8

A Guide to Naturalizing Semantics

BARRY LOEWER

Semantic predicates – *is true, refers, is about, has the truth-conditional content that p* – are applicable both to natural-language expressions and to mental states. For example, both the sentence "The cat is crying" and the belief that the cat is crying are about the cat and possess the truth-conditional content that the cat is crying. It is widely thought that the semantic properties of natural-language expressions are derived from the semantic properties of mental states.[1] According to one version of this view, the sentence "The cat is crying" obtains its truth-conditions from conventions governing its use, especially its being used to express the thought that the cat is crying. These conventions are themselves explained in terms of the beliefs, intentions, and so forth of English speakers.[2] In the following I will assume that some such view is correct and concentrate on the semantic properties of mental states.[3]

In virtue of what do mental states possess *their* semantic properties? What makes it the case that a particular mental state is about the cat and has the truth-conditions that the cat is crying? The answer cannot be the same as for natural-language expressions, since the conventions that ground the latter's semantic properties are explained in terms of the semantic properties of mental states. If there is an answer, that is, if semantic properties are real and are not fundamental, then it must be that they are instantiated in virtue of the instantiation of certain non-semantic properties. Recently a number of philosophers, whom I will call "Semantic Naturalizers," have attempted to answer this question in a way that they take to be compatible with *Naturalism*. Naturalism's central contention is that everything there is, every individual, property, law, causal relation, and so on, is ontologically dependent on natural individuals, properties, and so forth. It is not easy or straightforward to spell out the notion of ontological dependence; but for the purposes of this discussion I will understand it as including the claim that for each instantiation of property M there are instantiations of natural properties and relations, P, P*, …, that together with natural laws and causal relations among the P instantiations *metaphysically* entail M's instantiation. This characterization is intended to capture the idea that M is instantiated *in virtue of* the

A Companion to the Philosophy of Language, Second Edition. Edited by Bob Hale, Crispin Wright, and Alexander Miller.

P instantiations. Or, to put it metaphorically, Naturalism is the thesis that for God to create our world He needed only to have created the naturalistic entities and laws. Everything else follows from these.[4]

Naturalists are seldom explicit concerning exactly which properties are the natural ones. Their working account is that the natural properties are those expressed by predicates appropriately definable in terms of predicates that occur in true theories of the natural sciences.[5] Most contemporary naturalists think that all natural-science properties are identical to, or are exemplified in virtue of the exemplification of, fundamental physical properties. These are the properties that occur in laws of fundamental physics. This version of naturalism is physicalism; all God needed to do to create our world was to create the physical properties and laws and set the physical initial conditions. Whether or not they accept physicalism, Semantic Naturalizers assume that certain modal notions, specifically law, causation, and probability, are naturalistically respectable. Whether these notions can be grounded in contemporary physics (or physics and the other natural sciences), or even whether they may presuppose semantic concepts, is not without controversy. Of course, if these notions presuppose semantic notions then they cannot form the basis for a physicalistic or naturalistic reduction of semantics. At best one would have a metaphysical reduction of semantics.[6] Since this issue is seldom addressed by Semantic Naturalizers, and discussing it would involve us in controversial issues in metaphysics, I will, for the most part, ignore it in the following.

Semantic Naturalism is a *metaphysical* doctrine about the status of semantic properties.[7] Semantic Naturalizers also endorse an *epistemic* thesis that I will call "perspicuous semantic naturalism." It is the view that, at least in some cases, the metaphysical connections between naturalistic and semantic properties are sufficiently systematic and transparent to allow us to see that certain naturalistic conditions are sufficient for certain semantic properties. If Semantic Naturalizers were to find naturalistic conditions that are metaphysically sufficient for semantic properties, and know that they have found such conditions, they would show how semantic naturalism can be true and thus place the semantic within the natural order. This guide reviews recent naturalization proposals and the prospects of the naturalization project.

Although Naturalism in something like the above sense is widely endorsed in contemporary philosophy, there is also an active tradition that is inhospitable to semantic naturalism. Adherents to this tradition think that semantic and natural properties are so radically different from each other as to preclude the former from holding in virtue of the latter. Two lines of thought have been especially influential in this regard. One is that semantic properties are essentially normative. A putative example is that it is constitutive of the concept *cat* that it ought to be applied only to cats. Further, it is claimed, such essential normativity cannot be accounted for in purely naturalistic terms.[8] The second line of thought is that the principles that govern the attribution of semantic predicates lead to the indeterminacy of the semantic attributions even given all possible relevant evidence. For example, given all of a person's verbal dispositions (the supposed totality of relevant evidence), principles of attribution license alternative assignments of truth-conditions and references to that person's sentences and terms. It is a verificationist step, but perhaps one that is not inappropriate in this case, to the conclusion that there is no fact of the matter (within the range of indeterminacy) concerning reference and truth-conditions.[9]

There is not a philosophical consensus concerning how far, if any distance at all, these considerations go in undermining semantic naturalism. However, any adequate account of semantic properties will need to account both for the normativity that content properties

possess and for the determinacy of reference and truth-conditions. We will see these issues coming up in various ways in our survey of naturalistic theories.

But first we should note the consequences if semantic naturalism is false. Those who believe that it is false respond in two ways. One is to claim that there are no semantic properties (or that they are never instantiated). This view, Semantic Eliminativism (Churchland, 1981), thus preserves naturalism at the expense of semantics. The other response is to claim that there are semantic properties but they are metaphysically independent of natural properties. This view, Semantic Dualism (Davidson, 1980, especially pp. 207–224 and 245–260; McDowell, 1994), thus preserves semantics at the expense of naturalism. Neither option is very pretty. Eliminativism strikes some philosophers as self-refuting (Boghossian, 1990) and others (Fodor, 1987) merely as obviously false in light of the success of folk-psychological and cognitive-science explanations that employ semantic concepts.[10] Semantic Dualism seems incompatible with semantic properties playing a genuine causal role in producing behavior. If, as is widely believed, the natural sciences are causally complete, then there seems to be no room for causation (of physical effects) in virtue of properties metaphysically independent of natural properties (Papineau, 1993; Loewer, 1995). So the situation seems to be that while there are reasons to worry that semantic naturalism might be false, there are also reasons to doubt the alternatives. The semantic naturalist will resolve this paradox if he can produce a naturalization of semantic properties. That would be enough to quell doubts concerning semantic naturalism, since we would then know that the gap between the semantic and the natural can be bridged.

The mental states that have been the focus of naturalization proposals are the propositional attitudes: desire, belief, and perception (perceptual belief) (see Chapter 14, PROPOSITIONAL ATTITUDES). There are two parts to naturalizing a particular kind of propositional attitude. First is the specifying of natural facts in virtue of which it is an attitude of that particular type, such as a belief or a desire. Second is the specifying of the natural facts in virtue of which it has its semantic properties, such as its particular truth-conditional content. With regard to the first part, the view held by most semantic naturalizers is that the property of being a particular kind of attitude, such as being a belief, is a *functional* property (Fodor, 1987). Functional properties are higher-level properties instantiated by an individual x in virtue of x (or x's parts) and other entities instantiating lower-level properties that are lawfully or causally related to each other in certain specified ways.

Most semantic naturalizers also think that the property of being a belief (or other propositional attitude) involves an internal mental representation, and that this representation bears the state's semantic properties.[11] On this view, for example, the belief that the cat is crying involves a relation to an internal representation that has the truth-conditional content that the cat is crying. Some semantic naturalizers further propose that mental representations are elements in a *language* of thought, "Mentalese."[12] On this view, complex mental representations are composed of names, predicates, logical particles, and so on, arranged in syntactic structures. Naturalizing the semantics of Mentalese consists in specifying the natural facts in virtue of which simple Mentalese expressions possess their semantic properties, and then showing how the semantic properties of complex expressions are determined by their structure and the semantic properties of their constituents (Field, 1972 and 1978). While not every Semantic Naturalizer buys the language of thought hypothesis, it will often be convenient to presuppose it in what follows.

There are two conceptions of semantic content that have figured in recent discussions of naturalizing content, called "broad content" and "narrow content." "Broad content" refers to

the usual truth-conditional content of intentional mental states. Hilary Putnam (1975) posed thought experiments that have been taken to show that the usual truth-conditional content of certain thoughts fails to supervene on the thinker's intrinsic physical properties. Putnam imagined two people, Oscar and twin-Oscar, who are identical with respect to their intrinsic neurophysiological properties, but who differ in the following ways. Oscar lives on Earth and speaks English. Twin-Oscar lives on a twin-Earth and speaks twin-English. The primary difference between Earth and twin-Earth is that on the latter planet the liquid that fills the oceans, that quenches thirst, and so on is not H_2O but XYZ, a chemical compound indistinguishable from H_2O without chemical analysis. Putnam claims that Oscar's and twin-Oscar's utterances of "water is …" and the thoughts that each expresses with the sentence differ in their truth-conditions. Oscar's thought is true iff H_2O is … and twin-Oscar's thought is true iff XYZ is …. If this is correct, then intentional properties, at least in some cases, do not supervene on intrinsic neurophysiological properties or any properties that supervene on them (such as computational or syntactic properties). This view, *semantic externalism*, is now widely held for thoughts that involve natural-kind concepts like *water*.

