Chapter 19

Naturalistic Theories of Reference

Karen Neander

Introduction

''Bill Clinton'' refers to the man, Bill Clinton, and ''Paris'' refers to the city, Paris. In philosophy of language, the term ''reference'' is sometimes used only for naming relations like these, but sometimes it is used more broadly to include other relations, such as the relation that holds between a kind term (e.g., ''cats'') and its extension (all cats) or between a predicate (''red'') and a property (redness). In philosophy of mind, the term ''reference'' is usually used in the broader sense. On a representational theory of thought, my thought that cats are excellent hunters, or my thought that cats make me sneeze, involves a mental representation of cats, and cats are said to be its reference (or its referential or extensional content). Most broadly, a theory of reference is an attempt to describe the relation between a representation and what it represents. That is, it aims to describe what it is about the former in virtue of which it represents the latter.

Naturalism is an approach to philosophy that involves using science, ultimately physics, as our guide to the fundamental ontology of the universe. In practice, with respect to theories of reference, this amounts to not admitting such things as moral norms, mental states or semantic properties as fundamental, so that any appeal to them in an analysis of the reference relation must eventually be accounted for in other terms. Three alternatives to a naturalistic theory of reference are (1) to maintain that reference is not real (e.g., talk of reference is merely instrumental), (2) to maintain that reference can be analyzed in terms of fundamental but non-natural phenomena (e.g., Platonic concepts and propositions) and (3) to maintain that reference itself is fundamental (so that physicalism is false). As the title suggests, this entry looks only at naturalistic theories of reference.

Original and Derived Meaning

Naturalistic theories of reference tend to focus on mental representations because those who offer them tend to believe that the semantic properties of linguistic utterances ultimately derive from the semantic properties of mental representations. This is consistent with the derivation going to some extent in both directions; so social factors, including a community's linguistic conventions, might determine the reference of some mental representations, but it is generally thought that this presupposes certain psychological capacities (perception, memory, learning, and so on) that are themselves representational.

Furthermore, naturalistic theories of reference focus on the most basic (or simple) mental representations. Here, a basic representation is one that does not derive its referential content from that of other representations. The controversial classical view of concepts (e.g., Katz 1972) illustrates the distinction between basic and non-basic representations. On one version of this view, lexical entries in the semantic component of our linguistic system provide definitions for linguistic terms and their corresponding mental representations. For example, the lexical entry for "spinster" and for the corresponding mental representation (here denoted by the English term all in capitals) SPINSTER, might be ADULT, FEMALE PERSON WHO HAS NEVER MARRIED. On this view, the reference of SPINSTER is the intersection of the references of ADULT, FEMALE, PERSON, and NEVER MARRIED, so that the reference of the first is allegedly determined by the references of the last four. The representations used in the definition may in turn be defined but at some point this process must bottom out. On this view, some mental representations are basic; they are not defined and they do not derive their referential content from that of other representations. These mental representations must, according to the classical theory, derive their referential content in some other way.

Philosophers disagree as to which mental representations are the basic ones, the simples. They also disagree about the nature of the relation between the simples and other mental representations. As already remarked, the classical theory of concepts is controversial. However, most who offer naturalistic theories of content agree that some mental representations are basic, and that some of them must therefore possess their content without deriving it from the contents of other mental representations. Some have argued that virtually all mental representations corresponding to the morphemes of a natural language are simple (e.g., Fodor and Lepore 1992). Whereas others have supported the idea that there is a restricted set of simples which form the constitutive base for the references of other representations (e.g., Devitt 1996; Prinz 2002).

The fewer simples there are, the easier it is to give a first-stage naturalistic theory that accounts only for the referential content of the simples. There are, however, serious objections to a two-stage approach, which first explains how the simples

refer and then explains how complex concepts derive their referential powers from the simples. One is based on Quine's (1953) claim that we lack a principled analytic-synthetic distinction; the idea being that the second-stage requires a distinction between meaning constitutive relations and other relations among representations. There are a number of replies to Quine, but those who think that complex concepts can be constructed out of simpler concepts must also face Kripke's (1980) critique of description theories of reference.

Kripke's argument from ignorance and error is most relevant here. Kripke argued that we could be mistaken in almost everything we think we know about Aristotle and yet succeed in referring to Aristotle, and that even when no one knew the chemical analysis of water, our term "water" (or its cognate) referred to H_2O and only to H_2O. If so, not even implicit knowledge of some true definite description of Aristotle is needed to refer to Aristotle, and not even implicit knowledge of the essential properties of water is needed to refer specifically and exclusively to water. While Kripke's argument was directed at description theories of meaning for linguistic terms, his point generalizes. First, it seems to generalize to mental representations: a thinker, as well as a speaker, need have no mental description that accurately and exclusively characterizes (e.g.) water in order to use WATER to think exclusively and specifically about water. And second, the point seems to raise difficulties for more than description theories. It seems that *no* inner characterization of water, which accurately and exclusively characterizes water, is required for WATER to refer to water. Thus, whether the structure of the concept takes the form of a definition, a prototype, an exemplar, or a theory, it would seem that the content of component concepts could not be what determines the reference of WATER.

Some conclude that reference is always or almost always "atomistic." "Atomism" is variously defined, but here the reference of a mental representation is atomistic if it is determined independently of the references of other mental representations. Thus, to claim that reference is always or almost always atomistic is to claim that all or almost all concepts are simples. However, it would be too quick to come to this conclusion on the basis of the foregoing. For one thing, we need to consider what range of concepts the Kripkean point applies to, and for those within the range, we need to consider whether their reference might be partly determined by the references of other representations without being wholly so determined.

