Causal Theories of Mental Content

Forthcoming in Blackwell's *Philosophy Compass*

1. The problem of intentionality

Humans think *about* things: their friends, the World Series, elm trees, elections, and subatomic particles, to name a few. This raises what philosophers call 'the problem of intentionality'. How is it that human concepts, beliefs, desires, and thoughts can take objects, including nonexistent ones? How can any thought be *directed toward* something; and furthermore, what makes it the case that it is directed toward one thing – a property, an individual, a state of affairs, whatever – rather than another? Intentionality, this *aboutness*, is one of the distinctive and fundamental features of the human mind. As such, it has been the focus of sustained philosophical investigation, some of which will be surveyed in what follows.

Intentionality appears to pose a special problem for those who accept the contemporary scientific view of the universe. On this view, the universe contains nothing nonphysical, immaterial, or otherwise unnatural, i.e., nothing that cannot, at least in principle, be understood in terms of the natural order of cause and effect described by the physical sciences.[1] Contrast the aboutness of human thought with other physical relations, for example, *being next to*. The filing cabinet might be next to the desk, but it is not *about* the desk; it is not *directed* at the desk, it does not *mean* the desk, and it does not take the desk as an *object*. What in the physical world grounds the aboutness of concepts and mental states? If concepts (or mental representations, as I will more often

call them) and mental states are physical entities – states of the brain, for example – how do they attain an aboutness that seems otherwise absent from the physical universe?

One might attempt to avoid the problem of intentionality by embracing mind-body dualism. If the mind consists of a nonphysical, mental substance, then *being about x* could be a kind of simple property of that mental substance. The dualist solution appears mysterious and implausible, however, especially given the enormous success of materialist science, in general, and our ever increasing understanding of the physical basis of mental functioning, in particular. Better, then, to locate the mind and its distinctive attributes in the physical order of causes and effects, or so has been the view of those developing causal theories of mental content (CTs, hereafter).

Two kinds of question might be asked about the intentionality of the mental. One concerns the actual individual or the collection of actual things to which a given mental representation correctly applies. The mental representation $GOLD^2$ correctly applies to all samples of a particular element: all atoms with 79 protons each or, perhaps, all substantial collections of them. It is part of a CT's job to explain how this can be so. Philosophers often talk about a second and related aspect of intentionality, for which we might also expect a CT to account. Assume that when a mental representation properly applies to something – or would properly apply to something – it is in virtue of some property of the thing in question. In the case of gold, for example, it would seem that GOLD properly applies to all samples of gold *because* they have the property of being gold – not, for instance, because they have the property of being samples of Midas's favorite metal. Thus, GOLD seems to have a second kind of intentional content consisting in its privileged relation to an abstract property or natural kind. We might

conveniently think of this second kind of intentional content as an abstraction from the first, i.e., an abstraction from the collection of things to which a given mental representation in fact properly applies. Nevertheless, in some cases, this is clearly not the correct way to think about the relation between the two forms of intentional content. We should want to leave open the possibility that, for instance, UNICORN has intentional content of the second kind absent any actual unicorns – past, present, or future – from which a property might be abstracted.

The terms used to mark this distinction, or one much like it, vary. In connection with the first sort of content, readers will encounter such terms as 'reference', 'designation', 'representation', and 'extension', and in the first three cases especially, the related verb-forms: it is sometimes said that GOLD *refers* to all of the actual samples of gold, represents them, designates them, or has them all in its extension.[3] The second form of intentional content described above is sometimes described as a concept's 'referential content' or as its 'representational content'. The terms 'mental content' is also used in this regard, although frequently in a way that applies indiscriminately to both kinds of content. In contexts where the distinction at issue is not particularly germane, many authors also use 'reference' and 'representation' in this indiscriminate way.[4] When the question is, for example, whether my concept HORSE has as its content horses or rather trees instead, the distinction between, for example, reference and referential content is moot.

Keep in mind, too, the various forms mental representations themselves can take and the corresponding variations in the objects of those representations. Of greatest importance in this respect is the distinction between concepts as atomic mental units (e.g.,