"Narrow content" is a term introduced to designate content properties that do supervene on intrinsic neurophysiological properties (Fodor, 1981 and 1987). While Oscar and twin-Oscar's thoughts differ in broad content, they agree in narrow content. Some philosophers (Fodor, 1987) have argued that only narrow-content properties are implicated in intentional causation, and for this reason are required by an intentional science; but there is little agreement concerning exactly how to characterize it, or even whether there are such properties (Stalnaker, 1991). In any case, most of the naturalization proposals concern broad properties, specifically reference and truth-conditional content, so that will be our focus here.

What naturalistic facts are plausible candidates to serve as metaphysically sufficient for the semantic properties of mental representations? Putnam's twin-Earth thought experiments and Kripke's well-known theory of proper names (Kripke, 1972) both suggest that causal relations are involved in determining the references of predicates and names (see Chapter 35, REFERENCE AND NECESSITY, §4). Their considerations seem to carry over to mental representations corresponding to predicates and names. It is plausible that Oscar's mental representation "water" refers to H_2O partly in virtue of the fact that H_2O has caused or is apt to cause Oscar to think water thoughts. And it is also plausible that part of the account of what makes a person's mental representation "Aristotle" refer to Aristotle is that it possesses a causal history that originates with a baptism of Aristotle. Neither Putnam nor Kripke are sympathetic to the naturalization project, but their work is often taken as the starting point for naturalistic proposals. Causation and kindred notions like law, counterfactuals, and probability seem to be the "right stuff," if there is right stuff, out of which to try to build naturalistic accounts of intentionality.[13]

The Crude Causal Theory

I will begin our survey of specific naturalization proposals with the crude causal theory (CCT) for the reference of Mentalese predicates f. No one has ever held the CCT, but it will be useful to describe it and note its most obvious defects, since these are the problems that more sophisticated accounts are designed to solve.

(CCT) It is metaphysically necessary that (if tokens of f are caused by and only by instances of the property F then f refers to F).

The obvious problem with the CCT is that it doesn't allow for the possibility of tokening f or a sentence containing f that is not caused by F. This is called "the problem of error," since if f occurs as part of the perceptual belief that x is an f, then since f is caused by F it follows that the belief is true. But of course, a perceptual belief, such as the belief that x is a cat, may be caused by a small dog, not by a cat. The problem of error is a special case of the disjunction problem. The CCT implies that, whether or not f is a component of a belief, the disjunction of all the causes of f's tokens are the reference of f; so if f is caused by cats, small dogs, utterances of "cat," and so on, then CCT says that f refers to the property of being a cat or a small dog or an utterance of "cat," and so on. Clearly many of the causes of f need not be included within what it refers to. A naturalist successor to the CCT will need to find some way of naturalistically distinguishing the reference constituting causes from the others.

A second problem is that semantic relations are apparently more fine-grained than causal relations. This is the "fine-grainedness problem": f may refer to F and not G even though F and G are metaphysically or nomologically co-instantiated. For example, the properties of being triangular and of being trilateral are apparently distinct, but necessarily co-instantiated. Triangular things cause tokens of f just in case trilateral things do, but a predicate can refer to one property but not the other. Quine (1960) pointed out a pervasive type of property co-instantiation. When and only when the property of being a rabbit is instantiated, so is the property of being an undetached rabbit part. When one of these properties is causally linked to f, so is the other. This makes it quite difficult to see how a causal theory can account for the difference between thinking that 'there goes a rabbit' and thinking 'there goes an undetached rabbit part.'

Dretske's Information-Theoretic Account

Fred Dretske (1981) proposed a close relative of the CCT that identifies the truth-conditions of a belief state with part of the information that the state carries under certain circumstances. The notion of information can be defined this way: state type T carries information of type p iff there is a nomological or counterfactual regularity (perhaps a *ceteris paribus* law) to the effect that if a T occurs p obtains.[14] So, for example, the height of mercury in a thermometer carries information about the ambient temperature. Dretske's idea is to construct the content of beliefs out of the information that they carry under certain circumstances. An initial and crude formulation of the theory is:

(DRET) It is metaphysically necessary that (if B carries the information that p then B has the truth-condition that p).[15]

Versions of both the fine-grainedness and the error problem cause trouble for DRET. If B carries the information p and p implies q then it also carries the information that q. But, of course, one can believe that p without believing that q, even though p implies q. Dretske responds to this problem by identifying the content of a belief with the *maximal* information that it carries under certain circumstances. This is a little progress, but it leaves untouched the problem that if p and q are nomologically or metaphysically co-occurring then any state that carries information that p carries the information that q. So according to

DRET, no belief can have the exact content that there is a rabbit, since any state that carries the information that there is a rabbit also carries the information that there is an undetached rabbit part. Notice that it is of no avail to protest that a given believer might not even have the concept *undetached*, since that doesn't affect the fact that his belief state still carries the information that there is an undetached rabbit part. Dretske's attempts to handle this problem are not successful.[16]

The error problem arises for DRET in this way. According to DRET, the belief that p always carries the information that p, which means that whenever the belief is tokened it is true. Dretske's proposal for solving the error problem is to identify a subclass of the actual tokenings of B as the bearers of the information that constitutes B's truth-conditions. Tokens of B outside of this class have the same truth-conditional content as those within the class, although they may not carry the same information. This permits (but doesn't obligate) the latter tokens to be false. Dretske's initial specification of the class of tokens of the belief state that fix its truth-conditional content is the class of tokens that occur and are reinforced during what he calls "the learning period." His idea is that during this period a type of mental state becomes a reliable indicator of p, and so comes to have the content that p. So Dretske's official account is

(DRET*) It is metaphysically necessary that (if the maximal information carried by B during the learning period is p then any instance of B has the truth-condition p).

DRET* allows for errors, but its naturalistic credentials are questionable. The trouble is that *learning* seems to be a semantic notion. Dretske may think that it is possible to characterize the learning period non-semantically, but he can't just take this for granted. In any case, even if the learning period could be characterized naturalistically, the account is implausible, at least for some beliefs. There are some beliefs that are learned in circumstances in which the information they carry is not the belief's content. For example, when a child learns to token a belief with a content about tigers by seeing pictures of tigers, her belief states carry information about pictures, although their content is about tigers. Dretske's account will end up assigning the wrong truth-conditional contents to these beliefs.[17]

Optimal Conditions Accounts

A different way of specifying a belief's content is in terms of the information it would carry under epistemically optimal conditions (Stampe, 1977; Stalnaker, 1984; Fodor, 1990a; 1990b). The core idea of this approach is that there is a class of beliefs for which there are conditions – the epistemically optimal conditions – under which a person has the belief just in case it is true.

(OPT) It is metaphysically necessary that (if B is a belief of kind K then there are epistemically optimal conditions C_B such that B's truth-condition is p if, were C_B the case, then B would nomologically covary with p).

So, for example, if for subject A there is a belief state B, that under optimal conditions covaries with the presence of a red ball located in front of her, then B's content is that there is a red ball in front of A. In this case appropriate optimal conditions are that A's eyes are open, she is attending to what she sees, the lighting is good, and so on.

OPT allows for errors, since tokens of B that don't occur in epistemically optimal conditions need not be true. It also seems to supply truth-conditions with normative force, at least if epistemic optimality is a normative notion. But, like Dretske's theory, its specification of the meaning constituting conditions is not naturalistic. "Epistemically optimal" is clearly an intentional predicate. It is not at all clear that epistemically optimal conditions can be specified without reference to semantic notions. Different conditions are "optimal" for different beliefs. For example, epistemically optimal conditions for the perceptual belief that there is a red ball in the room include good lighting; but optimal conditions for the belief that there is a firefly in the room are that the lights are off. This example makes it obvious that the optimal conditions for acquiring true beliefs depend on the belief's content. Of course the naturalizer cannot appeal to the content of a belief in characterizing optimal conditions.