This issue complicates the assessment of naturalistic theories of reference. Different philosophers have alleged, against one or another such theory, that it cannot account for our reference to, say, non-existent objects (e.g., Santa and Satan), entities that cannot affect us or have not yet affected us (e.g., spatially very distant or future events), or entities that have diffuse impacts on us (e.g., the Big Bang or electrons). Such objections can succeed only if the corresponding representations (SANTA, SATAN, etc.) are simples, or are at least alleged to be simples by the target theory. If they are not simples, then accounting for their reference is a joint venture involving a naturalistic theory of simples plus an account of reference derivation for complex concepts.

The Causal-Historical Theory

What natural properties can serve to ground reference? Kripke's critique, along with Putnam's (1975) Twin-Earth thought experiment, have suggested to some that reference must involve causal relations between thinkers and what they think about. Twin-Earth is an imaginary planet somewhere in the universe that is as like Earth as possible, except that wherever there is H_2O on Earth there is an alien stuff on Twin-Earth, designated 'XYZ.' Liquid consisting of XYZ is indistinguishable from liquid consisting of H_2O, short of chemical analysis: it tastes the same, looks the same, quenches thirst, falls from the sky, etc. Everyone on Earth has a Twin-Earth doppleganger, including Oscar, whose Twin-Earth counterpart is Twin-Oscar. Neither knows about the chemical composition of the liquids on their planets, and nor do their respective communities, and Putnam claims that, nonetheless, Oscar's English term "water" refers to H_2O and not XYZ and that Twin-Oscar's Twin-English term "water" refers to XYZ and not H_2O.

One implication, says Putnam, is that meaning has two components: an internal one, which is the same for both "twins," and an external one – referential content – that is different. Since the twins are doppelgangers they possess the same internal characterization of what they call "water," and the environment must be responsible for the difference in referential content. But the mere fact that there is H_2O in Oscar's environment whereas there is XYZ in his "twin's" is not enough to account for the difference. This claim is supported by the intuition that if Oscar and Twin-Oscar were exchanged without their knowledge (e.g., by teletransportation during sleep) they would still have thoughts about water (H_2O) and Twin-water (XYZ) respectively. If so, this suggests that a history of causal interaction between a thinker and what the thinker thinks about may be required to fix reference.

Neither Putnam (1995) nor Kripke (1982) support the naturalism project but, with Donnellan (1970), they are responsible for sketching a causal-historical theory of referential content. The main idea of a causal-historical theory is that reference begins with an act of dubbing that involves a perception of or a description of the referent, which establishes a chain of reference as the name of the item is passed from one person to another in communication. For example, an infant, Richard Feynman (to use one of Kripke's examples) is named at birth, and his parents talk about him and introduce him to others and, as he matures and becomes known, word of him spreads. Someone might hear of him and later refer to him without remembering anything that is both true and distinctive about him. What is required, on this theory, is that we partake in a "... certain passage of communication reaching ultimately to the man himself" (Kripke 1980: 91). This passage of communication involves a causal chain, from speaker to speaker, and thinker to thinker.

A causal-historical theory seems to presuppose intentional psychological states that remain in need of analysis, as Kripke recognized. For example, the dubbing

ceremony seems to require an intention to name something, and what is named seems to depend on the content of the dubbing intention: e.g., on whether Feynman's parents intended to name him, a time-slice of him, or his crib, or babies in general. In addition, not just any passing on of a name will do. Reference borrowing needs to be distinguished from 'naming after' (as in the case of naming a pet after a famous person). Reference borrowing seems to require certain intentions, such as the intention to use the name to refer to the same individual as the one referred to by the person from whom the name was received.

Devitt (1981) has argued that a causal-historical theory can be naturalized if it is articulated in terms of causal relations of the right kind, although it will then be incomplete. Among other things, it will lack a solution to the ''qua-problem'' (whether what was named was Feynman, time-slices of Feynman, and so on). Devitt argues that it might nonetheless be part of an overall naturalistic theory.

The Crude Causal Theory

The causal-historical theory is not a complete fundamental theory, but perhaps further appeal to causal relations between representations and their representeds can ground reference (Stampe 1977). The crude causal theory (too crude to have been held by anyone) often serves as a starting point in thinking about how to further naturalize reference. The crude causal theory says that representations of a given type refer to the causes of its tokens: e.g., tokens of the type SKUNK refer to skunks if skunks and only skunks cause SKUNKs.

This theory has many problems, and as a further preliminary it will be useful to mention some of them. First is the *problem of error*. If I see a long tailed weasel striped from walking under a newly painted fence, I might think, ''there goes a skunk.'' I'd be mistaken, but the crude causal theory does not give that result. Instead, it entails that once a non-skunk has caused a SKUNK, SKUNKs refer to skunks *and* some non-skunks (e.g., some painted long-tailed weasels). Error is impossible on the crude causal theory. Second is the *problem of distal content*. When a skunk causes me to token SKUNK, so do light rays reflected from the skunk, as well as neural firings en route to my visual cortex. There is a causal chain involved in the causing of SKUNKs, not a single cause, and the crude theory does not identify the part of the chain that is referred to. Third is *the problem of intentional inexistence*: the represented may not be the kind of thing that can cause a representation because it might not exist. No unicorn has ever caused a UNICORN because there are no unicorns. Fourth is *the problem of the absent represented*. Even in cases of correct representation, the represented may not be among the causes – or at least, not the immediate causes – of the representation. Your talk of your pet dog might remind me of my long-dead pet cat, in which case your talk is the more immediate cause of my representation. But my FLUFF represents my pet cat and not your talk of your pet dog for all that. Fifth is the

qua-problem again: when a skunk causes a token of SKUNK, so might an undetached part of it (e.g. its tail or its face), or a spatio-temporal slice of it.

