." Since all the historical small moving dark spots were in fact edible bugs, the fact that frogs historically caught small moving dark spots will also explain why the representation consumer enables the frog to snap at and catch small moving dark spots: the frogs that caught small moving dark spots were nourished enough to survive and reproduce (and so the representation producers and consumers were reproduced).

Millikan's (1993a, p. 127) drawing a judgment, for explanatory purposes, about the "*relevant* correspondence" is revealing. This is one of the general problems with teleosemantic accounts of representations' semantic content; they depend upon the judgments of theorists trying to explain content, such as the judgment we make about whether objects the frog snaps at are relevantly similar to those its ancestors snapped at that led to their selective advantage. Similarly, Dretske (1994, p. 480), argues that a neurological mechanism that is a natural indicator of F situations and G situations by virtue of correlating with such situations, becomes a representation R, that signifies external condition F (rather than G), because it assists in producing an *intelligent* response to the presence of Fs; the kind of response that leads to a selective advantage in F situations but not in G situations.[9] R gains the function of representing the F-ness of things, he argues, by some further neurological process using it to produce and control actions that are *appropriate* responses to the presence of Fs.[10] These actions are *successful* because they employ mechanisms that signify the F-ness of things, where "successful" is defined in terms of the kind of "fitness" for which evolutionary mechanisms select. I should point out here that Dretske's aim is to show that representations as he defines them, without mention of intelligence or appropriate behavior, produce "intelligent" or "appropriate" behavior because of the mechanics of natural selection.[11] However, the problem is that we cannot decide on the appropriate

[9] "Something not only becomes the thought that F by assisting in the production of an *intelligent* response to F, it assists in the intelligent response *because* it signifies what it does" (Dretske, 1994, p. 480, initial emphasis mine).

[10] "If R is drafted to shape output because it supplies the information about when and where that output is *appropriate*, then . . . part of R's job, its function, is to supply this needed information" (Dretske, 1994, p. 480, emphasis mine).

[11] Dretske describes his aim this way in personal communication about an earlier draft of this paper.

descriptions of the representation's content, in response to kinds of changes to the frogs' conditions that I describe above, purely by reference to the conditions of natural selection. The decisions about whether the frog snapping in such a case counts as correct representation or misrepresentation can only be based on our judgments about whether snapping in this context is an intelligent or appropriate or rational thing for the frogs to do in this new context.

The only reason to declare snapping at a beebee pellet an exception to the norm (a case of misrepresentation) rather than conforming to it is that *we judge*, perhaps rightly, that frogs which snapped at too many beebee pellets would lose out in an evolutionary competition against frogs that did not. This judgment, however, is not based only on the facts about the conditions of natural selection that led to the reproduction of frogs with representation consumers that guided snaps. It is also based on *our valuing a certain aspect or a particular description of the process of natural selection.* For instance, it's based on our presumption that *if* there had frequently been beebee pellets in the frogs' environment, frogs *would have* evolved the ability to distinguish beebees from flies, and would have come to snap only at the flies. Either that or the species would have died out. Perhaps this is true. However, consider a species of frogs that have not yet encountered beebee pellets. In such a case, there is a very large (potentially infinite) set of descriptions of all the "successful" snaps so far; descriptions that differ from one another only in how they would rule on future frog snaps. It is only after we judge a frog snapping at a beebee to be a mistake, based on our assessment that frogs should not eat lead pellets, that we might have reason to select one of these descriptions over others. But we have no naturalistic, non-question-begging reason — one that applies independently of our explanatory interests — to rule that this is a mistake. Similarly, the other examples above of grey areas between frogs snapping at members of the exact same species of fly as their ancestors snapped at and their snapping at a beebee pellet, show that *our judgments* about which actions would count as errors and which would count as successes undergird any distinction between correct representation and misrepresentation, and thus they also undergird any ascription of content to the representation in question. The normativity problem cannot be avoided simply by appeal to the facts about the history of natural selection and facts about the kinds of objects the mechanism has guided the creature's actions towards in the past.

Of course, these facts about history of natural selection are not completely irrelevant either. The answers we give to the above questions will be influenced by these facts. But *these facts alone* cannot determine the precise answers we give. The normativity problem is still at play, because we cannot make decisions about these borderline cases by considering *only* the bare facts.

Someone might interpret the charge I am making here to be this: instead of appealing only to bare facts of natural selection, teleofunctional semanticists beg the question by (somewhat tacitly) appealing to intentional psychology to ascribe contentful thoughts to the frog, and then tailor their teleosemantic theories to justify those ascriptions of content.[12] Ascriptions of function, and so content, to mechanisms that causally produce actions make tacit appeal to the practice of giving intentional states to rationally explain actions. We theorists beg the question in assuming that the frog makes a mistake. We then ascribe intentional states to the frog that explain this (mistaken) action; we ascribe the belief that the beebee pellet was an EDIBLE BUG or FOOD or a FLY. Someone might interpret me to be arguing simply that this assumption should not drive our conclusions about content, but follow from them; we have equally good reasons for not treating the action as a mistake, and ascribing to the frog the (correct) belief that it was snapping at a SMALL MOVING DARK SPOT. Beginning with what we assume to be obvious cases of correct and incorrect application of a representation has obscured our view.

This is perhaps part of the picture. However, the problem also goes much deeper than this. My point here is that there is no fact about the content of the frog's internal state, independent of human theorizing about it. *Content is not an intrinsic feature of the world, but a feature of our explanations of the world.* The same goes for errors. The laws of nature do not make mistakes. Physical objects just do what they do, and cause what they cause. In fact, even declaring that the mechanism in question is a representation — that it is something whose function is to carry information about something else — is also dependent on our explanatory activities. From the perspective of a theorist trying to explain an event, there is a strong temptation to explain causal systems by appeal to functions. However even ascriptions of function like this depend on our explanatory activities and judgments and the norms of our scientific practices.

Valerie Hardcastle (2002) gives a compelling argument that functions are not naturalizable, in that they cannot be reduced to purely physical properties of a system, independent of human theorizing and explaining. Ascriptions of function, she argues, are like all other scientific observations in that they are dependent on a theory whose adherents assume background conditions, adopt explanatory goals, accept theoretical postulates and so on. Functions are properties that are important relative to some framework. If we shift frameworks, we shift what we take to be the function. Hardcastle gives the examples of the Morrow reflex and the palmomental reflex in infants. An infant exhibiting the Morrow reflex will swing her arms up and around when startled. This

[12] Thanks to Tim Schroeder for pointing out this interpretation of my position. Neander (1995, p. 131) points out that it has been a common assumption that the frog must represent, and must misrepresent a beebee; this has become almost a desideratum of any theory of content that it show that the frog misrepresents when it snaps at a beebee.

reflex historically enabled our long distant tree-dwelling ancestors' infants to grab a branch if they fell. Contemporary doctors interested in infants' development and health can use the Morrow reflex as a gauge of cortical development, since with increased myleination of the cortex the reflex disappears. Infants exhibiting the palmomental reflex will curl their top lips if their palm is stroked. There does not appear to be any selective advantage to this reflex which also disappears with cortical development; it appears to be an accident of wiring in very young infants. Thus from a teleological perspective the Morrow reflex has a function while the palmomental reflex has no function. However, from a medical perspective both of these reflexes have the function of indicating — and can misindicate — the level of cortical development. Hardcastle argues that the allegedly natural normativity in the system's function, what this reflex is supposed to do, depends on the interests of those examining it. She uses this example and others to argue that:

> Relative to each explanatory framework that uses the language of functions, we find a naturalistic notion of normativity. The notion varies as the particular notion of function does. But in each case the function of T is to do E in O because E is necessary for answering the question of what O is doing. (Hardcastle, 2002, p. 153)

There is no such thing as the function of a biological mechanism, *simpliciter*, independent of human judgments, questions and explanatory interests. This pragmatic approach to functions, she argues, grounds functions in the activities and practices of scientists who give good empirically and theoretically responsible reasons for ascribing such functions.