Not only are epistemically optimal conditions for a belief sensitive to the belief's content, but for most beliefs, if they possess optimal conditions at all, these conditions involve other beliefs. Whether or not a person's belief state reliably covaries with a state of affairs depends on what other beliefs that person has. For example, a person who fails to believe that fossils are derived from once-living organisms, or who believes that the Earth is 6,000 years old, will not reliably form beliefs about the age of a fossil. If there are optimal conditions for forming beliefs concerning the age of fossils, those conditions will involve having certain beliefs and not having certain other beliefs. To assume that optimal conditions can be characterized naturalistically looks as though it begs the naturalization problem rather than solving it.[18]

Teleological Theories

Teleological theories propose to explain the truth-conditional content of mental states, especially certain desires and beliefs, in terms of their biological functions. A crude teleological theory (CTT) for belief is:

(CTT) It is metaphysically necessary that (if O is an organism and B is one of its belief states and it is B's biological function to carry the information that p then B has the truth-conditions that p).[19]

The concept of a biological function is defined in terms of natural selection (Wright, 1973; Neander, 1991) along the following lines: it is the function of biological system S in members of species s to F iff S was selected by natural selection because it Fs.[20] S was selected by natural selection because it Fs just in case S would not have been present (to the extent it is) among members of s had it not increased fitness (that is, the capacity to produce progeny) in the ancestors of members of s.[21] So CTT says that if B was selected because it carried the information that p, then B has the truth-condition that p.

CTT is naturalistic and allows for error. In fact, it is compatible with almost all tokens of B being false, since all that is required is that B was selected because it carried the information that constitutes its content; and that could be so even if most past and no present tokens of B are true. It also seems to supply truth-conditional content with normativity. Just as a heart ought to pump blood, B ought to be tokened only if it carries the information that p. There are, however, a number of problems with CTT. One is that it directly applies only to beliefs composed out of innate concepts, since only beliefs involving innate concepts could possess a biological function. Perhaps the notion of biological function can be extended

beyond features selected by natural selection; but that remains to be seen. A second, and more worrying problem, is that it either fails to assign determinate contents or assigns contents that are much too thick-grained to be the truth-conditions of beliefs. This problem has been discussed mostly with respect to the belief, or proto-belief, of animals, especially a frog's (the hope being that extension to a human's will come when the bugs are worked out).

Suppose that B is an internal state of a frog that is responsive to stimuli and that controls the frog's snapping behavior. Tokens of the state B in the frog's ancestors generally carried a great deal of information including: that flies are present, that small moving black things are present, that food is present, and so on. Furthermore, since these various conditions were reliably co-instantiated in the environment in which the frog evolved, they are all equally good candidates to be the information that it is the function of B to carry. So CTT implies either that B's content is indeterminate among components of the package or that its content is the whole package of information.

It is not clear whether this is an objection to teleological accounts, since it is not clear what beliefs or desires, if any, frogs have. But it is an objection if teleological accounts are incapable of delivering more fine-grained contents than the one they apparently attribute to the frog. More elaborate theories of content that promise to solve this problem are due to Millikan (1984; 1986; 1989) and Papineau (1993). Both accounts, especially Millikan's, are rather elaborate. Here I will just briefly sketch Papineau's approach.

(PAPB) If D is a desire and B a belief and p is the (minimal?) state of affairs whose obtaining guarantees that actions based on B and D satisfy D then B has the truth-condition p.

If we suppose that the frog desires to catch a fly, and that this desire together with B lead to his snapping, then B's truth-conditional content is the minimal state of affairs that will guarantee that snapping will result in catching a fly. In this case it is a belief with something like the content *if I snap then I will catch a fly*. Of course, PAPB is not naturalistic, since it appeals to the concept of satisfying a desire and that is a semantic concept. Papineau attempts to remedy this by providing a naturalistic account of the contents of desires.

(PAPD) If q is the minimal state of affairs such that it is the biological function of D to operate in concert with beliefs to bring about q then D is the desire that q.

Papineau's idea is that if the desire of type D was selected because it contributed by acting in concert with beliefs to bringing about q, then q is D's content. Let's suppose that the content of A's desire D is that she eats an apple. On a particular occasion, D (together with beliefs) may cause the moving of A's hand, A's eating an apple, A's eating a fruit, and A's being nourished. Papineau suggests that the moving of the hand (to grasp the apple) isn't among D's functions, since there are occasions when D was selected (A's ancestors who possessed D had increased fitness, or D was reinforced in A) even though D didn't cause their hands to move. On the other hand, Papineau supposes that whenever D was selected A ate an apple, ate a fruit, was nourished, and so on. He suggests that the most specific of these features of the behavior which led to D's being selected is D's content; that is, eating an apple.

There are a number of worries that one might have concerning Papineau's account. One is that it applies, at best, only to certain beliefs and desires. PAPB provides contents only to means–ends beliefs (although Papineau suggests how the account can be extended to other beliefs). Many desires could not have been selected for by natural selection, since they are

desires that possess impossible satisfaction conditions, or desires for situations that have never obtained, or have obtained too recently to be selected. It is hard to see how the desire to not have any children (or the desire that no one has any children) could have been selected for on the basis of bringing about its content. Perhaps these objections are not all that damaging if PAPD is intended just as a sufficient condition that applies to a certain class of desires. But then we will need a naturalistic specification of that class of desires. More damaging to PAPD is that possessing the function of bringing about x is not a sufficient condition for D's being the desire to bring about x. Suppose that D is the desire to eat an apple. It is compatible with this that there have been occasions when D led not to apple-eating but to pear-eating (some ancestors of A mistook pears for apples). It is plausible that eating pears (pears being as nutritious as apples) led to increased fitness, in which case D's function is to cause (together with beliefs) eating apples or pears. PAPD yields the result, contrary to our assumption, that D is the desire to eat apples or pears. There seems to be no reason why a desire could not have as its function causing, together with beliefs, some situation that differs from its content. If PAPD is incorrect then PAPB, even if it is correct, is no longer adequate as a naturalization of belief.

It is plausible that the human cognitive system contains subsystems that have the functions of producing states that bring about certain effects, and producing other states that carry certain information (and work in concert with the first kind of state to produce effects). But there is no reason to suppose that these states are individuated exactly in the same way that beliefs and desires are. Truth-conditional content seems much more determinate and fine-grained than anything that teleology is capable of delivering. This is made obvious by considering that there cannot be any selectional advantage for creatures whose beliefs are about rabbits over those whose beliefs are about undetached rabbit parts; yet our contents are so fine-grained as to distinguish these belief states.

Fodor's Asymmetric Dependence Theory

Fodor (1990b) proposed a variant of the causal (or informational) account that is intended to be a naturalization of the reference of a simple Mentalese predicate. It appeals to the idea that the meaning-constituting causes are those which, in a sense to be soon explained, are resilient. It will simplify exposition of his theory to define two technical notions. The law $Q \rightarrow C$ (Qs cause Cs) *asymmetrically depends* on the law $P \rightarrow C$ just in case if Ps didn't cause Cs then Qs would not cause Cs but if Qs didn't cause Cs then Ps would still cause Cs. *C locks onto P* just in case (1) it is a law that Ps cause Cs, (2) there are Qs (= Ps) that cause Cs, and (3) for any $Q \neq P$, if Qs cause Cs then Qs causing Cs asymmetrically depends on Ps causing Cs.[22] If C locks onto P then $P \rightarrow C$ is resilient in that it survives the breaking of $Q \rightarrow C$ for Qs other than P. Fodor's proposal, then, is:

(ADT) It is metaphysically necessary that (if C locks onto P then C refers to P).

Suppose that it is a law that cows cause "Cow"s (or rather the word's Mentalese counterpart), that other things also cause "Cow"s, and that such causal relations asymmetrically depend on the 'cow → "Cow"' law. Then, according to ADT, "Cow" refers to cow. ADT handles the error and disjunction problems this way. Horses on a dark night can cause "Cow"s even though the horses on dark nights are not among the reference-constituting causes of "Cow"; that is, the law that horses on a dark night → "Cow"s depends on the law

that cows → "Cow"s. If a horse caused "Cow" is a constituent of the belief "There is a cow," then the belief is false. Of course, this account of error is correct only if ADT is correct. If ADT is not correct then it may count some erroneous beliefs as true, or some true beliefs as erroneous.