There will not be space to discuss how each theory, outlined below, aims to deal with each of these problems. The following sections focus mainly on their responses to the problem of error, the problem of distal content, and the problem of intentional inexistence, but this should not be taken to suggest that the other problems are less important.

Another problem that gets mentioned in this context is *the fine-grainedness problem*. The issue is that content seems to be more fine-grained than causation is. Our thoughts can distinguish between metaphysically co-extensive properties such as being triangular and trilateral or being a rabbit and being a collection of undetached rabbit parts but, it is alleged, causation cannot. However, it is not clear that this is a problem for a naturalistic theory of *reference*. On some views, if 'X' and 'Y' refer to properties that have identical causal powers, then X and Y are one and the same property. On this view, if we cannot discriminate between the causal powers of triangularity and trilaterality, TRIANGULAR and TRILATERAL refer to the same property. So, if they differ in meaning, the difference must be elsewhere, such as in Putnam's internal component, in their inferential role perhaps.

The Asymmetric Dependency Theory

Fodor (1990b) aims to solve the above-mentioned problems. He starts with the problem of error. We solve this problem, he suggests, when we distinguish the *right* from the *wrong* causes, and he adds that the wrong causings are dependent on the right causings and not *vice versa*. More carefully, representations of a type, *R*, refer to *X*s and not non-*X*s if non-*X*s can cause *R*s only because *X*s can, and not vice versa. It's because skunks can cause SKUNKs that non-skunks can too, and not vice versa, and that's why SKUNKs refer to skunks. According to Fodor, there's one ceteris paribus causal law to the effect that skunks can cause SKUNKs, another to the effect that (e.g.) some long-tailed weasels can too, and a dependence between the two that is synchronic, not diachronic, such that the second law depends on the first.

Some philosophers have found this theory hard to test. The theory requires that one-kind-of-event' s-causing-*R* (where "*R*" stands for a representation) depends on another-kind-of-event's-causing-*R* and these higher-order dependencies are somewhat mysterious. On occasion, Fodor explains them in terms of counterfactuals: in the nearest possible world where skunks cannot cause SKUNKs, painted long-tailed weasels cannot either, but in the nearest possible world where painted long-tailed weasels cannot cause SKUNKs, skunks still can. As Fodor puts it, the link between SKUNKs and skunks is more *resilient* than the link between SKUNKs and non-skunks. Some have proposed counter-examples based on possible world scenarios, but Fodor argues that they are mistaken about

which content is correct or about the direction of dependency (see e.g., the papers by Baker, Block and Boghossian, as well as Fodor's replies, in Loewer and Rey 1991). This debate might show that our intuitions in these cases are too malleable to be useful, and/or that the theory needs to be more fully specified.

Another problem for the asymmetric dependency (ASD) theory is the problem of systematic error and ignorance. The problem arises because the theory requires that thinkers have certain synchronic dispositions such that the relevant asymmetric dependencies hold true of them. It is neutral as to how they hold, but hold they must. ASD requires that, if we think of skunks, we must be disposed to have skunks cause SKUNKs in us – otherwise, non-skunks could not cause SKUNKs in us only because skunks can. The problem is that we can think about skunks even if we lack the ability to recognize them – even if we are not disposed to have skunks cause SKUNKs in us. Further, ASD seems to entail that, if I misconceive the nature of skunks, my mental representation of skunks will not refer to skunks but to the kind that best matches my conception (i.e., my misconception of what a skunk is like).

Fodor maintains that the ASD theory allows reference to unicorns, since the relevant counterfactuals hold in nearby possible worlds: if there were unicorns, they would cause UNICORNs, and non-unicorns could only cause UNICORNs because unicorns would. But this move, which might be essential for accommodating reference to non-existent entities, has a high price. For now suppose that Tom believes that skunks have a sweet scent and pink polka dots. Were there such a creature – call it a squunk – it would cause him to token SKUNK. Worse, a skunk would cause him to token SKUNK only if he mistakenly thought it had squunky features. In other words, skunks could cause SKUNKs in Tom only because squunks could. So the squunk-to-SKUNK connection seems more resilient for Tom than the skunk-to-SKUNK connection.

One could invoke Fodor's ceteris paribus clause to maintain that the asymmetric dependencies pertain to suitably well-informed people, but this invokes something intentional, which undermines the naturalistic aims of the theory. It might also be circular, for we must ask what is it to be suitably well informed. Is being suitably well informed about skunks a matter of having the capacity to recognize skunks as skunks? If so, it seems we must peek at the content of SKUNK to apply the ceteris paribus condition. One might also claim that the skunk-to-SKUNK connection is more resilient than the squunk-to-SKUNK connection, even for someone as benighted as Tom, because Tom would respond to appropriate instruction by revising in favor of the skunk to SKUNK connection. But this also invokes something intentional and apparently circular. Instruction is an intentional notion, and Tom will only revise appropriately if he is instructed correctly. A better response might be to suggest that SKUNK is not an apt subject for a fundamental theory of reference, on the grounds that it is not plausibly a simple. While Fodor's atomism may not allow that response, others may want to consider its merits.