Hardcastle points out (p. 152) that some might interpret this as a relativistic license to ascribe just about any function at all to a system. Others might argue that this account does not distinguish functions from dispositions or accidents, and so is not really an account of functions at all. Hardcastle responds that the function of a system is not completely relativistic, because which function one should ascribe depends on criteria set by the discipline asking the questions. For each discipline, there are agreed-upon norms (Hardcastle [pp. 152–153] calls them "criteria") regulating what counts as an adequate explanation or an adequate empirical result. There are methodological criteria, theoretical precepts, background assumptions, measurement techniques and so forth that constitute a normative practice that grounds the activities of researchers and enables them to critique and to support one another's conclusions. Any scientific observation is made against such a normative background. Conclusions about the function of a mechanism are no different. Thus, the agreed-upon norms specifying defensible ways to describe a mechanism's function for evolutionary biologists may be quite different from those norms agreed-upon by anthropologists or doctors. But among scientists who agree-upon the norms regulating their particular discipline, there should be little disagreement about the way the function ascribed to the system should be described.

Hardcastle, notably, concludes that this pragmatic approach to functional norms in nature cannot extend to ethical norms, since scientists work within well-established frameworks while ethicists "are still struggling to fix the framework itself" (p. 154). A notion of normativity will be a naturalistic notion, for Hardcastle, only if we restrict the account to "functions that stem from scientific concerns of the world" (p. 153). Ethical norms, she concludes, do not abide by such a restriction and so are not naturalizable in this sense.

One of the central concerns of the account I present here is whether a cognitive science that rests upon the concept of representation, and thus on the concepts of semantics and intentionality, can be a naturalistic science. Thus, Hardcastle's defining naturalism in terms of science will not help us here, since this would make any science naturalistic by definition. And as Hardcastle reminds us (p. 147), science itself is a human activity, embedded in human explanatory activities, values, assumptions and practices. These all, of course, have human norms and human intentional states at their heart. So unless we can find a way to give a naturalistic account of human norms and human intentional states, no scientific account of functions can be truly naturalistic. As I have shown, we cannot use teleofunctions to naturalize the semantic content and intentionality of human intentional states. The barrier is the apparently vicious circularity in these explanations: we are trying to naturalistically explain the aboutness of intentional states. But it turns out that explanations which appeal to natural selection bestowing biological functions on mechanisms *also* appeal to human norms and human intentional states. Theorists who make judgments about the functions of biological mechanisms use the norms of their particular discipline to make judgments about the appropriate ways to describe such functions. We need to look deeper for a naturalistic account of semantics and intentionality, then.

Intentional Psychology is Normative, Not Just Descriptive

The above conclusions give us reason to switch explanatory directions. In answering the intentionality question and the semantic question by beginning with low-level biological states that supposedly have contents by virtue of their proper functions, and building up to more complex intentional states from there, we ran into a dead end, since the functions of low-level biological mechanisms depend on the intentional states of the people who ascribe functions with determinate contents to these mechanisms. For this reason, it seems to me that teleosemantics approaches the project from the wrong direction. A more promising approach is to begin with the social and normative practice of ascribing intentional states to people to explain their actions, and to give a rationale for ascribing semantic content and intentionality to biological mechanisms, and especially to neurological mechanisms, from within this practice. Furthermore, as I'll show in the following section, this approach is not as non-naturalistic a position as many might accuse it of being. I'll show that there is good reason to see such an approach to semantic content as compatible with naturalism. And interestingly, a different kind of appeal to evolution supports this optimism.

An approach that begins with the human practice of ascribing intentional states as reasons for action begins with relinquishing strong realism about representations. Many are reluctant to consider this approach for that reason (one of my aims here is to undermine such reluctance). There is a fairly heated debate about whether we should be realists about the intentional states we appeal to in giving explanations of people's actions. Realists about representation take it as a fact that some mechanisms *just are* contentful representations whose tokenings have intentionality; this is an intrinsic fact about some mechanisms and their physical relationships to objects. Searle, Fodor, Dretske, and Millikan take this route. Millikan, for example, argues

> If it really is the function of an inner representation to indicate its represented, clearly it is not just a natural sign, a sign that you or I looking on might interpret. It must be one that functions as a sign or representation *for the system itself.* (Millikan, 1989, p. 284, emphasis original)

I have argued above that we have reason to be more skeptical: nothing is *intrinsically* a representation. Describing particular brain mechanisms as representations with contents that apply correctly or not to their targets is part of *a human normative activity of explaining* certain movements of complex systems as purposeful actions and ascribing intentional states to such systems to explain and predict their movements. Daniel Dennett (1987, 1991) has argued for something like this position. Although I disagree with some aspects of Dennett's position (Cash, 2008), his "milder-than-mild realist" position about beliefs and other intentional states is a illustrative place to begin.[13] He argues that no creatures *intrinsically* possess intentional states. Intentional states are theoretical entities, attributed by observers to agents to explain and predict patterns in the agents' behavior. They are attributed by creatures like human beings; creatures able to adopt the intentional stance to one another and to themselves. In a world without observers able to adopt the intentional stance, however, there would be no intentionality.

[13] Dennett contrasts his own views with four exemplars of different extremes of realism about beliefs and other intentional states: Fodor's industrial strength Realism (with a capital "R"), Davidson's regular strength realism, Rorty's milder-than-mild irrealism and Paul Churchland's eliminative materialism (Dennett, 1991, p. 30).

To Dennett, intentional states are abstract variables an observer posits as intervening between stimulus and response, in order to explain a pattern in that creature's behavior. A claim, for example, that Joe has a particular intentional state, is not, to Dennett, a claim that is established as true by being reduced to properties of Joe's neurology and environment. Rather, its truth conditions are to be found in relationships between Joe's actions and environment, and the observer's (tacit or explicit) "theory"; a theory that specifies relationships between an entity's behavior, environmental conditions and intentional states. (These relationships, if made explicit, would be of the form "If an agent does A in conditions C then the agent believes X, desires Y and intends Z.") By attributing such reasons to the agent based on the agent's behavior, the theory enables that observer to make rational sense of that agent's behavior in that context.

The decision to adopt the intentional stance towards a creature and to attribute particular contentful intentional states rather than others, for Dennett, is up to the attributing individual alone. This decision is justified pragmatically, by virtue of the success of the behavioral predictions that such attributions facilitate. For example, Dennett (1991) argues that two individual observers could come up with "two different systems of belief attribution to an individual which differed *substantially* in what they attributed." He continues, arguing that:

> no deeper fact of the matter could establish that one was a description of the individual's *real* beliefs and the other not. In other words, there could be two different, but equally real patterns discernible in the noisy world. The rival theories would not even agree on which was pattern and which was noise, and yet nothing deeper would settle the issue. The choice of a pattern would indeed be up to the observer, a matter to be decided on *idiosyncratic pragmatic* grounds (1991, p. 49, final emphasis added).