Along with the theory, Fodor provides some commentary that helps to understand it. One point is that the law connecting a property to a predicate that refers to it is a *ceteris paribus* law. That is, it holds only as long as certain unspecified conditions obtain. Presumably this means that only under certain kinds of circumstances do cows actually cause A's mental representation "Cow." Presumably these conditions are that cows are perceptually salient to A, A's perceptual system is in good working order, and so on. A second point involves the dependence relation between causal laws. Sometimes Fodor says that it is a basic relation among laws that cannot be explained in other terms. But sometimes he explains it in terms of counterfactuals; Q → C depends on P → C just in case if P → C had not obtained then neither would Q → C have obtained. Fodor insists that the counterfactual be understood *synchronically*, not *diachronically*. If A learned to recognize cows on the basis of pictures of cows, then it may be that cow → "Cow" depends diachronically on cow-picture → "Cow." That is, it is true that if there hadn't been a causal connection between pictures of cows and A's "Cow"s, there wouldn't be a connection between cows and A's "Cow"s. But Fodor thinks that synchronic dependence goes in the opposite direction. Once A has acquired "Cow" then cow → "Cow" is more resilient than cow-picture → "Cow." A third point is that the account of reference is *atomic*. By this is meant that it is metaphysically possible for A's Mentalese predicate C to lock onto P, even if C bears no inferential or causal relations to any of A's other symbols, or even if A's Mentalese vocabulary contains only the predicate C. Fodor welcomes this surprising feature of his account, since he thinks that there are reasons to hold that inferential or causal relations among thoughts are not constitutive of the thought's semantic properties (Fodor and Lepore, 1992).

There are two questions that need answers to evaluate Fodor's theory. First, is it genuinely naturalistic? And, second, is C locking onto P really a sufficient condition for C's referring to P? Answering these questions is made difficult by the fact that the central notions in Fodor's account – *ceteris paribus* laws and asymmetric dependence between laws – are technical notions that are not clearly defined.

There are two places to worry whether ADT is genuinely naturalistic. First, supposing that it is a law that P → C then it is reasonable to believe that its *ceteris paribus* conditions include having and not having certain other intentional states. We noticed a similar point in our discussion of optimal-conditions theories. Does this make P → C non-naturalistic? Not necessarily. If the fact that P → C is a law is naturalistically reducible, then it too is a naturalistic fact. But do we have any reason other than the belief that semantic naturalism is true to think that P → C is naturalistically reducible?

Second, and more worrying, is whether the dependency relations that Fodor requires are naturalistic. These dependency relations are not themselves the subject of any natural science; so Fodor cannot claim, as the teleosemanticist does, that he is explaining a semantic notion in terms of a scientifically respectable notion, that is, a biological function. Further, it is not obvious that the synchronic counterfactuals that Fodor appeals to when explaining asymmetric dependence have truth-conditions that can be specified non-intentionally. Why is Fodor so certain that the counterfactual (synchronically construed) *if cow* → "*Cow*" *were broken then cow-picture* → "*Cow*" *would also be broken* is true? Perhaps if the first law were to fail "Cow" would change its reference to cow-picture and so the second law would

still obtain. If so, then while "Cow" refers to 'cow,' ADT would say that it refers to 'cow-picture.'[23] Fodor cannot respond by saying that in understanding asymmetric dependence the counterfactual should be understood as holding the actual reference of "Cow" fixed, since that would be introducing a semantic concept into the explanation of asymmetric dependence. I do not think that these points show that ADT is not naturalistic; but they do show that the burden is on Fodor to argue for the naturalistic credentials of the dependency relation. Fodor sometimes seems tempted to just take the dependency relation to be metaphysically primitive and declare that it is part of the natural order (Fodor, 1991). One could see some irony in calling on such elaborate metaphysical notions to defend scientific naturalism.

Is the fact that C locks onto P sufficient for C to refer to P? It is difficult to answer this question without having a clear characterization of asymmetric dependence. The intrepid philosopher who thinks that she has devised a counter-example to ADT runs the risk of being told by its inventor that she has gotten the dependency relations wrong. There are a number of such putative counter-examples in the literature (Baker, 1991; Boghossian, 1991; Adams and Aizawa, 1994; Gates, 1996) and answers to the counter-examples by Fodor (1991; 1994).[24] Instead of going into the details of these objections I will sketch two general worries about the account.

We attribute propositional attitudes to one another on the basis of folk-psychological generalizations and general information about what people tend to believe, desire, and so forth under certain circumstances. So, for example, if A is a normal human being looking at a cow 100 feet away, then we expect A to believe that there is a cow in front of her. If, in fact, there is not a cow but a cleverly made cardboard cow-façade, then we expect A to at first believe that there is a cow, but that when she moves closer to the cardboard cow and examines it she will cease to believe that there is a cow. Our ability to attribute beliefs, desires, and so on to each other depends, at least in part, on generalizations like these. When testing a theory of intentionality we appeal to such generalizations. We ask whether it is possible for the putative naturalistic sufficient condition for A's believing that p to be satisfied while our folk-psychological generalizations give the result that A doesn't believe that p. The problem I see with ADT is not that there are clear cases in which C locks onto P, but C fails to refer to P; it is rather that, as far as I can see, ADT doesn't engage folk psychology. For all we know, an assignment of beliefs to A employing ADT and an assignment employing the usual folk-psychological principles may diverge radically. I am not arguing that they must or do diverge, but that Fodor has provided no reason to think they don't. The worry isn't an idle one, since it is not at all clear what asymmetric dependence has to do with our folk-psychological principles of belief-attribution. If ADT is to carry conviction we need some account of why it is that the contents it assigns will match those assigned by folk psychology.[25]

The second problem is the familiar one of the inscrutability of reference that seems to bedevil all naturalistic theories. If cow → "Cow"s is a law, then so is undetached-cow-part → "Cow" (and laws involving various other properties metaphysically co-instantiated with cow: Quine, 1960). Neither one of these putative laws asymmetrically depends on the other since they hold in exactly the same possible worlds. So it looks like if a predicate locks onto any property it either locks onto all those properties that are metaphysically co-instantiated, or onto the disjunction of all these properties (Gates, 1996).

One response to the problem is to declare that properties like undetached-cow-part, temporal state of a cow, and so on are not eligible to enter into laws and causal relations. Without a naturalistic justification of this claim the response is another instance of borrowing from

metaphysics to buy naturalism. Fodor, to his credit, has not taken this route, but has suggested an addition to ADT to cope with the problem (Fodor, 1994). He argues that the inferential relations among sentences containing the predicate "Cow" will differ (for a thinker whose Mentalese contains the truth-functional connectives) depending on whether "Cow" refers to cow or to undetached cow part. By adding further conditions on the inferential relations borne by sentences to each other, he proposes to specify sufficient conditions for "Cow" to refer to cow (and no other property). The account is too complex to deal with in detail here. I will just say that, at best, Fodor's proposal excludes some properties from being the references of "Cow," but fails to single out cow as the unique reference.

Causal-Role Semantics

Causal-role (aka "conceptual role" and "inferential role") semantics (CRS) is another approach to naturalizing semantics that deserves mention, albeit only a brief one here. The mention is brief because although causal-role semantics has been in the air for some time (Sellars, 1974; Harman, 1982; Field, 1978; Loar, 1981; Block, 1986) no one has actually proposed a CRS that is naturalistic and assigns specific truth-conditions to mental states or representations. The basic idea of CRS is that the semantic properties of a mental representation are partially constituted by certain causal or inferential relations between that and other mental representations. If only causal relations among mental representations are taken into account, then at best CRS is an account of narrow content. To turn it into an account of broad content, causal relations between mental representations and external items need to be added.

CRS should be distinguished from theories of interpretation like Davidson's (1984) that also ground truth-conditions in causal relations among mental representations (or natural-language representations) and external events. Davidson's theory of radical interpretation places constraints on the contents of a person's propositional attitudes. The most important one is that a correct theory of interpretation should assign mostly true beliefs. But the account is not a naturalization, since the semantic concept *truth* is used in formulating the constraint. (On Davidson's theory, see further Chapter 13, RADICAL INTERPRETATION.)

The immediate difficulty with CRS is that most of the actual causal roles of a Mentalese sentence do not seem necessary for it to possess its truth-conditions. For example, a person's Mentalese sentences "There is a cat" and "There is an animal" might have their usual truth-conditions even though the person has no disposition to infer the latter from the former. Given externalism, CRS cannot adequately specify sufficient conditions for a sentence to possess particular truth-conditions solely in terms of its causal connections to other sentences. It will also need to invoke causal connections with external items. But this brings it back to the problem of specifying exactly which causal connections are content-constituting. CRS has made no distinctive contribution to answering this question naturalistically. The prospects for a naturalized CRS do not look good (Fodor and Lepore, 1992).

CRS seems to fare better as an account of what makes it the case that logical expressions possess their meanings. For example, it is plausible that dispositions to infer S from S#R, and to infer S#R from the pair of premises S and R, are relevant to making it the case that "#" is conjunction. But elaborating this into a naturalistic sufficient condition of "#" to be conjunction is not completely straightforward. The most obvious difficulty is characterizing those causal relations that count as *inferences* without appealing to *truth*.