Fodor thinks his ASD theory has a good shot at dealing with the absentee problem (Fodor 1990). Basically, the claim is that cases of absentee representation depend asymmetrically on cases where the represented is present. He also offers a

solution to the problem of distal content. Consider a case where a SKUNK has a proximal cause that is a certain pattern of retinal firings, *RF*. Fodor claims that *RF* can cause SKUNKs to be tokened only because skunks can too. On the face of it, this seems wrong, since skunks cannot cause SKUNKS except through the mediation of more proximal causes, including *RF*. However, Fodor points out that skunks can cause SKUNKs without the assistance of *RF* in particular, since any number of patterns of retinal firings could mediate between a skunk and a SKUNK. Under the right circumstances, a mere sniff, or glimpse of brown fur, or rustle in the grass, could cause us to token SKUNK. Indeed, says Fodor, there is no closed disjunct of retinal impressions or sensory inputs – *RF* or RF^1 or RF^2 or $RF^3 \ldots RF^n$ – that could satisfy the ASD theory's requirements for referential content. While this might be true, we might still ask if it goes to the heart of the matter. Suppose we discover that a certain neural process immediately prior to SKUNK tokening *is* required for SKUNK tokening. This seems possible, but we would not then conclude that SKUNK referred to this more immediate neural process (Loar 1991).

Teleosemantics

The naturalistic theories on which there has been most work done of late are the teleosemantic theories. These are a diverse class of theories that share the claim that the contents of mental representations are determined (somehow) by the functions of the systems that (it depends on the version) produce or use the representation.

The relevant notion of function is given an etiological analysis (following Wright, 1976). According to such an analysis (e.g., Neander 1991) items of a type have the function of doing what that type of item was selected for doing. For example, our pineal glands have the function of secreting melatonin at nightfall because their doing so in the past contributed to their past selection. In the case of innate representational capacities, the relevant selection process is neo-Darwinian natural selection, so that the function of a system is to do whatever ancestral systems did which caused systems of that type to be preserved and/or proliferated in the population. These functions are referred to as ''normal functions'' or ''proper functions,'' following biological talk of the proper functioning or normal functioning of a system.

Theories that appeal to this notion of function are known as ''teleological theories'' because the notion has a teleological flavor. To say that hearts have the function of pumping blood seems equivalent to saying that they are *for* pumping blood or even (metaphorically) that their *purpose* is blood pumping. On an etiological analysis, they *are* literally *for* pumping blood in so far as this is what they were selected for. However, the relevant notion is not literally purposive, and if mental content is to derive from past selection, it must ultimately derive from non-intentional processes of selection, such as neo-Darwinian natural selection.

However, according to some proponents of teleosemantics, other forms of selection can also serve to ground appropriate (content determining) functions: for example, there is also talk of meme selection, cultural selection, and learning or conditioning doing so.

The relevant notion of function is also said to be normative in the sense that it permits the possibility of malfunction: a token trait can have a function that it lacks the ability to perform. For example, my pineal gland can have the function of secreting melatonin at nightfall even if it is unable to do so. If it cannot secrete melatonin because it is malfunctioning, we might say that it is "supposed to" secrete melatonin, nonetheless. However this is cashed out in descriptive rather than prescriptive terms on an etiological theory, for all it means on such a theory is that pineal glands were selected for doing so. Even if my pineal gland cannot secrete melatonin, it belongs to a homologous type that was selected for secreting melatonin, and thus secreting melatonin is its function. Some prefer to reserve the term "normative" for prescriptive contexts only. According to this more restrictive use of the term, only statements that entail ought-claims without the addition of further premises count as normative statements. Function ascriptions are not claimed to be normative in this sense by those who advocate teleological theories of content. On an etiological analysis of proper function, no ought-claims follow from function ascriptions alone. (Consider that some HIV genes are adapted for disabling our immune system, but it doesn't follow that they should do so, or that we should help them to.) Content ascriptions might also be normative in only the more liberal, descriptive sense. Indeed, those who try to naturalize content usually assume that this is so. Representations can misrepresent, beliefs can be true or false, and so on, but content ascriptions might not entail ought claims without the addition of further premises. If so, the attempt to naturalize semantic norms is not an attempt to derive ought-statements from is-statements.

Teleosemantic theories are offered by among others: Dennett (1969: ch.9; 1987: ch.8; 1995: ch.14); Dretske (1986, 1988, 1991); Fodor (1990a), although he repudiated his offering long before it was published; Israel (1987); Jacob (1997); Matthen (1984); McGinn (1989: ch. 2); Millikan (1984, 1989, 1991, 2000); Neander (1995, forthcoming); Papineau (1984; 1987; 1993: ch. 3); Price (2001) and Sterelny (1990). The theories of Papineau and Millikan were the earliest detailed versions of teleosemantics and these are outlined in this section. Some of the above theories are very different to these and are discussed in the last section.

Millikan maintains that the content of a representation is determined, not by *its* function, but by the functions of its consumers – the systems that used the representation to perform their proper function. These consuming systems may or may not be cognitive systems. A much discussed example is the frog's response to anything appropriately small, dark, and moving past its retina by flicking out its tongue and attempting to catch and swallow it (Lettvin et al.: 1959). One consumer of its perceptual representation is its digestive system, another is the neural components of the frog's brain that control its tongue snapping. Contents,

Millikan claims, concern the conditions required for the consumers to perform their function in the historically normal way. That is, it concerns the conditions required for them to do what traits of the type did in the past when they contributed to their own selection on those occasions when the representation was used in doing so. Specifically, Millikan maintains that the frog's perceptual representation represents frog-food (and not, say, something small, dark, and moving) because it was only when frog-food corresponded to the representation that the frog's digestive system or its tongue snapping mechanism succeeded in feeding the frog and hence succeeded in contributing to the preservation and/or proliferation of such systems in frogs. On this account, misrepresentation therefore occurs when the stimulus is not frog food (e.g., when the frog snaps at a BB, a small plastic pellet).