Thus attributing particular beliefs, intentions and desires to a system (such as an animal, human being, or machine) is justified, to Dennett, if doing so gives an individual observer the pragmatic advantage of being able to successfully explain and predict the system's behavior.

The pragmatic advantage of the ability to explain and predict others' actions by ascribing intentional states to them, or the ability to "mindread" as it's often referred to, seems to have been of such selective importance that this ability is the genetic inheritance of almost all human beings, and is practically automatic for all adults.[14] All children, with the possible exception of autis-

[14] "Mindreading" here is a contrast with "behavior reading," where agents only notice connections between what happened to people and how they respond behaviorally. Mindreaders can attribute intentional states to explain reasons for others' behavior (Whiten, 1996, 1998).

tic children (Carruthers, 1996; Leslie, 1991), develop this ability somewhat automatically. The "graduating" achievement in the development of mindreading abilities is taken to be the child's ability to recognize that someone else's beliefs are different from their own (at this point they recognize others' beliefs *as beliefs* that can be incorrect), which occurs between three and a half to four years old. The almost inevitable development at the appropriate time of these abilities, leading to the eventual development of the ability to ascribe intentional states as reasons for actions, is evidence of a long history of selection within human cultures, for the ability to explain actions by ascribing reasons for acting. These abilities are the genetic inheritance of almost every human child.

Genes do not act alone, however. This ability to ascribe intentional states to others also depends on the child developing in a context where people treat one another as persons with intentional states. This development happens in a context in which children are socialized into an interactive, and especially a linguistic, culture in which intentional states are an important social currency. People interact with one another relying on others' ability to ascribe intentional states as reasons for the agent's actions. (An increasingly popular account of language [e.g., Cash, 2004] holds the production and interpretation of linguistic utterances to depend on the interpreter's ability to recognize the speaker's reasons for making the noises they make.)

This shared context of people ascribing intentional states to one another as reasons for actions, suggests that Dennett has missed (or at least significantly underplayed) an important dimension of the situation when he rests his account of intentional states on pragmatic justification of the observer's theory, by virtue of the successful predictions it enables that individual observer to make. As I'm about to show, the relationships between intentional states and behavior are not, as Dennett supposes, simply in the eye of an *individual beholder's theory*, who justifies the theory pragmatically. Recall the way Hardcastle anchors the ascriptions of a function to a mechanism in a scientific discipline's shared background knowledge and practices; these practices constitute normative criteria for the appropriateness of ascribing functions to a biological mechanism and for what functions it is appropriate to ascribe. Similarly, ascriptions of contentful intentional states to agents are based in the observer's *linguistic community's shared, normative criteria* for the appropriateness of ascribing intentional states to an agent, and for what contentful intentional states it is appropriate to ascribe.

The norms of the social, linguistic practice of ascribing contentful intentional states as reasons for actions spring from a fundamental aspect of human life, that we hold ourselves and one another responsible for our actions. This responsibility, to oneself and to others, brings out the more important side of the normativity that Dennett underplays. Predictions of the actions of another member of our linguistic community are not simply based on causal regularities (observable patterns, to Dennett) specifying what a rational agent with a particular intentional state *will* do. Rather — this is the point that Dennett seems to miss or downplay — such regularities are a side-effect of the fact that the theory relating actions to intentional states is shared and normatively enforced. The "patterns" we expect to see manifest when we ascribe intentional states to one another specify what a rational, responsible agent with those intentional states *should be interpreted as committed to doing.*

Many people have argued recently that the notions of mental content and of linguistic meaning are normative notions. While there is some debate about what kind of norms are involved (Cash, 2008), the central thesis — that when we talk of meanings and of the content of people's thoughts we are saying something normative — has much merit. This normative dimension of meaning and content was made explicit in Kripke's (1982) discussion of Wittgenstein's rule-following problem, which presents the normativity problem for content that I discussed earlier. Brandom (1994, 2000, 2001) further develops Kripke's thesis and its consequences. Drawing heavily on Wilfred Sellars and Immanuel Kant, Brandom presents a muscular argument for the thesis that ascribing an intentional state to another is ascribing a normative status; this status tracks the agent's commitments to act in such and so ways and tracks what the observer is entitled to expect from the agent in terms of what the agent does and says. This is part of the normative social practice of giving and asking for reasons for actions. I have not space to present the details of Brandom's argument here, but I intend to present enough of his conclusions to illustrate the plausibility of this perspective.

My purpose here is to show that if we conclude (as I have argued we should) that teleofunctional semantics will not provide a reductive naturalistic account of semantic content and intentionality, then we need a different approach to naturalizing semantics and intentionality. An approach that tries to reduce semantic content to a functional property won't succeed because it presumes human normative explanatory practices. But if Brandom is right that intentional states are normative statuses, ascribed as part of the normative practice of giving reasons for actions, then intentional states do not get their content from the functions of the neurological states that implement them or that they supervene upon. Rather, intentional states are ascribed to people as part of a normative practice, and thus the contents of any neurological states that these intentional states supervene upon or are implemented by will also be derived from this social normative practice.

Brandom draws from Kant the insight that human judgments and human actions have a feature in common: they are both things we are *responsible* for. (This responsibility, to Kant, is what distinguishes our actions from those of animals.) We are responsible for our judgments and actions in the sense that we are responsible for giving *reasons* for them. Importantly, for Brandom, these reasons for actions and for judgments are constituted by (often tacit) normative practices governing the inferential use of the concepts we employ in giving these reasons. Brandom thus moves past Dennett's pragmatic justification for adopting a "theory" of the relations between behavior and intentional states, instead seeing the theory as having the same normative basis as that which undergirds a community's shared rules for the appropriate use of linguistic expressions and that which undergirds a scientific community's shared criteria for explanatory adequacy. The "theory" the observer uses to ascribe intentional states to others, Brandom argues, is a set of norms *shared* by a linguistic community; one that includes both agent and observers.

According to the norms of these interpretive practices, particular actions (or at least the intention to perform them) *should* flow from being in a particular intentional state:

> To say this [that a person has a particular reason for acting] is not yet to say that the one who has such a reason *will* act according to it, even in the absence of competing reasons for incompatible courses of action. What follows immediately from the attribution of intentional states that amount to a reason for action is just that (ceteris paribus) the individual who has that reason *ought* to act in a certain way. This "ought" is a *rational* ought — someone with those beliefs and desires is rationally obliged or committed to act in a certain way. (Brandom, 1994, p. 56)

To Brandom, these obligations and commitments arise from the general social injunction that *one ought to act rationally* (which includes the injunction to explain others' actions rationally). These obligations and commitments also arise from the fact that the norms instituting rules of rational inference from actions to intentional states and from intentional states to actions are shared by both agents and observers.