the concept CAT) and mental states such as beliefs and desires, which express complete thoughts. Presumably the latter states have various atomic concepts as constituents, e.g., the belief that cats are mammals in some way has the concepts CAT and MAMMAL as constituents (although we should not assume without question that the structure of the vehicles of representation mirrors the structure of what is represented – see Dennett, 1991). When we are concerned with complete thoughts, we must recast the distinction between representation and representational content. Rather than contrasting the individuals or collections of individuals to which a mental representation properly applies, on the one hand, with the properties by virtue of which the mental representation properly applies to them, on the other, we might contrast states of affairs with truth-conditions. States of affairs are often thought of as the situations obtaining in the actual world – the facts – to which, for example, a true belief in fact corresponds. In this sense, the state of affairs is the compound thing in the world to which the complete thought correctly is applied: Tom's belief that gold is an element is about a state of affairs in the world – the state of affairs the obtaining of which makes Tom's belief true. In contrast, the representational content of a belief consists in its truth-conditions, which specify or determine what it would take to make the belief true. On this way of describing things, a false belief represents nothing, for there is no actual state of affairs to which the belief applies or corresponds. Nevertheless, the false belief has representational content in the sense that it has truth-conditions; it is associated with something – perhaps an abstract entity – that specifies what must obtain in order for the belief to be true.

In what follows, 'mental representation' should be read as designating either conceptual atoms or complete thoughts, as the context demands. Some CTs take

conceptual atoms to be of primary importance (Fodor, 1990, 1998, Papineau, 1993, p. 82) while others take complete thoughts – beliefs and other mental states that would seem to have subject-predicate structure or express propositions – to be initial bearers of mental content (Millikan, 1984). Furthermore, some CTs focus, in the first instance, on specific causal interactions (the tiger in front of Sarah caused the activation – or the 'tokening' – of her concept TIGER), in which case representational content might be seen as parasitic on representation (Dretske, 1981, 1988, Rupert, 1999). Other theories focus to a greater extent on the law-like relations between properties (Fodor 1990), which suggests the primacy of representational content.

2. What should we want from a CT?

An appeal to causal relations does not, by itself, solve the problem of intentionality. Causes and effects permeate the universe, but intentionality does not. Thus, a CT must identify the particular form or pattern of causal relations that establishes, determines, or constitutes an intentional relation.

In the discussion to follow, I employ two evaluative criteria. First, a CT must explain misrepresentation. Humans frequently misrepresent the world, often in cases where the item or state of affairs being misrepresented caused the tokening[5] of the mental representation in question. To illustrate: In the typical subject, the perception of a horse can cause the tokening of HORSE; however, many other things can cause HORSE-tokenings in her – for example, saddles or cows on dark nights. Obviously, some of these causes are not horses. Imagine a theory stating that, if some kind of thing can cause mental representation *R*, then *R* represents that kind of thing. On this theory, misrepresentation seems impossible, at least as regards the misrepresentation of

something causing a present tokening of *R*. Any object or property that can cause the tokening of *R* is, thereby, in the extension of *R*.[6]

The second criterion makes the straightforward demand that a CT assign *correct* extensions to mental representations. The simple causal story considered above clearly fails by this standard. On such a view, HORSE might represent a horse, a saddle, a cow on a dark night, or, for that matter, an utterance of the word 'horse'. The list is to include every kind of thing that can cause the tokening of HORSE in the subject in question. The disjunctive form taken by such a list – HORSE is properly applied to a thing if it is either a horse or a saddle or an utterance of 'horse', etc. – makes salient the oddness of the list, and this oddness underlies what Fodor (1987, p. 102) labels the 'disjunction problem'. Insofar as the disjunction problem extends beyond the problem of misrepresentation discussed above, the disjunction problem evokes a more general demand for correctness. It might be that some concepts have conditions of application most naturally stated in disjunctive form (consider the concept STRIKE from baseball). Nevertheless, we take very many of our concepts *not* to be of this sort. Given that the simple causal story implies otherwise, it assigns the wrong extensions.

This demand for the correct assignment of truth-conditions has a positive and a negative side. The latter requires that our CT not make patently erroneous content assignments: any CT entailing that, even though it seems to me I am thinking about the 2004 World Series, I am actually thinking about crop failure in Ethiopia faces a nearly decisive objection. Thinking now in positive terms, we simply demand that a CT assign the content we intuitively take to be correct.

Application of only these two evaluative criteria substantially limits the discussion. Nevertheless, it fairly captures the dimensions on which CTs have typically been evaluated, even if a genuinely satisfactory CT must ultimately meet other demands.[7]