Like Millikan, Papineau maintains that content is determined by the past use of representational states and by the environmental conditions that obtained when their use contributed to fitness. For Papineau, a desire's satisfaction condition is "... that effect which it is the desire's biological purpose to produce" (1993:58–9). By that he means that "[s]ome past selection mechanism has favored that desire – or, more precisely, the ability to form that type of desire – in virtue of that desire producing that effect' (1993: 59). The main function of beliefs, he maintains, is to collaborate with desires to cause behavior that bring about their satisfaction conditions. The truth condition of a belief, he tells us, is the condition that must obtain if the desire with which it collaborates in producing an action is to be satisfied by the condition brought about by that action. A desire that has the function of bringing it about that we have food has the content that we have food, since it was selected for bringing it about that we have food. If this desire collaborates with a belief to cause us to go to the fridge, the content of the belief is that there is food in the fridge if our desire for food would only be satisfied by our doing so if it is true that there is food in the fridge.

As Fodor (1990b) has argued, it is a problem for this formulation that some desires do not or cannot contribute to their own satisfaction (e.g., the desire for rain tomorrow and the desire to be immortal) and that others would not have been selected for any such contribution (e.g., novel desires, or an adolescent's desire to kill himself or herself).

Of course, Papineau and Millikan know that some mental representations are not "innate," or the direct result of ordinary natural selection. Papineau claims that learning (he seems to have conditioning in mind) is a process that is sufficiently similar to ordinary natural selection (Papineau 1993: 59–67) to ground content. But it is not plausible that we learn all of our new beliefs by something akin to natural selection (e.g., I might acquire a new belief by reading a sentence in a book). Millikan appeals to what she calls "derived" proper functions and to mapping rules that, when applied in new contexts, furnish novel representations (Millikan 1984: 41–3). To illustrate the notion of a derived proper function, Millikan mentions a mechanism in a chameleon, which she says has the proper function of matching the color of the chameleon to the color of the surface on

which it sits (within a certain range of colors). Millikan says that if a particular chameleon is sitting on a particular shade of green, the mechanism has the derived proper function of matching that particular chameleon to that particular shade of green, even if no chameleon has happened to sit on that particular shade of green before. Millikan claims that, along similar lines, belief-producing mechanisms (and belief-consuming mechanisms) can have the derived proper function to produce (or consume) particular beliefs, including novel beliefs never believed before.

Exactly how, or whether, this idea could determine contents for sophisticated representational states (e.g., beliefs in humans) remains obscure. But perhaps an analogy with literal mapping might help. Suppose we (intentionally) select a system for mapping terrain. This involves our selecting certain mapping rules. However, once we have done so, we can apply these rules to generate indefinitely many maps that represent indefinitely many terrains. We do not need to select a fresh set of mapping rules for each new map. Along similar lines, the idea might be that natural selection can select certain mapping rules (or respects of isomorphism), which a cognitive system can then exploit to produce indefinitely many novel representations. Millikan's response to the problem of novel concepts (e.g., the concept of an electron, when the concept was first proposed) is along the same lines. (Readers should also see Millikan's newer treatment of these issues, in Millikan 2000.)

Another option, to return to points made earlier, is to claim that teleosemantics only determines the reference of representational simples, and that no novel concept is a simple. This approach at any rate seems unavoidable for non-existent objects, which cannot be the reference of simples on a teleological theory. Those who support teleological theories of mental content can also opt for a combinatorial semantics, thus avoiding the problem of novel, impotent and destructive desires, mentioned above.

On the surface, Papineau and Millikan have a solution to the problem of distal content – i.e., of determining what is represented from among the items involved in the chain of more proximal and distal causes of a representation. In the case mentioned above, neither light rays nor neural firings, in the absence of frog food, feed the frog or contribute to fitness. Nor do they satisfy our desire for food when we walk to the fridge. However, on closer inspection, it is not clear whether the problem is solved. The question to ask, according to Millikan, is not "What feeds the frog?" but rather, "What condition was required for the consumers of the representation to perform their function in the historically normal way?" The problem is that not just food was required. Food that goes undetected is of no use to a frog and its detection requires that light hit the frog's retina. A similar problem arises for Papineau. The mere presence of food does not satisfy our desire for food, since the food must be perceived and eaten, and so it must stimulate our retinas or our olfactory nerves and pass into our gut. We return to this in (2) below.

Two main objections to teleosemantics are (1) the Swampman objection and (2) the so-called functional indeterminacy problem. (For discussion of lack of

epistemic access to selection histories and the theory's alleged commitments to adaptationism, see Fodor 1996, Peacocke 1992, and Neander 1999).