These norms constitute rules for inference from a description of what someone does to an ascription of intentional states that would count as good reasons for doing that. And since these norms are mutually known, an agent who acts that way knows they have *entitled* observers to infer that they have these intentional states. These norms also constitute rules for inference from intentional states ascribed as reasons for their actions, to further actions expectable of someone with those intentional states. Agents who have entitled observers to ascribe these intentional states to them thus also undertake a *commitment* to perform the kinds of actions that inferentially follow from those intentional states. For example, imagine that while I am aware that you are observing me, I act in a way that licenses you to infer that I want the light on and that I believe that my flipping the switch will turn the light on. (I could do this very explicitly, for instance, by telling you that I want and believe this; though many non-linguistic actions could also, less explicitly, license you to attribute these intentional states to me.) Imagine also that I am very close to the light switch, such that you recognize that it is easy for me to flip the switch myself rather than asking or expecting anyone else to do it, and I am aware that you recognize this fact. By having acted this way, while aware that you are observing me, I have *entitled* you to ascribe to me this belief and desire. I have also thereby entitled you to expect me to have the intention to flip the switch that inferentially follows from these intentional states. According to the norms of our shared practice of giving and asking for reasons for actions, you *should* ascribe to me the intention to flip the switch (rather than any other intentional state). It would be irrational for you to ascribe any other intentional state to me. Likewise, especially since I know that you should ascribe this intention to me, it would be irrational of me to not (try to) flip the switch. By acting this way, I have *committed* myself to flipping the switch, at least to the extent that if I fail to (try to) flip the switch I should recognize that my behavior entitled you to expect that I would do it, and that it would not be out of place for you to ask me to explain why I did not do it (i.e., to ask me to make my competing reasons explicit).

Note that the agent and observer can be the same person. My own conception of myself as a rational responsible agent is to some degree dependent on my own expectations of how I should conduct myself, according to the standards of rational behavior endorsed by the communities with which I identify and which socialized me. I ask myself about my reasons for my actions and ascribe to myself reasons for my actions.

This gives a very Kantian perspective on people's actions: we do what we ought to do, and we do it *because* we recognize that this is what we ought to do. The force of the "ought" here comes from the general injunction to be rational, or at least to be seen by other people as a rational agent, as someone who lives up to their commitments. And the content of what we ought to do comes from our previous actions, from the linguistic community's norms that legislate: (a) the intentional states others ought to ascribe to us as reasons for those actions, and (b) the actions that someone with those intentional states rationally ought to perform in the current situation.

On this normative view, then, the "theory" of the relation of intentional states to actions that Dennett argues we use to predict actions thus turns out to be less like an individual's scientific hypothesis, and more like a community's moral code. It enables prediction of the actions of other members of our linguistic community, because those others willingly conform to its strictures, holding themselves responsible for a certain degree of consistency among their actions, especially when observed by others. People willingly conform for reasons similar to the reason they willingly conform to the norms and conventions of their language. We are raised in this normative context, and are innately disposed to conform to the practices of the community that raises us.

As we get older, conformity to these norms becomes more and more important for us; conforming is as important as being understood, being trusted to be responsible enough to make our own decisions, being seen as "rational," and other social goals that are (in part) achieved by living up to one's commitments and obligations. We also conform, importantly, because of sanctions that others may apply when we fail to behave as one ought to behave. We are socialized into conformity, through sanction as well as through such rewards as being trusted. Dennett appears to neglect the way these obligations and commitments arise from a shared practice of holding one another responsible for our actions, and of the role reasons play in that practice. What Brandom (2000, p. 81) calls "the social, implicitly normative game of offering and assessing, producing and consuming, reasons" is a *shared* normative practice of ascribing intentional states as a means of keeping track of what we expect one another to be *committed* to doing and saying.

From this point of view, a person's movement only gets to be an "action" by being the kind of event for which explanations that appeal to reasons are appropriate. Asking for reasons gives way to asking for causes at about the same point that holding people responsible for actions gives way to treating what happened as a mere bodily process for which one need not be held responsible (we do not generally ask for reasons for sneezes). The kinds of events for which someone can be held responsible are the kinds of events that this normative system of giving reasons for actions categorizes as "actions." For bodily movements (such as sneezes, hiccups, going unconscious when struck on the head, shivering when cold, etc.) we do not ask for reasons, but instead ask for causal explanations, since we do not view these as voluntary actions for which the person could be held responsible. This is another way of saying that being an agent who performs actions and who has intentional states is a normatively constituted status. It is by virtue of participation in the practice of giving and asking for reasons for actions that one qualifies as an agent, and as having intentional states.

The status of being an "action" and the status of having intentional states are very similar to any other normatively constituted status. For instance, no event is "intrinsically" a murder. An event only gets to be a murder by satisfying criteria set out by a community's normative ethical and legal system. Similarly, if the practice of giving and asking for reasons is a normative practice, then no movement is intrinsically an action (as opposed to a bodily movement), and nothing intrinsically has intentional states. Something qualifies as having intentional states by virtue of movements it makes being constituted as actions for which intentional states are appropriately ascribed as reasons.

An interesting implication of this normative approach to intentional states is that someone might deny that the case of the frog that snaps at a beebee counts as a genuine case of misrepresentation, since frogs are incapable of participating in this interpersonal normative linguistic practice. Creatures like frogs cannot be expected to conform to the norms of the practice of giving reasons for actions, and to live up to the commitments to future behavior entailed by their previous behavior. On the account I have offered, can we say that frogs — or human infants, for that matter — have intentional states at all?

One response to this question could be to take these as reasons to see the frog as not representing at all, and so not to see a frog as misrepresenting when it snaps at a beebee pellet. Perhaps we should call the frog's snapping a simple bodily movement; a reflex of some kind we can explain in purely causal terms without the need for invoking intentional states as explanations. This example, then, would not be a case of a mechanism with semantic content, and thus not a genuine case of misrepresentation. On this view, we should only look for contentful representations (and so genuine misrepresentation) in cases in which both ascriber and ascribee of intentional states are fully accredited members of a normative linguistic community.

A more reasonable response to this question, however, is to distinguish between the way non-linguistic creatures like frogs simply *have* intentional states, in the sense that observers can reasonably ascribe them, and the way that fully accredited members of a linguistic community can consciously and reflectively *self ascribe* intentional states, qua intentional states. The reflective ability to ascribe intentional states to *oneself* as well as to others is an important aspect of the development of the kind of mindreading skills I mentioned earlier, which human children gradually develop over their first few years of life. Non-linguistic creatures like frogs and infants lack this prerequisite for participating in the practice of giving and asking for reasons, and for understanding entitlements and undertaking commitments of the kind instituted in this practice. But such creatures might be said to "have" beliefs desires and intentions, in the passive sense that an observer can ascribe intentional states to them. Having the second-order ability to ascribe a belief reflectively to oneself — and to understand it *as a belief* and hold oneself to the commitments entailed — is uniquely the domain of socialized, language-using human beings who have developed the ability to mindread and who have been socialized into their linguistic community's practice of giving reasons for actions. Frogs, human infants, and other non-linguistic creatures could sensibly be seen as intentional *patients* but not intentional *agents* (adapting Regan's [1980] distinction between moral patients, who deserve moral consideration but who cannot accord such consideration to others, and moral agents, who can do so). We humans, we seekers of explanations, could have explanatory reasons for — somewhat generously and metaphorically — treating "simpler" creatures like frogs as *agents performing actions*, and ascribing intentional states to such creatures, based on the similarity of form their movements have to the kinds of human actions for which we should ascribe intentional states as reasons. But when doing so, we should understand that the parallels are relatively loose between fully accredited intentional agents who act on commitments to act, and intentional patients which may be *described as* acting for reasons, but which are incapable of acting *for* reasons.

Using Evolution to Naturalize Normativity

If all talk of semantic content is inextricably situated within the normative practice of ascribing contentful intentional states as reasons for actions, what hopes are there, then, for incorporating intentionality into a naturalized cognitive science? Can we appeal "honestly" to the allegedly representational powers of the brain when explaining human cognitive abilities? I think there's hope aplenty. But realizing that hope requires us to let go of a rather strongly held, but misguided, conception of what is required to "naturalize" a phenomenon.