2. Informational theories of content

2a. Dretske

Fred Dretske (1981) developed one of the earliest and most influential CTs by appropriating concepts from information theory.[8] This is an applied mathematical theory pertaining to the amount of information – typically the average amount – generated or transmitted by processes of interest to electrical engineers (for example, telephone signaling). On a specific occasion when a signal is transmitted, the transmitting source is in one of its possible states; so, too, is the device receiving that signal, and this latter state – the state of the receiver – may reveal more or less about the state of the source. The *amount* of information the receiver's state represents about the source depends, roughly speaking, on the extent to which the receiver's state homes in on the state of the source (assuming the source can take on a range of possible states). If the state of the receiver is consistent with a wide variety of sources of the state, then the state of the receiver represents less information about the source than if the state of the receiver had been consistent with only one or two states of the source. As an illustration, consider a case in which an English speaker passes a one-word note to another English speaker. The end of the word is illegible; all that can be made out is 'pe', with a smudge following it. The resulting state of the receiver – the visual apparatus of the person reading the note – is consistent with a substantial range of English words: 'pet', 'pen', 'percolate',

'pedestrian', and many more. Thus, the state of the receiver does not come close to pinpointing the state of the source—i.e., the state of mind of the person who wrote the note. In contrast, if the note had shown the letters 'perenniall' followed by a smudge, the state of the receiver would have carried as much information as possible in this situation; for it rules out all possibilities except that the person writing the note had 'perennially' in mind.

In more technical terms, a receiver carries the greatest possible amount of relevant information about the source if the conditional probability of the obtaining of some particular simple state of the source given the state of the receiver equals one. (This situation is symbolized as P(S | R) = 1; the probability of S-given-R equals one.) This is simply a way of saying that the state of the receiver pinpoints a single, simple state of the source: it is guaranteed (i.e., the probability equals one) that the source is in that specific state, given the condition of the receiver's being in the state it is in; relative to the range of possible states at the source, the state of the receiver could not have carried a more detailed message.[9]

To work these ideas into a CT, Dretske first defines the informational content of a receiver-state to be *any* state S – even a complex one – of the source such that P(S | R) = 1. That is, the informational content of the receiver state is whatever information is carried by the state of the receiver, even if that information is, returning to the example given above, only that the state of the source could be 'get', or it could be 'bet', or it could be 'ballet', etc. Informational content alone does not assign the right mental contents; for in many cases, the content of a mental representation is something more

specific than a disjunction of all of the states of the world consistent with the subject's state. I return presently to this concern.

First, though, consider a closely related problem. Carrying the information that something is a mammal also carries the information that it is an animal; thus, if the probability of something's being a mammal, given the obtaining of a particular receiver state, equals one, then so is the probability of that thing's being an animal given the same state of the receiver. Nevertheless, a mental representation of a thing as a mammal should be distinct from the representation of that thing's being an animal. These are, as we might say, two distinct thoughts. Dretske recognizes this problem and demands more of semantic content – i.e., mental content – than merely informational content. The *semantic* content of a mental representation *R* is the *most specific* piece of information *S* such that $P(S \mid R) = 1$.

As with any CT, Dretske's informational theory must account for misrepresentation. At first blush, Dretske's view is in a boat with the simple causal theory considered earlier, even when attention is limited to Dretskean semantic content. Suppose that someone mentally represents that something up ahead is an aardvark. Suppose also that the mental representation tokened on this occasion, call it AARDVARK, can be tokened in conjunction with an anteater's being up ahead, but not in conjunction with any other state of the environment besides the presence of an aardvark or an anteater. In this case, the conditional probability of an aardvark being in the local environment given the tokening of AARDVARK is less than 1, and similarly in the case of anteaters. In contrast, the probability of there being either and aardvark or an anteater present, conditional upon the representation's being tokened equals one. Assuming that no more

specific piece information has as high a probability, Dretske's theory entails that the content of the mental representation in question is *that thing is either an aardvark or an anteater*. Oddly, then, whether the subject encounters an aardvark or an anteater, so long as the subject tokens AARDVARK, the subject accurately represents the animal as something that is either an aardvark or an anteater, instead of accurately representing an aardvark or misrepresenting an anteater as an aardvark – as the case may be. This results generalizes, so that in any case where we think the subject misrepresents the environment, Dretske's informational theory seems to tell us that, instead, the subject accurately represents a "disjunctive" state of affairs.

Dretske handles this problem by positing a learning period during which semantic content is determined. It is only during this period that representation *R*'s semantic content is determined in accordance with Dretske's conditions stated above. Thereafter, the content of symbol-type *R* is the semantic information that tokens of *R* carried during the learning period. Thus, once the learning period ends, it is possible that *R* will be misapplied, i.e., it will be applied to something about which it did not carry information during the learning period. One way to maintain the proper correlation during the learning period is via a supervisory mechanism. For example, the intervention of a teacher might ensure that whenever AARDVARK is tokened, an aardvark is present. After the learning period has ended and the supervisory mechanism is no longer active, it is possible for the subject to token AARDVARK in the absence of an aardvark, e.g., when an anteater enters the subject's visual field, without anteaters thereby being in the extension of AARDVARK.