(1) Swampman-style examples have been around for a while (see Boorse 1976). But Swampman, in particular, is a creature of Donald Davidson's imagination. Swampman begins his existence as a molecule-for-molecule synchronic doppelganger of Davidson at time *t*. He comes about as a result of a purely random collision of elementary particles. Crucially, he is not a copy of Davidson in any causal sense; the resemblance between them is just a stupendous coincidence. Moreover, Swampman lacks any selection history whatsoever: he has no evolutionary history and at the start at least no learning history. The troublesome intuition is that when Swampman first pops into existence, he has thoughts, perceptions, and so on, just like Davidson's at *t*, whereas according to teleosemantics, he lacks all such intentional or representational states, since his systems lack the appropriate function-conferring histories.

In defense of teleosemantics, a number of points have been made. One is that teleosemantics is usually meant as an account of what referential content *really is*, not what we conceive of it as being. Were he to exist, Swampman's states would superficially resemble states with referential content. Everyone can agree on that. But it can be argued that a deeper analysis can show that Swampman would not have the same kind of states as we have. A deeper analysis could show that content requires a selection history. Just so, were XYZ to exist, it would superficially resemble water. But a deeper analysis of water has shown that the XYZ liquid would not be the same kind of thing as water.

Some respond to this by maintaining that what is of most interest, in the case of referential content, is the larger category that includes our referrings as well as Swampman's analogous call-them-what-you-will. However, this shifts the subject, since we have been interested in the norms of reference, and according to teleosemantics the norms of reference depend on history. So, unless Swampman really has referential content, Swampman is beside the point. In support of the intuition that Swampman really has referential content, the most difficult intuition to resist is that it will *seem* to Swampman that he is thinking about all sorts of things. And if things seem a certain way to Swampman, it would appear that he has contentful states – if it seems to him that he is thinking about cats, then presumably he at least has a representation that refers to his thought, or at any rate to his Swampish equivalent. Again, however, proponents of teleosemantics can claim that appearances could be deceptive, and that Swampman's seemings would not be the same kind of thing as our seemings.

Most proponents of teleosemantics reject the idea that we should care about Swampman intuitions. It would be enough, they claim, if we could find a theory of referential content that was successful for real creatures. While this would certainly be an achievement, some argue that we should still care about Swampman, on the grounds that scientific classifications should be based on similarities and differences in causal powers (e.g., see Antony 1996). If we classify mental states according to their content, and content depends on history, then we do not classify

mental states on the basis of their causal powers, since two states can have the same causal powers but different histories or the same histories but different causal powers. This introduces a debate called the "Methodological Individualism debate." One problem with Methodological Individualism is that it is radically revisionist. Biology has many historically based classifications (e.g., species, clades, and physiological kinds, which are often based on function or homology or both). If, for instance, one were to classify Swampman's "kidney" with Davidson's, on the basis of their having the same causal powers, one would do so at the expense of excluding the (malfunctioning) kidneys of many real people. (For further discussion of Swampman, see the essays in the issue listed under Antony 1996, plus Braddon-Mitchell and Jackson 1997, and Papineau 2001.)

(2) The second objection to teleological theories of content that will be considered here is the objection that function ascriptions are too indeterminate to determine content. Consider the toad, similar to the aforementioned frogs (Ewert 1983). Motivated toads will hunt and try to eat anything with a suitably worm-like configuration: i.e., they will hunt and try to eat anything that's, roughly, small, elongated and moving parallel to its longest axis. They cannot discriminate between toad food and cardboard cutout rectangles with the right configuration, but in their ancestral environment, things with that configuration were often enough toad food, so a device that responded to these features sufficed for the job. We can describe the function of the relevant part of the toad's perceptual system in different ways: e.g., as detecting toad food and as detecting things with the right configuration of features. What does the toad's perceptual representation represent then? Does it misrepresent the cardboard cutout as a toad food? Or does it correctly represent the cardboard cutout as an item in worm-like motion?

As Fodor (1990b) sees it, the problem is that natural selection cannot discriminate between co-extensive features. If it's adaptive to snap at *F*s and *F*s are co-extensive with *G*s, then it's equally adaptive to snap at *G*s. This leaves it up to us how we choose to describe the function of the toad's systems. We can equally well describe them as having the function of snapping at *F*s or at *G*s. And if function ascriptions are indeterminate, they cannot determine content. We have not naturalized content if the content depends on our choice of description. The standard reply to this version of the problem is that something *can* be selected *for* occurring in the presence of *F*s and not for occurring in the presence of *G*s if it was *F*s and not *G*s that played a causal role in the selection. Selection *for* is a causal notion. A type of item is selected for doing that which caused the type of item to be preserved and proliferated in the population. (See Fodor 1996, for his updated version of the problem, and Neander 1999, for a reply.)

Nonetheless, a problem remains because traits are selected for complex causal roles, and so more than one property can be causally efficacious in selection (Neander 1995). Components of the toad's perceptual system were selected for helping to feed the toad *by* detecting a certain configuration of visible features, and so both properties of the stimulus – both its being nutritious and its having a certain configuration of visible features – played a causal role in this case of

selection. A problem remains for teleosemantics because we can obtain different function ascriptions by focusing on different aspects of the complex causal roles for which traits were selected. And so it remains true that if we must choose among different descriptions of the relevant function, we have not naturalized content by appealing to functions.

Some think that teleosemantics must isolate a unique correct function ascription if teleosemantics is to work (Enc 2002). And some have tried to do so by adding further conditions to an etiological analysis of functions (e.g., Price 2001). However, proponents of teleological theories of content need not insist that function ascriptions are determinate. That is to say, they can allow that traits are selected for complex causal roles, and that the functions of traits can be described in different ways by focussing on different aspects of these complex causal roles. They can do this because they can appeal to other things in addition to functions that can make content more determinate.