A concern to avoid viciously circular arguments has led many philosophers of mind to feel logically forced into the kinds of reductive approaches to naturalizing intentionality and semantic content discussed earlier. As the quote from Fodor that I presented earlier attests, many philosophers hold that such reductive approaches are the only respectable game in town. Barbara Von Eckardt (1993, p. 206) also makes this point explicitly, arguing that "a conventional ground [for a neurological representation's content] is ruled out at the outset because of cognitive science's commitment to naturalism." Conventions (of which norms are a subset) are not a naturalizable ground, for Von Eckardt, because conventions are supposed to depend on the intentional states of agents who follow them, and we can't explain intentionality by appeal to something that depends upon agents' intentional states. Von Eckardt adds that because of this constraint, reductive accounts are "at the heart of most of the current approaches to the content-determination question" (1993, p. 206).

It does *seem* that the explanation of intentionality I have offered exhibits the kind of circularity that drives such philosophers to reductionism. I have suggested here that a person's intentional states are a socially and normatively constituted status, a social currency we humans ascribe to keep track of what we each are committed to doing and saying. If this is so, then these intentional states depend on a preexisting practice of ascribing intentional states into which we were socialized as we grew from infancy, as well as depending on the intentional states of other participants in this practice. Furthermore, the participants in that preexisting practice were themselves socialized into their elders' practice of ascribing intentional states, and so on and so on. The attempt to explain intentional states by appeal to a practice that itself depends upon intentional states, however, only *appears* to be viciously circular. This appearance can be dispelled, and thus need not be a barrier to seeing intentionality (normatively constituted) as a natural phenomenon.

The intentionality question and the semantic question cannot be answered by attempting to explain intentionality away, by *reducing* it to scientifically respectable naturalistic properties of the purported representation and of its history. If having intentionality is having a normatively-conferred status, then *naturalistically justifying* the claim that a neurological state or process inside a person has a determinate semantic content would require *naturalistically justifying* particular norms. This, I have argued here, is futile. Hume (1739/2000) is right to conclude that one cannot derive an "ought" from an "is" like this; in order to have normative conclusions, one needs at least some normative premises. Attempts to naturalistically justify the normative distinction between correct representation and misrepresentation such as Dretske's and Millikan's teleofunctional accounts, commit what G.E. Moore (1903) calls the "naturalistic fallacy," by trying to reduce normative assessments to naturalistic properties. Such attempts, I have argued here, fail to be purely reductionist. Ascriptions of function to biological systems either fail to be determinate enough to misrepresent, or tacitly depend for a specific content upon the normative practice of giving reasons for actions and on the norms of the scientific practices whose participants identify functions of neurological mechanisms. Premises regarding what events count as actions (for which intentional states rather than causes are appropriate explanations) come from this practice. These practices also supply the normative premises that identify some of those actions as "mistakes" for which mistaken intentional states can be given as reasons. This normativity cannot be reduced away. In addition to these problems with assigning functions to naturally evolved systems, there is the further problem that human intentionality is inextricably normative. It is constituted by a normative practice, whose norms regarding entitlements to ascribe intentional states and the commitments of one to whom such ascriptions are appropriately made are not reducible to any natural facts.

To incorporate representational explanations into a naturalistic cognitive science, then, we need a better conception of what is required for a phenomenon to be given a naturalistic explanation; one that does not assume that naturalism requires reduction to "something else" as Fodor argues. Rather than attempting to give a *naturalistic justification for particular norms* by trying to reduce these norms to physical properties, we can give a *naturalistic nonreductive account of normativity in general.* Such an account will not justify any particular norms, but will naturalistically account for the existence of norm-governed practices (whatever their norms happen to be), such as the practice of giving contentful intentional states as reasons for actions.

This is where an appeal to evolution by natural selection can help. There appears to be a vicious circularity in explaining people's intentional states by appeal to the norms of the practice of ascribing intentional states, which themselves depend upon the intentional states of people who participate in this normative practice. But this apparently vicious circularity can be discharged, by showing it to be a non-vicious recursive circularity. It can be discharged by giving a naturalistic evolutionary account of the transition from a world where there were no human beings, and no normative practices, and thus no practice of ascribing intentional states as reasons for actions, to a world where such normative practices are commonplace.

It's useful to pause and reflect on why we feel we need a naturalized account of semantics and intentionality. Descartes held that the mind was a non-physical substance, immune from explanation by science. Brentano (1874/1973) added to the problem by declaring intentionality the characteristic that distinguishes mental from physical phenomena. Philosophers of mind have been trying to recover from this position, to show how we can explain human cognitive abilities, including intentional states, without assuming that mental phenomena are different in kind from physical phenomena, and without appealing to *supernatural* entities like Cartesian minds. An evolutionary approach can show that the practice of giving and asking for intentional states as reasons for actions has a natural origin, as natural as the origin of human beings, the origin of mammals, or of multi-cellular organisms. Nothing supernatural is involved in offering such explanations.

This evolutionary approach to naturalizing normativity would work for all forms of human normativity; it could show a natural origin for linguistic norms, for ethical norms, for the norms of scientific practices, and for the norms of the practice of giving intentional states as reasons for actions. With linguistic norms, for instance, we cannot naturalistically *justify* the particular norms for the appropriate usage of expressions that a linguistic community shares and follows, as "better" than some other possible linguistic norms. This account might naturalistically explain how the norms came to be the norms that they are, but it will not justify these norms as the best norms to have. For example, there is no naturalistic justification that can be given for why the English symbol "dog" (rather than some other noise or mark) is appropriately used to refer to canine mammals that people keep as pets. Nor can any naturalistic justification be given for why English speakers should not use gendered nouns yet French and Spanish speakers should. These norms are somewhat arbitrary; arrived at through historical accidents. The particular norms of a particular language cannot be justified naturalistically. However, we can give a nonreductive evolutionary account of how humans evolved to become the kinds of creatures that congregate in societies whose members share some linguistic norms *rather than not having a language at all* (Fitch, 2005).

We also cannot naturalistically justify one set of social, etiquette, religious and ethical norms as better than any other set of possible norms, since our criteria for "better" will themselves incorporate normative assumptions (for example, assumptions about the "intrinsic good" of sociability, piety, social harmony, trustworthiness, happiness, respect for persons, life, certain forms of equality, and so on). However, we can give a nonreductive evolutionary account of how humans came to be the kinds of creatures that congregate in societies whose members share and enforce social, etiquette, religious, and ethical norms, *rather than having no such norms at all.*[15]

Similarly, we cannot naturalistically justify any particular set of the potentially infinite sets of possible inferential norms relating actions to intentional states and relating intentional states to actions. We cannot explain why we engage in one practice of giving and asking for reasons, rather than a different possible normative practice. But it is possible to give a nonreductive evolutionary account of how humans came to be the kinds of creatures that are able to treat movements as actions for which agents can be held responsible; creatures that are able to evaluate the propriety and rationality of patterns of actions, making sense of such patterns of actions by ascribing intentional states to one another (and to themselves) as reasons for actions, *rather than being the kinds of creatures that do not ascribe intentional states to one another at all.*

For any community's norms, it is at least theoretically possible to give a nonreductive naturalistic historical account of how they came to have the particular norms they have, whatever those norms happened to be. While this account will not justify those norms as better than any other possible set of norms that community might otherwise have adopted, it would show human normative practices to be natural phenomena, in the sense of "natural" that counts; that which contrasts with "supernatural." There is no need for such phenomena to be natural in the sense of "natural" that contrasts with "cultural."