Conclusion

None of the naturalization proposals currently on offer are successful. We have seen a pattern to their failure. Theories that are clearly naturalistic (such as CCT) fail to account for essential features of semantic properties, especially the possibility of error and the fine-grainedness of content. Where these theories are sufficiently explicit we have seen that they are subject to counter-examples. In attempting to avoid counter-examples, semantic naturalists place restrictions on the reference (or truth-condition) constituting causes or information. But in avoiding counter-examples these accounts bring in, either obviously or surreptitiously, semantic and intentional notions, and so fail to be naturalistic.

Of course, the failure of naturalization proposals to date does not mean that a successful naturalization will not be produced tomorrow. But another possibility, and one that philosophers have recently begun to take seriously (such as McGinn, 1993), is that while semantic naturalism is true, we may not be able to discover naturalistic conditions that we can *know* are sufficient for semantic properties; that is, perspicuous semantic naturalism may be false. It may be that the naturalistic conditions that are sufficient for semantic properties are too complicated or too unsystematic for us to be able to see that they are sufficient. Or, it may be that there is something about the nature of semantic concepts that blocks a clear view of how the properties they express can be instantiated in virtue of the instantiation of natural properties. This position, though it may be correct, is not by itself intellectually satisfying. The least we would like to know is exactly why we cannot know which natural properties are sufficient for semantic properties.[26] As of now, we don't know whether semantic naturalism is true and, if it is true, we don't know whether we can know, of any particular proposed naturalization, that it is correct: though, as we have seen, we can know of some that they are incorrect.[27]

Notes

1 Proponents of this view include Grice (1957), Lewis (1969), and Fodor (1975). For a contrary view see Davidson (1984), who holds that mental and public language semantic properties are interdependent, and that neither is metaphysically prior to the other.

2 The program of accounting for the semantic properties of natural language in terms of those of mental states is identified with Paul Grice (1957) and Stephen Schiffer (1972). A detailed account in terms of conventions can be found in Lewis (1969). See also Chapter 3, INTENTION AND CONVENTION IN THE THEORY OF MEANING.

3 So in the following, "semantic property" means semantic property of an intentional mental state or event.

4 The proposition that Fx is metaphysically entailed by conditions K just in case K together with a characterization of the nature of F logically imply Fx. The best-understood example of this is the realization of a functional property F by lower-level property instantiations. In this case it logically follows from the functional nature of F, the nature of the Ps, and causal relations among the Ps that whenever the Ps are instantiated M is also instantiated.

5 This characterization is vague with respect to what counts as an appropriate definition, as a property, and as the natural sciences. Removing the vagueness raises a number of problems that would take us too far afield to discuss.

6 Hilary Putnam (see, e.g., 1992) has long maintained that causal and nomological concepts are inextricably bound up with intentionality, and for this reason attempting to naturalize semantics is a misconceived project.

7 Although it is a metaphysical doctrine, it is also contingent, since its truth doesn't rule out possible worlds in which some properties are instantiated but not in virtue of the instantiations of natural properties.

8 There are two issues that are often mentioned by those who think that normativity considerations derail semantic naturalism. One is that grasping a concept involves being in a mental state that obligates one to applying the concept only to items in its extension. It is difficult to see how any purely natural state can entail such an obligation (Kripke, 1982; Boghossian, 1989; see also Chapter 24, RULE-FOLLOWING, OBJECTIVITY, AND MEANING). The other consideration is the claim that the attribution of intentional concepts is constrained by normative principles of rationality and charity (see Chapter 13, RADICAL INTERPRETATION). Davidson (1980; 1984) starts with this claim and tries to fashion it into an argument against the existence of nomic connections between intentional and non-intentional properties. There is little agreement about exactly what Davidson's argument is or even whether its conclusion conflicts with naturalism. Even so, it has been influential, and is often cited or repeated by those skeptical of naturalization (McDowell, 1994).

9 Quine's (1960) arguments for the indeterminacy of translation and for the inscrutability of reference, and Putnam's (1978) so-called model-theoretic argument are instances of this line of thought (see Chapter 26, INDETERMINACY OF TRANSLATION, and Chapter 27, PUTNAM'S MODEL-THEORETIC ARGUMENT AGAINST METAPHYSICAL REALISM).

10 A sophisticated version of eliminativism maintains that robust semantic properties don't exist (or are uninstantiated) but that deflationary semantic predicates can be used to specify reference and truth-conditions. A robust semantic property is a property that may enter into causal explanations and exists independently of our concepts and definitions. In contrast, a deflationary truth predicate, "DT," for a language L is defined by providing a list of the conditions under which the predicate applies; for example, "Snow is white" is DT iff snow is white; "Snow is green" is DT iff snow is green; etc. More generally (p)("p" is DT iff p) where the quantifier is substitutional. An important feature of DT is that, unlike robust truth, it applies only to the language for which it is defined. There is no reason to suppose that items in the extension of a deflationary predicate have anything, in particular causal and explanatory powers, in common. It seems to follow that deflationary semantic notions cannot be employed in causal explanations or play an explanatory role in an intentional cognitive science. The attraction of deflationism (the view that the only instantiated semantic predicates are deflationary ones) is that it both allows us to use semantic predicates for certain purposes (e.g., for infinite conjunction and disjunction) and is compatible with Naturalism. Skepticism concerning deflationism arises from the worry that deflationary truth and reference are too thin to do the work that we want done by semantic concepts. For discussion see Horwich (1990) and Field (1986; 1994).

11 Proponents of this view usually distinguish between explicit and implicit propositional attitudes. Only the former involve relations to mental representations. The latter are dispositions to produce explicit attitudes (Fodor, 1987, ch. 1).

12 Field (1978) and Fodor (1975; 1987) are important sources of this view. Fodor proposes it as an empirical hypothesis that provides the best explanation of certain features of human thought, specifically systematicity and the capacity to engage in logical reasoning.

13 Causation, laws, counterfactuals, and so on are not themselves items mentioned in physics, and it is controversial whether they supervene on physical facts. Even so, Fodor and other naturalizers would consider it a successful naturalization if they could show that intentional properties supervene on these properties. However, Putnam (1992) has complained that notions of law and causation presuppose intentional notions. While this may be true on some accounts of these notions, it is not true on others. For example, on some accounts, probabilities are rational degrees of belief. Obviously, explaining semantic properties of beliefs in terms of degrees of

belief would not contribute to naturalization. On other accounts, probabilities are objective, mind-independent features of the world. In this case there seems to be no danger of circularity, though one may wonder at employing so metaphysical a notion in the cause of naturalism. However, these issues are too complicated to develop here.

14 Dretske (1981) characterizes information in terms of probabilistic relations. There are numerous problems with his account that are avoided by the characterization used here. Also see Loewer (1987).

15 Dretske's formulation characterizes belief functionally as states that guide behavior in certain ways. He doesn't commit himself to a language of thought account of beliefs.

16 This is forcefully argued in Gates (1996).

17 Dretske (1988) suggests a teleological characterization of the state tokens whose information fixes the beliefs content. His basic idea is that those instances of the belief state that produces behavior that is reinforced are the ones whose informational content fixes the belief's semantic content. While this is a naturalistic characterization of the class, it is questionable whether it assigns appropriate contents. It is easy to imagine situations in which a false token of a belief produces behavior that is reinforced. For further discussion of Dretske's theory see Loewer (1987) and McLaughlin (1993).

18 This point is developed in Loewer (1987) and more thoroughly in Boghossian (1991).

19 Some teleological accounts employ a more general characterization of information. S carries the information that p iff P(p/S occurs) > P(p/S doesn't occur).

20 Selection by conditioning (i.e., by reinforcement) also figures in accounts of function devised by some teleosemanticists (Dretske, 1988).

21 For example, the biological function of the heart is to pump blood (not to make a thumping sound) since it is that property of pumping blood (not making a thumping sound) that accounts via natural selection for the presence of hearts. Notice that something may have the function to F even if it doesn't F or seldom Fs. It should be noted that it doesn't follow that every biological system that does something useful has that as its function (or that it has any function). Only those things that a system does that lead to an increase in fitness are its functions. So, for example, it is not obvious that certain cognitive abilities are the product of any function.

22 Fodor sometimes also adds the requirement that the law $P \rightarrow C$ is instantiated. This is supposed to give the result that Oscar's Mentalese "water" refers to H_2O and twinOscar's Mentalese "water" refers to XYZ. However, this addition may not be needed if the dependency relations concerning laws involving Oscar's and twin-Oscar's mental expressions are different.

23 Boghossian (1991) argues that locking on is either not sufficient for reference or is not naturalistic. His argument shows that to get the counterfactuals that underlie the locking-on relation to come out right, the similarity relation relative to which they are evaluated must take into account *semantic similarities*.