Millikan claims that her focus on the consumers of representations helps. Only if the toad gets fed do the consumers of the prey-representation perform their proper function in the historically normal way, she maintains. However, the appeal to consumers is no help since the functions of consumers are just as complex (and hence "indeterminate") as the functions of producers are. The function of the frog's tongue-snapping mechanism, for example, can be described as snapping at frog food, or as snapping where and when the frog's brain tells it to snap.

What might be doing more real work for Millikan is an implicit reliance on what was *most* crucial for a contribution to fitness (see Millikan 1991). In the past, for a contribution to fitness to occur, it was Normally (in the teleological sense) required that the stimulus have the right configuration of features, for otherwise it would not have been detected, *and* that it be nutritious, otherwise it would not have fed the frog. But the latter was more crucial than the former in the following sense. If the toad could have had the toad food minus the right configuration of features, it would have been perfectly fine, whereas if it could have had the right configuration of features minus the toad food it would have starved to death.

Both Millikan and Papineau might also try to appeal to what was, in this sense, most crucial for fitness, to try to handle the problem of distal content. What was most important for fitness: food or the light reflected from it? By the same reasoning, the answer is food, for if we could have had the food without the light all would have been well, whereas if we had only the light without the food, we would have starved to death. The difficulty here is that if we press the point we can go too far. If we could have had digested nutrients in the gut, without the food outside, all would have been well also, but if we had only the food outside, without the digested nutrients in the gut, we would have died. Digested nutrients are distal, but not appropriately distal.

Hall (1990) argues that Millikan's theory leads to overly specific content: a contribution to fitness didn't just require food. The food must have contained no deadly toxins, no viruses, bacteria, or parasites that tipped the balance between costs and benefits. In the toad's case, it also required that no fishing line be

attached and no crow lurking nearby, and so on and so forth. All this, it seems, must be included in the content. Pietroski (1992) also argues that the content generated is the wrong content. He asks us to imagine creatures, called "kimu," that are initially color blind. Due to a mutation, one kimu is able to see red and is attracted to the red glow of sunrise. It ascends the top of the nearest hill each morning to see the rising sun. By doing so, it happens to avoid the dawn-marauding predators below. The trait would be selected because it guided the creature and its descendents to predator-free spaces. Pietroski argues that, in this case, Millikan's theory gives the wrong perceptual content because it would not allow him to tell the story this way. Intuitively, the creatures see red and love to see red. But according to Millikan, they do not see and love to see red. They instead see and love to see predator-free spaces, even if they have never in their lives seen these predators (for Millikan's response, see Millikan 2000: appendix B). Neander (forthcoming) also argues that standard teleological theories of content generate the wrong content ascriptions. Her argument is based on an analysis of the kind of content needed to play a role in mainstream cognitive science. A careful look at neuroethological explanations of frog and toad perceptual systems, she argues, supports the view that standard teleosemantics cannot serve the purposes of information processing explanations of cognitive capacities.

Informational semantics

A different style of teleological theory is a theory that links contents less directly to contributions to fitness. On such theories content concerns the information that representations are supposed to carry, where the "supposed to" is teleological, or a matter of what something was selected for.

Dretske's theory (1981, 1986, 1988, and 1991) is the best known of these. He defines an informational relation, called "indication." Events of one kind, *R*, *indicate* events of another kind, *C*, just in case (within the relevant environment) if there is an instance of *R* then there is an instance of *C*. Dretske tells us that *C* need not be a cause of *R*; *C* and *R* might have a common cause, for instance. Nor need *C*'s connection with *R* be nomological. To use one of Dretske's examples, if there is someone ringing the doorbell whenever the doorbell rings, its ringing indicates that someone is at the door, even though there's no law that doorbells ring only if someone rings them. If squirrels start to ring doorbells because people start making them out of nuts, it would no longer be the case that doorbell ringings indicate that someone is at the door.

Mere indication is not sufficient for representation. There can be no error if representation is equivalent to indication, as Dretske explains, because "*R* indicates *C*" is incompatible with "*R* and not-*C*." So Dretske suggests, at a first pass, that representations are items that have the function of indicating. Since items don't always perform their function, misrepresentation would then be possible.

The initial proposal is that *R represents C* only if *R*s were recruited for indicating *C*s and for bringing about some behavior, *M*. It follows that *R* need only indicate *C* during recruitment and error is possible after that time and in other environments.

Dretske (1986) maintains that content is determinately distal in creatures with a capacity to acquire an indefinite number of epistemic routes to the same representation, thus ensuring that there is no time-invariant closed disjunct of proximal features that the representation was recruited to indicate. This is similar to Fodor's response, and it suffers from the same problem (see also Loewer 1987). In addition, it denies distal content to representations based on innate representational capacities.

Another problem stems from Dretske's stringent definition of "indication," according to which *R*s indicate *C*s only if "if *R* then *C*" (in the relevant environment) has a probability of one. This seems to force Dretske's theory to rely on an unrealistic distinction between recruitment and post-recruitment phases and places. To see this, note that nothing can be selected by a natural (as opposed to an intentional) process of selection for doing something that it does not do. So if *R*s are selected for indicating *C*s, they must actually indicate *C*s during this selection. It follows that during recruitment there can be no misrepresentation of *C*s by *R*s, or more neutrally, no *R*s in the absence of *C*s. Thus, during this period, which Dretske early on refers to as the learning period, there can be no representation of *C*s by *R*s either, since (Dretske says) representation requires the possibility of misrepresentation. It is only once recruitment of *R*s for indicating *C*s ceases that an *R* in the absence of a *C*, and hence mis/representation of *C*s by *R*s, becomes possible. This seems quite unrealistic because, if *R*s can occur without *C*s after the learning or recruitment period, *R*s could surely occur without *C*s during the learning or recruitment period (Fodor 1991a).