This claim that explicit human normativity can be given a naturalistic, evolutionary explanation I take to be rather plausible. It does not seem to need much in the way of support. The details of *how* the human cognitive abilities and cultural institutions that support such normative practices evolved (in contrast with the claim *that* they evolved) are all there is to dispute here.[16]

[15] On the evolutionary advantages of having ethical norms, see Campbell (1986), Gibbard (1990), Rosenberg (1990), Ruse (1986, 1990), and many of the papers and replies in Katz (2000). Most agree that particular ethical norms cannot be given any evolutionary justification. However, as Ruse (1990, p. 65) argues, that while no objective basis can be given for ethics, an argument can be made that having some ethics rather than none affords a selective advantage to a group. Campbell (1986, p. 24), adopts a similar tactic, arguing that "having some morality rather than none is justified for every member of the group if having some morality rather than none overwhelmingly improves the life prospects of everyone in the group."

[16] I refer the reader interested in the details of such an evolutionary account to Daniel Dennett's (2003) *Freedom Evolves* and to his (1995) *Darwin's Dangerous Idea.* Here Dennett tells a more individualistic story than I would, but connects many (not all) of the dots in an account of how creatures evolve normative practices like ours. Haugeland (1990) also includes a section (pp. 147 ff.) explaining how explicit norm-consulting practices like our ethical, linguistic and reason-giving practices could evolve out the kind of tacit norm-constrained practices we can observe nowadays in many groups of animals.

Because of this, I'll only give a very brief sketch in extremely broad brushstrokes.

The basic idea is that human forms of *explicit* linguistic, ethical, and reason-giving normativity depend on pre-existing cognitive abilities and cultural institutions, which depend upon the kinds of *tacit* norm-constrained but not norm-consulting behaviors we share with other animals. It's not explicit norm following "all the way down." Robert Brandom (1994, pp. 18–30) explains that human normativity is partially a norm-consulting practice: one of doing what one should do because we explicitly know that it's what we should do. However, such explicit norm-consulting practices are not all there is to human normativity (if it were, then we would have an infinite regress problem). Such explicit norm-consulting practices depend upon tacit norm-governed (but not norm-consulting) practices, in which we simply take or treat certain performances as correct or appropriate. Brandom uses this as the solution to the rule-following problem Kripke attributes to Wittgenstein. Explicit norm-consulting depends upon a background of tacit agreement about the kinds of performances that are acceptable. This agreement is enacted by participants in the practice tacitly agreeing that certain performances are legitimate, by simply treating those performances as acceptable or not acceptable. This applies especially to judgments about whether a rule has been followed or not.

We can apply the same considerations to the evolution of explicit normativity. The abilities to ascribe intentional states to one another and to use language (in part, to make our norms explicit) evolved from tacit forms of normativity. Such tacit forms of normativity can be seen in horse herds, chimpanzee troops, dog packs and other social groups of animals, in which behavioral conditioning ensures that young members of a group come to behave as others in their group behave. Members of such groups are able, for instance, to sanction others for doing things differently from how one ought to do them. Such herd animals do not need an explicit knowledge of the content of the rules, but simply the tacit ability to recognize and respond to "unacceptable" behavior. (Roberts [1997] apparently trains horses by using the norms he observes in wild horse herds, such as sanctioning a young foal who has misbehaved in the same way a matriarch horse would, by isolating him from the herd until he adopts a posture of contrition.) Furthermore, groups whose members share such tacit normative behavior evolved from ancestors whose members congregated in groups but did not even have the behavioral plasticity required for the behavioral conditioning needed to get a creature to make its behavior conform to the patterns of behavior of its conspecifics.

Tacit norm-constrained — but not norm-following — behavior simply requires a group with a certain amount of what Haugeland (1990) calls "conformism" in its members: the tendency to imitate patterns of behavior observed in other members of the group and a certain degree of sanctioning as well — the "positive tendency to see that one's neighbors do likewise, and to suppress variation" (p. 147). If group members have enough behavioral plasticity to learn from experience, then elder conspecifics can sanction certain performances; that is, they can respond to certain kinds of performances in ways that make it less likely that that kind of behavior will be repeated. If this is so, then patterns of behavior will emerge in the group. Boyd and Richerson (1992), furthermore, point out that such sanctioning behavior is also something that can be approved of or sanctioned. Their evolutionary computer model shows that if a group's members employ what Boyd and Richerson call a "moralistic" strategy, in which individuals who refrained from sanctioning others when they should have sanctioned them are themselves liable to be sanctioned, then *any* pattern of behavior can become evolutionarily stable in the group, even if it is individually costly and also confers no advantage to the group.

This conformism, combined with evolutionary processes selecting against groups which happen to have less efficient or effective practices than others, can produce groups whose practices give the group as a whole an advantage. The practices of different groups, and natural selection between separate groups with different practices that enable the group to prosper to varied degrees, are often cited as a source of evolutionary group selection (Sober, 1984, 1992; Wilson, 1997; Wilson and Sober, 1994). Groups of interacting individuals whose members share a common way of interacting, Wilson and Sober (1994) argue, can prosper because their style of interaction gives the group as a whole a competitive advantage over groups with different styles of interaction.

Gesturing very broadly here, it's possible to see in such a situation the conditions required (though I do not pretend to have described all the sufficient conditions) for the evolution of language and more explicit forms of normativity, including the practice of ascribing reasons for actions. As the group's tacit norms become more and more complex — imagine a group with norms approaching the complex social and political situation of modern wild chimpanzee troops, with alliances and patterns of dominance and practices of retribution — we can see a situation in which having more efficient ways of predicting how particular others might react to one's actions becomes an advantage. The ability to observe patterns in others' behavior and to privately apply one's own labels to simplify the process of keeping track of what one can expect them to do would be a serious advantage to an individual.[17] And in a group where many individuals have this skill, the practice of doing something

[17] See Dennett's *Kinds of Minds* (1996, pp. 124–125), in which he uses Andrew Whiten's (1993, p. 385 ff., 1996, p. 283 ff.) examples of the cognitive economy gained in ascribing intentional states as abstract intervening variables between others' observed actions and the responses one can thus expect from them to one's actions. It turns a problem of remembering *n* times *m* associations, in to one that requires *n* plus *m* associations.

knowing that observers will take this action to be part of a pattern and use that pattern to predict further actions could be an advantage. It could help both in being predictable and so eliciting cooperative teamwork, and in manipulating others' predictable reactions to one's actions to elicit responses that are to one's advantage (Krebs and Dawkins, 1984).

In a social group in which this skill is widespread, it could also be an advantage to be able to *share* these observed patterns of behavior, and perhaps to share public labels by which the patterns can be labeled and reidentified. The states of awareness or ignorance and the intentions and desires ascribed to others can be given common public labels. These labels can be used, for example, to avoid punishment by making clear the difference between what one did and what one was trying to do. This could be useful in giving excuses, especially in cases when good intentions might make a difference to potential sanctioning reactions. The correct use of such labels could be taught to younger members of the community as they learn the community's norms. In such a context, we could see the beginnings of human-style language, which depends upon the practice of ascribing intentional states as reasons for action.