24 One of Boghossian's counter-examples to Fodor's theory is particularly persuasive. He imagines a natural kind concept K and laws $X \rightarrow K$ and $Y \rightarrow K$ where X and Y are different substances that are nomologically indistinguishable by us (they behave differently only in black holes). It may then be that neither of these laws asymmetrically depends on the other. Fodor's theory would have the consequence that K refers to the disjunction $X \vee Y$. But surely in the imagined situation K might refer only to X in virtue of the role it plays in physical theory.

25 This point is developed at length in different ways by Carl Gillett and Andrew Milne in dissertations at Rutgers.

26 Boghossian (1990) argues that belief holism (the fact that which situations are apt to cause one to acquire a particular belief depends on one's other beliefs) prevents us from certifying that any naturalistic condition on content-constituting causes or information is correct.

27 I am grateful to Paul Boghossian, Jerry Fodor, Gary Gates, Carl Gillett, and Fritz Warfield for helpful discussion and (not always heeded) advice.

References

Adams, F., and K. Aizawa 1994. "Fodorian semantics." In *Mental Representation*, edited by S. P. Stich and T. A. Warfield, pp. 223–242. Oxford: Blackwell.

Baker, L. 1991. "Has content been naturalized?" In Loewer and Rey, 1991, pp. 17–32.

Block, N. 1986. "Advertisement for a semantics for psychology." In *Studies in the Philosophy of Mind*, edited by P. French, T. Uehling, and H. Wettstein. *Midwest Studies in Philosophy*, vol. 10. Minneapolis: University of Minnesota Press.

Boghossian, P. 1989. "The rule following considerations." *Mind*, 98(392): 507–549.

Boghossian, P. 1990. "The status of content." *Philosophical Review*, 99(2): 157–184.

Boghossian, P. 1991. "Naturalizing content." In Loewer and Rey, 1991, pp. 65–86.

Churchland, P. 1981. "Eliminative materialism and the propositional attitudes." *Journal of Philosophy*, 78(2): 67–90.

Davidson, D. 1980. *Essays on Actions and Events*. Oxford: Clarendon Press.

Davidson, D. 1984. *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press.

Dretske, F. 1981. *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Dretske, F. 1988. *Explaining Behavior*. Cambridge, MA: MIT Press.

Field, H. 1972. "Tarski's theory of truth." *Journal of Philosophy*, 69(13): 347–375.

Field, H. 1978. "Mental representation." *Erkenntnis*, 13: 9–61.

Field, H. 1986. "The deflationary conception of truth." In *Fact, Science, and Value*, edited by G. McDonald and C. Wright, pp. 55–117. Oxford: Blackwell.

Field, H. 1994. "Deflationist views of meaning and content." *Mind*, 103(411): 249–285.

Fodor, J. 1975. *The Language of Thought*. New York: Thomas Y. Cromwell.

Fodor, J. 1981. "Methodological solipsism." In *Representations: Philosophical Essays on the Foundations of Cognitive Science*. Brighton: Harvester.

Fodor, J. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press.

Fodor, J. 1990a. "Psychosemantics, or where do truth conditions come from." In *Mind and Cognition*, edited by W. Lycan, pp. 312–338. Oxford: Blackwell.

Fodor, J. 1990b. *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.

Fodor, J. 1991. "Replies." In Loewer and Rey, 1991, pp. 255–319.

Fodor, J. 1994. *The Elm and the Expert*. Cambridge, MA: MIT Press.

Fodor, J., and E. Lepore. 1992. *Holism: A Shopper's Guide*. Oxford: Blackwell.

Gates, G. 1996. "The price of information." *Synthese* 107(3): 325–347.

Gillett, C. 1997. "Naturalization: Physicalism and Scientific Theory Appraisal." PhD diss., Rutgers University.

Grice, P. 1957. "Meaning." *Philosophical Review*, 66(3): 377–388.

Harman, G. 1982. "Conceptual role semantics." *Notre Dame Journal of Formal Logic*, 23(2): 242–256.

Horwich, P. 1990. *Truth*. Oxford: Blackwell.

Kripke, S. 1972. *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Kripke, S. 1982. *Wittgenstein on Rules and Private Language*. Cambridge, MA: Harvard University Press.

Lewis, D. 1969. *Convention*. Cambridge, MA: Harvard University Press.

Loar, B. 1981. *Mind and Meaning*. Cambridge: Cambridge University Press.

Loewer, B. 1987. "From information to intentionality." *Synthese*, 70(2): 287–317. Reprinted in Stich and Warfield, 1994.

Loewer, B. 1995. "An argument for strong supervenience." In *New Essays on Supervenience*, edited by E. Savellos, pp. 218–225. Cambridge: Cambridge University Press.

Loewer, B., and G. Rey, eds. 1991. *Meaning in Mind: Fodor and his Critics*. Oxford: Blackwell.

McDowell, J. 1994. *The Mind and the World*. Cambridge, MA: Harvard University Press.

McGinn, C. 1993. *Problems in Philosophy: The Limits of Inquiry*. Oxford: Blackwell.

McLaughlin, B., ed. 1993. *Dretske and his Critics*. Oxford: Blackwell.
Millikan, R. 1984. *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.
Millikan, R. 1986. "Thoughts without laws: cognitive science with content." *Philosophical Review*, 95(1): 47–80.
Millikan, R. 1989. "Biosemantics." *Journal of Philosophy*, 86(6): 281–297. Reprinted in Stich and Warfield, 1994, pp. 243–258.
Milne, A. 1996. "The Alienation of Content: Truth, Rationality and Mind." PhD diss., Rutgers University.
Neander, K. 1991. "Functions as selected effects: the conceptual analyst's defense." *Philosophy of Science*, 58(2): 169–184.
Papineau, D. 1993. *Philosophical Naturalism*. Oxford: Blackwell.
Putnam, H. 1975. *Mind, Language and Reality (Philosophical Papers,* vol. 2*)*. Cambridge: Cambridge University Press.
Putnam, H. 1978. *Meaning and the Moral Sciences*. London: Routledge and Kegan Paul.
Putnam, H. 1992. *Renewing Philosophy*. Cambridge, MA: Harvard University Press.
Quine, W. V. O. 1960. *Word and Object*. Cambridge, MA: MIT Press.
Schiffer, S. 1972. *Meaning*. Oxford: Oxford University Press.
Sellars, W. 1974. "Meaning as functional classification." *Synthese*, 27(3–4): 417–437.
Stalnaker, R. 1984. *Inquiry*. Cambridge, MA: MIT Press.
Stalnaker, R. 1991. "Semantics for the language of thought." In Loewer and Rey, 1991, pp. 229–237.
Stampe, D. 1977. "Towards a theory of linguistic representation." *Midwest Studies in Philosophy*, 2(1): 42–63.
Stich, S., and T. Warfield, eds. 1994. *Mental Representation*. Oxford: Blackwell.
Wright, L. 1973. "Functions." *Philosophical Review*, 84(2): 139–168.

Further Reading

Field, H. 1977. "Logic, meaning, and conceptual role." *Journal of Philosophy*, 74(7): 379–409.
Kim, J. 1993. *Supervenience and Mind*. Cambridge: Cambridge University Press.
McGinn, C. 1982. "The structure of content." In *Thought and Object*, edited by A. Woodfield, pp. 207–259. Oxford: Oxford University Press.
Millikan, R. 1991. *White Queen Psychology and Other Essays for Alice*. Cambridge, MA: MIT Press.
Neander, K. 1995. "Misrepresenting & malfunctioning." *Philosophical Studies*, 79(2): 109–141.
Pietroski, P. 1992. "Intentionality and teleological error." *Pacific Philosophical Quarterly*, 73: 267–282.

Postscript

PETER SCHULTE

How can mental states be about things in the world? How is it possible that certain mental states are true (or satisfied) under some conditions, and false (or unsatisfied) under others? These questions continue to puzzle philosophers, especially those of a naturalistic bent, since all attempts to explain semantic properties in naturalistic terms face serious difficulties. This has led some theorists to conclude that the whole naturalization program is wrongheaded, but others still adhere to it: while admitting that the task of naturalizing semantic properties is more difficult than naturalists might have initially supposed, they see no reason to be pessimistic about the program as a whole. Hence, the debate about naturalizing semantics – or, more specifically, about naturalizing mental content – continues, and has even intensified in recent years. Since teleological theories are at the center of this debate, I concentrate on them in what follows.