In places, Dretske talks as though he is using a less strict notion of indication, one that permits talk of the "maximally indicated state." He does not elaborate much on this, but it is an interesting consequence that the content ascriptions he supports are different from those supported by a theory like Millikan's or Papineau's. Consider again the frog or toad. The maximally indicated (distal) state is presumably not frog or toad food but instead the presence of a stimulus with a certain configuration of visible features. The more maximally indicated state, at least, is something small, dark and moving (in the case of the frog) or something elongated and moving parallel to its longest axis (in the case of the toad).

Neander (1995) and Jacob (1997) have defended accounts that generate similar contents for simple systems. Neander (forthcoming) offers an informational teleosemantics that uses a more lenient notion than Dretskean indication. She stipulates that a mechanism *informs* a state or event of type *R* about something, *C*, to the extent that it does something to enhance the correlation between *R*s and *C*s by causally mediating between them. Neural components can causally connect *R*s and *C*s by being caused by *C*s to produce *R*s (as happens in perception) or by being caused by *R*s to produce *C*s (as happens in motor output). Neander claims that the referential content of a representational simple at the sensory/motor periphery

is what it is "supposed" to be informed about. That is, its content is what the systems that were adapted for informing it were adapted for informing it about. This proposal extends the scope of Dretske's theory, which only applies to perception, and it avoids the need for a distinct learning or recruitment period, in which if there is an *R* there must also be a *C*. The correlation between representations and their represenateds need never have been perfect on this version of informational semantics.

It is an intended implication of both Dretske's and Neander's proposals that the toad's perceptual representation has the content (roughly) *elongated thing moving parallel to its longest axis (at such and such a location)* and that the frog's representation has the content (roughly) *small, dark, moving thing*. As Dretske puts it, these are the more maximally indicated states, which the representation was selected for indicating. And, as Neander puts it, these are what the relevant perceptual systems were selected for informing the representations about. They were not selected for informing them about packets of nutrients, as Neander defines "informing," because they were causally insensitive to the presence or absence of nutrients. Neander (forthcoming) argues that this is the right result for an information-processing account of the toad's perceptual capacity, on the grounds that such content ascriptions, unlike those generated by more standard teleological theories of content, can play a role in information-processing accounts of the relevant cognitive or perceptual capacity.

Toads are not in error, on this account, if they snap at a suitably sized cardboard rectangle moved parallel to its longest axis, but they are in error if they snap at a stunned worm, a cricket, or a millipede that is dangled by its tail and moved perpendicular to its longest axis, which can happen in cases of severe neurological damage to the toad (e.g., ablation of parts of or all of its thalamus). This is the inverse of the results that the teleological theories of Millikan and Papineau aim to deliver.

Like the others before her, Neander also offers a solution to the problem of distal content. To say that (e.g.) the frog represents something small, dark and moving is not the same as saying that the content of its representation is proximal, since (after all) some small, dark, moving things are insects and insects are appropriately distal. However, there is a prima facie problem for her theory, regarding distal content. If a mechanism was selected for informing a representation about a distal item, it was also selected for informing it about the proximal items that carry the information about the distal item to the representation. It was *by* being informed about these that it was informed about the more distal item. Neander proposes that appropriately distal content is found at the end of the causal chain that the representation is supposed to be informed about. Take it as given that neural components were adapted for informing *R*s about *C*s. If so, she says, they were also selected for informing *R*s about more proximal items (*P*s) that carry information about *C*s to *R*s. But, she maintains, there is a difference: the neural components were adapted for informing *R*s about *P*s because they were adapted for informing *R*s about *C*s, but they were not adapted for informing *R*s about *C*s because they were adapted for informing *R*s about *P*s.

Some (even in the face of Pietroski's argument) will find the kinds of contents ascribed by theories like Millikan's and Papineau's more intuitive than those ascribed by theories like Dretske's and Neander's. Millikan also argues that theories of the latter kind preclude the possibility of representing kinds with hidden natures or essences (Millikan, 2000, appendix B). However, Dretske's and Neander's theories are offered as theories for representational simples, and it is questionable whether kinds with hidden natures are plausibly represented by simples. On an account of this sort, kinds with hidden essences, as well as non-existent kinds, must be represented by complex concepts.

The main problem with this kind of approach is the modesty of its scope. This is a bottom-up approach – one that seeks to account for the contents of representational simples and one that leaves the bulk of the work for the second-stage theory, which aims to explain how complex concepts can be derived from simpler ones. While this might turn out to be the correct approach, it is certainly a long way to the top and it is far from clear that we can get from here to there. (A somewhat more detailed introduction to different kinds of teoleological theories is given in Neander 2004.)

Along with consciousness, intentionality has been thought to be the mark of the mental. Both have been traditional philosophical puzzles. Many philosophers are inclined to think that consciousness is the less tractable of the two, and that we have some idea, at least, of how to proceed toward an understanding of mental representation. As we have seen, a number of ideas have been put forward. There are, however, serious difficulties with all of the presently available naturalistic theories of content and, to say the least, much work remains to be done. Some think that what is needed is more work on the same basic ideas, whereas others think that a radically novel idea, unlike anything proposed so far, is still needed.