I cannot give more than these very broad gestures at the way human explicit mindreading practices may have co-evolved along with human language. But even if the details remain somewhat incomplete and controversial, the fact remains that it is possible to give some kind of evolutionary account like this; an account of how explicit human normativity evolved out of the kinds of tacit normativity we find in contemporary primate social groups, which itself evolved due to the existence of individuals whose behavior could be influenced by how other members of their group behaved. If this kind of evolutionary account can be given, then we would have support for the thesis that normative practices are naturalistically respectable entities, in the sense of "naturalistic" that contrasts with "supernatural." We could thus rebut the charge of vicious circularity against appeals to a normative basis for an explanation of intentionality and semantic content. The account is circular, but not viciously so.

Thus explanations that appeal to neurological states and processes being representations with intentionality and semantic content can be supported as naturalistic in the sense discussed above. However, this means that any neurological mechanism which cognitive scientists identify as causally responsible for particular actions will not be a representation "intrinsically." Semantic content will be a normatively ascribed status *ascribed* to neurological mechanisms as part of our attempts to integrate the causal explanations emerging from cognitive science with our normative practice of ascribing reasons for actions to whole persons (Cash, in press). The possibility of identifying a (e.g., neurological) causal mechanism that we might want to call a representation with content depends upon first normatively individuating a set of causal happenings as actions that have a common intentional state as a reason. Only after first normatively individuating such a set of situations in which a common intentional state may be ascribed to a group of subjects as a reason for their actions, could cognitive scientists then look for a neural state or process or mechanism causally related to all and only these actions. In such a situation, it would be reasonable to *extend* the norms of the practice of ascribing intentional states to whole persons as reasons for their actions, such that once cognitive scientists identified such a sub-personal mechanism it would be appropriate to ascribe semantic content to it; a content derived from that normatively ascribed to the agent.

Compare the claim that a particular knife is a murder weapon. It is only from a normative perspective that individuates an event as a murder that the knife can qualify as a murder weapon. Scientists could investigate the causal details of the event, but it would be a causal account of a normatively individuated event. A forensic scientist's proof that "exhibit A" is the murder weapon would be a scientific proof of the causal role of the knife in an event that qualifies as a murder by meeting criteria set by a normative ethical and legal system. It would be a mistake to claim that the knife was intrinsically a murder weapon, if "intrinsically" entailed independence from human beings' activities of normatively individuating events as "murders." Similarly, it would be a category mistake to say that the mechanisms that cognitive scientists will eventually discover to be causally responsible for actions are themselves *intrinsically* or naturally representations, and *objectively* have a particular content. These will be, like murder weapons, mechanisms that play a causal role in a normatively individuated event. Thus cognitive scientists might find explanatory reasons for *considering* particular types of neurological mechanisms or activities to be representations that have contents, whose tokenings have external objects as intentional targets. But a tokening of such a mechanism can only qualify as having a determinate semantic content and a particular intentional target, by playing a causal role in an event (a bodily movement) that is identified normatively as an action for which intentional states with that target and that content should be ascribed as reasons.

Thus explanations of cognitive abilities that appeal to people's ability to represent determinate contents could play a respectable role in cognitive science. And explanations of these abilities by appeal to particular neurological states and processes that support them would also have a respectable place in cognitive science. And in such a context, it may well make sense to ascribe particular semantic contents to such neurological states and processes, and to ascribe particular intentional targets to tokenings of such states and processes. However, all of this will be done from within the human normative linguistic practice of identifying certain movements as actions and ascribing contentful intentional states as reasons for those actions.

Conclusions

I have shown here that there are two possible tactics in using evolution to "naturalize" intentionality: a reductive form and a non-reductive form. The more popular reductive strategy employed by teleofunctionalists such as Millikan and Dretske does not succeed in reducing the normativity away to the kind of "something else" Fodor (1987, p. 97) insists upon. While attempting to justify claims that a neurological state or process has a determinate enough content to distinguish correct from incorrect representation, I have argued, these attempts to reduce the normativity of semantic content to the proper functioning of biological mechanisms depend upon the norms defining scientific disciplines and upon the normative practice of giving and asking for reasons.

However, on the approach I have sketched here, one can be a realist about intentionality and semantic content without having to be a reductionist, in spite of Fodor's claims to the contrary. One can be a realist about content in the same way one can be a realist about murders, promises, and runs scored in baseball games. These all exist because they are constituted within shared human normative practices. In a similar way, one can be a realist about human actions and about reasons for actions, and about neurological representations with determinate semantic contents, whose tokenings can have intentional targets. These are constituted within the human practice of giving reasons for actions. A representation's semantic content would be derived from the content of the reasons ascribed to the agent, and the intentional target of that representation's tokening on a particular occasion will likewise be derived from the intentional target ascribed to the person's intentional state.

One can likewise be a naturalist about such normatively constituted semantic contents by accepting that although this content depends upon this normative practice of giving reasons for actions, this practice itself has a naturalistic evolutionary explanation. Nothing supernatural or mysterious is involved here; nothing is in need of "naturalizing" before an explanation that referred to this mechanism as a representation could be accepted as a respectable element of a cognitive science explanation. The charge of vicious circularity, in explaining intentionality by appeal to a normative practice that depends upon the intentional states of the norm-following agents that ascribe functions and/or contentful intentional states, can be dispelled. A naturalistic account of the evolution of explicit norm-following practices can dispel this apparently vicious circularity by showing it to be the same kind of non vicious, recursive, bootstrapping cyclic process that produced all evolved creatures.

The theorists engaged in debates about the intentionality question and the semantic question, should not be seen, as they often see themselves, as engaged in *ontological* debates about whether there really are representations in the brain, and about what the semantic content of a particular representation really is. We should instead see them as engaged in *legislative* debates about the merits of extending the normative practice of ascribing intentional states as reason for their actions. They can better be seen as debating *principled* reasons to ascribe a particular semantic content to a mechanism within a person that is causally related to the person's actions, and for ascribing an intentional target to particular tokenings of that mechanism. Perhaps our interpreting mechanisms according to what we take to be their evolved biological functions will be an important consideration in such ascriptions of content, for instance. We should not forget, however, that the answers to these questions will be derived from the human normative practice of giving intentional states as reasons for actions; a practice with a respectable natural history.

References

Boyd, R., and Richerson, P.J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology*, *13*, 171–195.

Brandom, R. (1994). *Making it explicit: Reasoning, representing and discursive commitment.* Cambridge, Massachusetts: Harvard University Press.

Brandom, R. (2000). *Articulating reasons: An introduction to inferentialism.* Cambridge, Massachusetts: Harvard University Press.

Brandom, R. (2001). Modality, normativity and intentionality. *Philosophy and Phenomenological Research*, *63*(3), 587–609.

Brentano, F. (1973). The distinction between mental and physical phenomena [A. Rancurello, D.B. Terrell and L. McAlister, Trans.]. In *Psychology from an empirical standpoint.* New York: Humanities Press. (Original work published 1874)

Campbell, R. (1986). Can biology make ethics objective? *Biology and Philosophy*, *11*, 21–31.

Carruthers, P. (1996). Autism as mind-blindness: An elaboration and defense. In P. Carruthers and P.K. Smith (Eds.), *Theories of theories of mind* (pp. 257–273). Cambridge: Cambridge University Press.

Cash, M. (2004). Unconventional utterances? Davidson's rejection of conventions in language use. *Protosociology*, *20*, 261–295.

Cash, M. (2008). Thoughts and oughts. *Philosophical Explorations*, *11*(2), 93–119.