Teleological Theories: Basic Distinctions

It is useful to distinguish between two questions which semantic naturalizers must address (Sterelny, 1995, p. 254). The first question concerns *representational status*: What distinguishes representational states, that is, states with semantic content, from non-representational states? Or, more precisely, in virtue of which natural facts do certain internal states qualify as representational states? Call this the 'status question.' The second question concerns *content determination*: Given that R is a representational state, why does R have the content that p rather than some other content? To put it differently, which are the natural facts that determine the content of R? Call this the 'content question.' Of course, those two questions are intimately related, and many theorists answer them both at once (e.g., Millikan, 1984), but it is important to emphasize that they can be discussed separately.

Traditionally, semantic naturalizers have focused on the content question, and they have continued to do so in recent years. One of the main debates in this area is the debate between proponents of two different teleological approaches to content determination, the *input-* and the *output-oriented* approach. These approaches differ fundamentally in the way they specify the content of descriptive representational states (i.e., 'belief-like' states with a mind–world direction of fit). Consider the descriptive representational state R. According to the input-oriented approach, we have to look 'upstream' at the functions of the mechanisms that generate R, or to R's functional relations to causes or conditions in the world, in order to specify the semantic content of R. One example of an input-oriented account is the "crude teleological theory" (p. 180), which equates R's semantic content with the information that R is supposed to carry (Dretske, 1988). According to the output-oriented approach, on the other hand, R's content is primarily dependent on 'downstream factors' – for example, on the biological function of mechanisms that respond to R, or on R's functional relations to behavior. Papineau's (1993; 1998) teleological theory is a variant of the latter approach, since he proposes that belief content depends on desire content, and that the content of a desire is the most specific state of affairs it is supposed to bring about (p. 181).

The Content Question: Input-Oriented Theories

An attractive new version of the input-oriented approach is Karen Neander's (2013) causal-informational version of teleosemantics. It should be noted that her theory is restricted to a subclass of descriptive representational states – namely, to perceptions. According to Neander, the content of a perceptual state is determined by a certain function of the perceptual mechanism that produces it (i.e., by a function of its 'producer'). More precisely, it is determined by the producer's *response function*: the function to generate R in response to certain environmental conditions p. So Neander's teleological theory is

(NTT) If R is a perceptual state and R's producer has the function of producing R in response to the state of affairs p, then R has the content that p.

Consider again the case of the frog (p. 181), and assume that N is the neural state that (a) is normally produced when there's a small, dark, moving object in the frog's visual field, and (b) normally triggers prey-catching behavior. According to Neander's account, N has the

content that a small, dark, moving object is present, because the frog's visual system has the function of producing N in response to a small, dark, moving object (cf. Neander, 2013, p. 31).[1] *Responding to a small, dark, moving object by producing N* is what, among other things, the frog's visual system has been selected for.

At this point, one might ask: Can't we also ascribe the function of *responding to a nutritious object ('frog food') by producing N* to the frog's visual system? What's more, isn't this description in some sense superior, given that it is only because small, dark, moving objects often contained nutrients that the system was selected for responding to them? This seems to raise the notorious indeterminacy problem all over again (p. 181). However, it should be noted that Neander uses the term 'responding' in a strictly causal sense (Neander, 2013, p. 23). A mechanism can only be selected for responding to some x's being F by producing R if x's being F is *causally relevant* for bringing about the mechanism's production of R. But in the case of the frog, the object's being nutritious is *not* causally relevant for bringing about the visual system's production of N, only the object's being small, dark, and moving is (as can easily be tested by varying the properties independently). Hence, the system can only have been selected for responding to small, dark, moving objects. This is true even though the evolutionary *reason* why the system was selected for doing this consists in the fact that small, dark, moving objects were (often enough) nutritious. (Since Neander also analyzes information in terms of causation, she argues that NTT can be restated as a version of informational teleosemantics. This claim, however, is not essential to her theory.)

Neander's input-oriented account of perceptual content is thus of great interest, since it promises to solve the indeterminacy problem that was fatal to earlier input-oriented theories. Still, there are several objections that can be raised against it. One worry concerns the notion of a response function (Millikan, 2013). Another worry centers on a second problem of indeterminacy, often called the 'distality problem.' This problem becomes apparent when we consider that the frog's N-state normally stands at the end of a long chain of causes – it is caused by a pattern of retinal stimulation, which is caused by a pattern of light waves, which in turn is caused by the presence of a small, dark, moving object. So which of these 'normal causes' constitutes the content of N, according to NTT? Appealing to the visual system's response functions is of no help here: the system has the function f_1 to produce N in response to a certain pattern of retinal stimulation, but also the function f_2 to produce N in response to a certain pattern of light and the function f_3 to produce N in response to a small, dark, moving object (it performs f_3 *by* performing f_2, and f_2 *by* performing f_1). Thus the content of N turns out to be indeterminate. Neander (2013, pp. 33–35) tries to solve this problem by slightly modifying NTT, but it remains to be seen whether this modified account really yields plausible content ascriptions for all relevant cases.

The Content Question: Output-Oriented and Mixed Theories

The most discussed version of the output-oriented approach, and the most discussed version of teleosemantics generally, is Millikan's biosemantics. Millikan first presented her theory in *Language, Thought, and Other Biological Categories* (1984), but has extended and refined it ever since (cf. Millikan, 1989; 2004; 2009). According to Millikan, representations are basically signals passing from a producing mechanism ('producer' or 'sender') to a consuming mechanism ('consumer' or 'receiver'), and descriptive representations are those signals which have to vary systematically with conditions in the environment in order for

the consumer to be able to fulfill its functions. More formally, but still simplified, Millikan's teleological theory can be summarized as follows:

(MTT) R_i is a descriptive representation with the content that p iff (i) R_i belongs to a family of states R_1, …, R_n (the 'R-signals') which stand midway between two cooperating devices, a producer and a consumer, (ii) for the consumer to fulfill its biological functions in a normal way, different R-signals must correspond systematically to different conditions in the external world, and (iii) R_i must correspond to condition p.[2]

To determine the content of a descriptive representation, it is thus crucial to look at the biological functions of the *consumer* of that representation. Since the consumer is located 'downstream' from R, this makes MTT into a version of the output-oriented approach.

The difference between MTT and an input-oriented theory like NTT becomes clearer when we consider what MTT says about the frog case. Here, the relevant family of representations consists entirely of (possible) N-tokens occurring at different times: <N, t_1>, …, <N, t_n>. The producer of these states is the frog's visual system, and their consumer is the prey-catching mechanism – the mechanism responsible for the frog's snapping behavior. The main function of this mechanism is to catch prey, and thus ultimately to provide the organism with nutrients. Since the mechanism is activated by N-states, a covariation between N-states and small (i.e., swallowable) nutritious objects is necessary for the mechanism to fulfill its functions, that is, the representational states < N, t_1>, …, <N, t_n > must correspond to the conditions < small nutritious object present, t_1>, …, <small nutritious object present, t_n>. By contrast, it is irrelevant for the consumer's well-functioning whether the object in front of the frog is dark or moving: the prey-catching mechanism, once activated by N, would fulfill its functions equally well if the small nutritious object were light and stationary. (Of course, it would normally not *get* activated under these conditions, because N would not occur, but from Millikan's output-oriented perspective, this makes no difference to N's content.)

The same point can be put another way. In the evolutionary past, earlier tokens of the prey-catching mechanism have often provided their possessors with objects that were small, dark, and moving. Many of those objects were flies, and most of the flies were nutritious. But according to Millikan, it is only because the prey-catching mechanism provided its possessor with (small) *nutritious* objects that it was evolutionarily successful, so it acquired only the function of catching those objects, not the function of catching small, dark, moving objects or flies. Hence, MTT implies that an N-token occurring at t_1 has the content that there is a small nutritious object present at t_1 (cf. Millikan, 1991, p. 163).

Many objections have been raised against Millikan's proposal. A very influential criticism stems from Neander (1995, pp. 126–127). She argues that MTT, rigorously applied, yields content ascriptions that are *overly specific*. When we take a closer look at the frog case, for example, we find that the fact that really explains the evolutionary success of the prey-catching mechanism is not the fact that earlier tokens of the mechanism provided their possessors (often enough) with small nutritious objects *tout court*, but rather the fact they provided them with small objects that contained enough nutrients to make up for the calories lost in catching and digesting them, and that were, in addition, digestible, free from poison, not contaminated with deadly pathogens, and so on. So the mechanism must have the *function* to provide the organism with objects that have all these properties, and this entails, according to MTT, that all these properties enter into the content of N. To be sure, these consequences are highly implausible.

In recent work, Millikan has replied to Neander's objection by bringing in the functions of representation *producers* and the normal mechanisms by which they perform these functions (Millikan, 2004, pp. 85–86; 2009, p. 404), but it is not entirely clear how this reply is supposed to work in detail and whether it is consistent with Millikan's official output-oriented approach.