Cash, M. (in press). Normativity is the mother of intention: Wittgenstein, normative practices, and neurological representations. *New Ideas in Psychology.*

Cummins, R. (1996). *Representations, targets and attitudes.* Cambridge, Massachusetts: MIT Press.

Dawson, M.R.W. (1998). *Understanding cognitive science.* Malden, Massachusetts: Blackwell.

Dennett, D. (1987). *The intentional stance.* Cambridge, Massachusetts: MIT Press.

Dennett, D. (1991). Real patterns. *Journal of Philosophy*, *88*(1), 27–51.

Dennett, D. (1995). *Darwin's dangerous idea.* New York: Simon and Schuster.

Dennett, D. (1996). *Kinds of minds.* New York: Basic Books.

Dennett, D. (2003). *Freedom evolves.* New York: Viking Penguin.

Dretske, F. (1988). *Explaining behavior: Reasons in a world of causes.* Cambridge, Massachusetts: MIT Press.

Dretske, F. (1994). If you can't make one, you don't know how it works. In P.A. French, T.E. Uehling, and H.K. Wettstein (Eds.), *Midwest studies in philosophy XIX — philosophical naturalism* (pp. 468–482). Notre Dame, Indiana: University of Notre Dame Press.

Fitch, W.T. (2005). The evolution of language: A comparative review. *Biology and Philosophy*, *20*(2–3), 193–203.

Fodor, J. (1987). Meaning and the world order. In *Psychosemantics* (pp. 97–133). Cambridge, Massachusetts: MIT Press.

Fodor, J. (1990). A theory of content I: The problem. In *A theory of content and other essays* (pp. 51–88). Cambridge, Massachusetts: MIT Press.

Gibbard, A. (1990). *Wise choices, apt feelings: A theory of normative judgment.* Cambridge, Massachusetts: Harvard University Press.

Hardcastle, V.G. (2002). On the normativity of functions. In A. Ariew, R. Cummins, and M. Perlman (Eds.), *Functions: New essays in the philosophy of psychology and biology* (pp. 144–156). Oxford: Oxford University Press.

Haugeland, J. (1990). The intentionality all-stars. *Philosophical Perspectives*, 4, 383–427.

Hume, D. (2000). *A treatise of human nature, being an attempt to introduce the experimental method of reasoning into moral subjects.* Oxford: Oxford University Press. (Original work published 1739)

Katz, L.D. (Ed.). (2000). *Evolutionary origins of morality.* Bowling Green, Ohio: Imprint Academic.

Krebs, J.R., and Dawkins, R. (1984). Animal signals: Mind reading and manipulation. In J.R. Krebs and N.B. Davies (Eds.), *Behavioural ecology: An evolutionary approach* (second edition, pp. 380–401). Oxford: Blackwell.

Kripke, S. (1982). *Wittgenstein on rules and private language.* Cambridge, Massachusetts: Harvard University Press.

Leslie, A. (1991). Theory of mind impairment in autism: Evidence for a modular mechanism of development? In A. Whiten (Ed.), *Natural theories of mind* (pp. 63–78). Oxford: Basil Blackwell.

McGinn, C. (1989). *Mental content.* New York: Basil Blackwell.

Millikan, R.G. (1984). *Language, thought, and other biological categories: New foundations for realism.* Cambridge, Massachusetts: MIT Press.

Millikan, R.G. (1986). Thoughts without laws: Cognitive science without content. *Philosophical Review*, 95, 47–80.

Millikan, R.G. (1989). Biosemantics. *Journal of Philosophy*, 86(6), 281–297.

Millikan, R.G. (1993a). Compare and contrast Dretske, Fodor and Millikan on teleosemantics. In R.G. Millikan (Ed.), *White queen psychology and other essays for Alice* (pp. 123–133). Cambridge, Massachusetts: MIT Press.

Millikan, R.G. (1993b). Truth rules, hoverflies, and the Kripke–Wittgenstein paradox. In R.G. Millikan (Ed.), *White queen psychology and other essays for Alice* (pp. 211–239). Cambridge, Massachusetts: MIT Press.

Millikan, R.G. (Ed.). (1993c). *White queen psychology and other essays for Alice.* Cambridge, Massachusetts: MIT Press.

Millikan, R.G. (1998). Proper function and convention in speech acts. In L.E. Hahn (Ed.), *The philosophy of P.F. Strawson* (pp. 25–43). Chicago: Open Court.

Millikan, R.G. (2002). Biofunctions: Two paradigms. In A. Ariew, R. Cummins, and M. Perlman (Eds.), *Functions: New essays in the philosophy of psychology and biology* (pp. 113–143). Oxford: Oxford University Press.

Moore, G.E. (1903). *Principia ethica.* Cambridge: Cambridge University Press.

Neander, K. (1991). The teleological notion of function. *Australasian Journal of Philosophy*, 69, 454–468.

Neander, K. (1995). Misrepresenting and malfunction. *Philosophical Studies*, 79, 109–141.

Papineau, D. (1987). *Reality and representation.* Oxford: Blackwell.

Perlman, M. (2002). Pagan teleology: Adaptational role and the philosophy of mind. In A. Ariew, R. Cummins, and M. Perlman (Eds.), *Functions: New essays in the philosophy of psychology and biology* (pp. 263–290). Oxford: Oxford University Press.

Pfeifer, R., and Scheier, C. (1999). *Understanding intelligence.* Cambridge, Massachusetts: MIT Press.

Regan, T. (1980). Animal rights human wrongs. *Environmental Ethics*, 2(2), 99–120.

Roberts, M. (1997). *The man who listens to horses.* New York: Random House.

Rosenberg, A. (1990). The biological justification of ethics: A best-case scenario. *Social Philosophy and Policy*, 8(4), 86–101.

Ruse, M. (1986). *Taking Darwin seriously.* Oxford: Basil Blackwell.

Ruse, M. (1990). Evolutionary ethics and the search for predecessors: Kant, Hume, and all the way back to Aristotle? *Social Philosophy and Policy*, 8, 58–85.

Sober, E. (1984). Holism, individualism, and the units of selection. In E. Sober (Ed.), *Conceptual issues in evolutionary biology* (pp. 184–299). Cambridge, Massachusetts: MIT Press.

Sober, E. (1992). Models of cultural evolution. In P. Griffiths (Ed.), *Trees of life: Essays in philosophy of biology* (pp. 17–40). Dordrecht, The Netherlands: Kluwer.

Von Eckardt, B. (1993). *What is cognitive science?* Cambridge, Massachusetts: MIT Press.

Whiten, A. (1993). Evolving a theory of mind: The nature of non-verbal mentalism in other primates. In S. Baron–Cohen, H. Tager–Flusberg, and D.J. Cohen (Eds.), *Understanding other minds: Perspectives from autism.* Oxford: Oxford University Press.

Whiten, A. (1996). When does smart behaviour reading become mindreading? In P. Carruthers and P.K. Smith (Eds.), *Theories of theories of mind* (pp. 277–292). Cambridge: Cambridge University Press.

Whiten, A. (1998). Evolutionary and developmental origins of the mindreading system. In J. Langer and M. Killen (Eds.), *Piaget, evolution and development* (pp. 73–102). Lawrence Erlbaum.

Wilson, D.S. (1997). Introduction: Multilevel selection theory comes of age. *The American Naturalist, 150* (Supplement), S1–4.

Wilson, D.S., and Sober, E. (1994). Re-introducing group selection to the human behavioral sciences. *Behavioral and Brain Sciences, 17*, 585–654.