Even if the problem of overly specific contents can be avoided, another question remains: Does MTT entail content ascriptions that are *plausible*? Pietroski (1992) construes a case of hypothetical creatures called 'kimus' and argues that the content ascriptions entailed by MTT for this case are in conflict with our pre-theoretic intuitions. Defenders of MTT, however, can simply reject those intuitions as irrelevant (cf. Millikan, 2009, pp. 405–406). Consequently, Neander (2006) pursues a different strategy to attack the plausibility of the content ascriptions entailed by MTT. She considers the states that govern prey-catching behavior in toads, and argues that the content ascribed to these states by MTT (namely that there is a small nutritious object present at t_1, or something along these lines) is implausible from the perspective of mainstream cognitive science, since this content ascription does not fit with standard information-processing explanations of the toad's discriminatory capacities.

Not every teleological account of content can be neatly categorized as a version of either the input-oriented or the output-oriented approach. Some theorists defend mixed theories, holding that 'upstream' *and* 'downstream' factors are important for determining the content of descriptive representations. Carolyn Price (2001, pp. 89–103) and Nicholas Shea (2007, p. 418), for example, add to a broadly Millikanian theory of content the requirement that R must carry *information* about p in order to have the content that p. The addition of an input-requirement of this kind may help to rule out overly specific or otherwise implausible contents: one could argue, for example, that the frog's N-state does not carry information about the absence of pathogens in frog food, so that *not being contaminated with pathogens* cannot enter into N's content. Whether this strategy is successful depends, however, on the details of the theory, and especially on the definition of 'information' that is employed.

The Status Question

So far, the focus of this postscript has been on teleosemantic answers to the content question. But it is important to note that some theories also provide answers to the status question. Most prominently, Millikan's MTT specifies the conditions that are necessary and sufficient for conferring the status of a descriptive representation on R. These conditions include, besides clauses (ii) and (iii) which are also crucial for determining R's precise content, the condition that R must stand midway between two mechanisms, a producer and a consumer, which are designed to cooperate with each other (clause (i)). Thus, at least when it comes to representational status, Millikan's theory is not purely consumer- or output-oriented.[3]

Millikan's answer to the status question has been criticized in several different ways. Some theorists argue that MTT is *too restrictive*, because it requires representations to have cooperating producers and consumers (Sterelny, 1995; Stegmann, 2009). This requirement appears to exclude non-cooperative animal signals like the 'stotting' of gazelles, a behavioral display which (*prima facie*) indicates to approaching predators that they have been recognized, and this may seem problematic (for a reply on Millikan's behalf, see Artiga, 2014). Stegmann (2009) uses similar cases to motivate a purely output-oriented version of teleosemantics, where representational content *and* representational status are determined exclusively by consumer mechanisms and their functions.

Most critics, however, argue that Millikan's status requirements are *too liberal*. Price (2001, pp. 93–96) and Shea (2007, pp. 427–430) describe (hypothetical) organisms whose feeding behavior is triggered by a randomly generated internal state. Since food is abundant in the habitat of these organisms, this way of producing feeding behavior is evolutionarily successful. Millikan seems to be committed to treat the randomly generated internal state of such an organism as a descriptive representation of food, which does not seem adequate. Price and Shea suggest that this defect can be corrected by adding an informational input-requirement to MTT (see above).

But Millikan's status requirements may also be too liberal in a more fundamental way. Burge (2010) argues that Millikan's theory describes many simple systems as representational even though this description yields no explanatory benefits, thus drawing the 'lower border of representation' too low. He even suggests that this defect is shared by all teleological accounts of representation, but this assessment is arguably due to an impoverished conception of the resources available to the teleological approach (Schulte, 2015).

Conclusion

Explaining content in a naturalistic way is not an easy task, and for all we know, it might turn out to be impossible. But the naturalistic proposals that have been formulated and defended in recent years surely deserve serious consideration, and should not be dismissed out of hand. This goes for the theories presented above as well as for those omitted here like, for example, the teleological accounts of Dan Ryder (2004), Mohan Matthen (2005), and Manolo Martínez (2013), or the non-teleological account of Robert Rupert (1999).

Alternatives to Naturalized Semantics have also found new supporters in the past few years. Some adherents of the phenomenal intentionality paradigm accept that semantic content is a primitive feature of mental states, not capable of naturalistic explanation (e.g., Strawson, 2008), and proponents of radical embodied or enactive approaches to cognition deny that there are any contentful mental states at all (e.g., Noë, 2009). But neither of these alternatives has gained wide acceptance among contemporary philosophers. So if one thing is clear, it is that the discussion about Naturalized Semantics will continue for many years to come.[4]

Notes

1 Here and in the following, I'm disregarding a number of empirical details about the case, for example, the fact that the frog's perceptual states also represent the *location* of the object in the visual field.

2 Millikan (1984, pp. 96–97; 2009) spells out the last two conditions in terms of "mapping functions" or "rules of correspondence," but to introduce this terminology here would complicate the matter unnecessarily.

3 For a different take on the status question, see Papineau (2003).

4 I would like to thank Hannah Altehenger, Fabian Hundertmark, Insa Lawler, and Alex Miller for helpful comments.

References

Artiga, M. 2014. "Signaling without cooperation." *Biology & Philosophy*, 29(3): 357–378.
Burge, T. 2010. *Origins of Objectivity*. Oxford: Oxford University Press.
Dretske, F. 1988. *Explaining Behavior*. Cambridge, MA: MIT Press.
Martínez, M. 2013. "Teleosemantics and indeterminacy." *Dialectica*, 67(4): 427–453.
Matthen, M. 2005. *Seeing, Doing, and Knowing*. Oxford: Oxford University Press.
Millikan, R. 1984. *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.
Millikan, R. 1989. "Biosemantics." *Journal of Philosophy*, 86(6): 281–297. Reprinted in Stich and Warfield, 1994.
Millikan, R. 1991. "Speaking up for Darwin." In *Meaning in Mind: Fodor and His Critics*, edited by B. Loewer and G. Rey, 151–164. Oxford: Blackwell.
Millikan, R. 2004. *Varieties of Meaning*. Cambridge, MA: MIT Press.
Millikan, R. 2009. "Biosemantics." In *The Oxford Handbook of Philosophy of Mind*, edited by B. McLaughlin, A. Beckermann, and S. Walter, pp. 394–406. Oxford: Oxford University Press.
Millikan, R. 2013. "Reply to Neander." In *Millikan and Her Critics*, edited by D. Ryder, J. Kingsbury, and K. Williford, pp. 37–40. Chichester: Wiley-Blackwell.
Neander, K. 1995. "Misrepresenting & malfunctioning." *Philosophical Studies*, 79(2): 109–141.
Neander, K. 2006. "Content for cognitive science." In *Teleosemantics*, edited by G. MacDonald, and D. Papineau, pp. 167–194. Oxford: Oxford University Press.
Neander, K. 2013. "Toward an informational teleosemantics." In *Millikan and Her Critics*, edited by D. Ryder, J. Kingsbury, and K. Williford, pp. 21–36. Chichester: Wiley-Blackwell.
Noë, A. 2009. *Out of Our Heads*. New York: Hill and Wang.
Papineau, D. 1993. *Philosophical Naturalism*. Oxford: Blackwell.
Papineau, D. 1998. "Teleosemantics and indeterminacy." *Australasian Journal of Philosophy*, 76(1): 1–14.
Papineau, D. 2003. "Is representation rife?" *Ratio*, 16(2): 107–123.
Pietroski, P. 1992. "Intentionality and teleological error." *Pacific Philosophical Quarterly*, 73: 267–282.
Price, C. 2001. *Functions in Mind: A Theory of Intentional Content*. Oxford: Oxford University Press.
Ryder, D. 2004. "SINBAD neurosemantics: a theory of mental representation." *Mind & Language*, 19(2): 211–240.
Rupert, R. 1999. "The best theory of extension: first principle(s)." *Mind & Language*, 14(3): 321–355.
Schulte, P. 2015. "Perceptual representations: a teleosemantic answer to the breadth-of-application problem." *Biology & Philosophy*, 30(1): 119–136.
Shea, N. 2007. "Consumers need information: supplementing teleosemantics with an input condition." *Philosophy and Phenomenological Research*, 75(2): 404–435.
Stegmann, U. 2009. "A consumer-based teleosemantics for animal signals." *Philosophy of Science*, 76(5): 864–875.
Sterelny, K. 1995. "Basic minds." *Philosophical Perspectives*, 9: 251–270.
Stich, S., and T. Warfield, eds. 1994. *Mental Representation*. Oxford: Blackwell.
Strawson, G. 2008. "Real intentionality 3: why intentionality entails consciousness." In *Real Materialism and Other Essays*, pp. 281–305. Oxford: Clarendon Press.