2b. Objections to Dretske's purely informational theory

As an account of the content of human mental representations, Dretske's purely informational story seems implausible. Humans do not acquire concepts (or primitive mental representations) by a two-step process. There is no time in the acquisition of a concept when a whistle blows (or what have you) to indicate that learning has ended, the concept's content is determined, and now misrepresentation is possible (Fodor, 1990, p. 41).[10] Perhaps of greater concern is that, even if there were a cleanly delimited learning period, the conditional probabilities of interest would not have the value one. Children make many mistakes in the acquisition and early use of concepts; in some cases – such as overgeneralization – those mistakes can have a principled basis (that is, the sort of basis in natural law that is relevant to informational relations, according to Dretske). Furthermore, even if an attentive parent follows the learner around, the parent is himself subject to misrepresentation; thus, the conditional probabilities of interest will not be one: the parent may sometimes mistake an anteater for an aardvark. In fact, much human learning involves imperfect covariation (Slater, 1994). The child might token BIRD as the result of perceiving smallish creatures with wings; however, some things that are not birds (e.g., bats) fit that description, too, and the child might not become aware of this until well after any plausible learning period, relative to BIRD, has ended. It is likely, then, that during the learning period, many uncorrected mistakes are made, which entails, on Dretske's pure-informational view, that the concepts in question have the wrong extensions.

2c. Fodor's Asymmetric Dependence view

Jerry Fodor begins with the idea that misrepresentation involves an asymmetry: *R*'s responsiveness to what it misrepresents depends on *R*'s responsiveness to what it actually

represents, but not vice versa. When seeing a cow on a dark night, someone might apply her mental representation HORSE to that cow, but the application depends on a more fundamental relation between HORSE and horses. Upon this insight, Fodor builds his Asymmetric Dependence Theory (AD) (Fodor 1987, 1990). Let us say that *Q*s can cause *R*-tokenings and that *S*s can cause *R*-tokenings. What would make it the case that *R* properly applies only to *Q*s and not to *S*s? According to Fodor, it is that the following conditions are met:

1. It is a law that *Q*s cause *R*s.
2. It is a law that *S*s cause *R*s.
3. If it were not a law that *Qs* cause *Rs*, it would not be a law that *S*s cause *R*s (alternative construal: in the nearest possible worlds in which it is not a law that *Q*s cause *R*s, it is not a law that *S*s do) (i.e., if (1) were false, then (2) would be false).
4. If it were not a law that *Ss* cause *Rs*, it would still be a law that *Q*s cause *R*s (alternative: in the nearest possible worlds in which it is not a law that *S*s cause *R*s, it is a law that *Q*s do) (i.e., if (2) were false, (1) would still be true).

In summary, breaking the nomic connection (i.e., the connection grounded in laws of nature) between *Q*s and *R*s would sever the nomic connection between *S*s and *R*s, but not vice versa. Consider an example, fleshed out in possible worlds talk: in the nearest possible world in which it is not a law that horses cause HORSE-tokenings (i.e., tokenings of the mental representation that we would normally take to be the mental representation of horses), it is not a law that cows cause HORSE-tokenings; but in the nearest possible world where it is not a law that cows cause HORSE-tokenings, it is still a

law that horses cause HORSE-tokenings. On AD, this establishes that HORSE represents only horses, even though cows sometimes cause HORSE-tokenings (for instance, when seen from a distance at night).

Various issues must be ironed out here. Assume *S*s are cows and *R*s HORSES. In what sense are these nomically connected? It is consistent with the laws of nature that an *S* cause an *R*; there is, we might say, at least one law-respecting causal route from *S*s to *R*s. It would, however, be a very strange conception of laws of nature according to which cows and HORSE are *directly* related by natural law (cf. the discussion of the Natural-Kinds-Only assumption in Rupert, 1999, pp. 344-48). There may be a way to gerrymander properties or events such that the specific situations in which cows cause HORSE-tokenings (say, on dark nights when cows are at certain distances) are connected by natural law to HORSE-tokenings, but this seems a desperate strategy. Yet, Fodor needs such a story so that he can cash condition 4. To make good on condition 4., Fodor must explain which law-like relations hold between cows and HORSE and are such that, if they were altered, horses would still cause HORSE-tokenings. Causal relations hold in virtue of laws, and thus to break a causal connection one must, so to speak, change the laws.[11] In the most straightforward passages, Fodor focuses on the property *being a cause of HORSE-tokenings*. He claims that cows instantiate such a property and that breaking the nomic connection between cows and HORSE-tokenings is a matter of breaking the connection between the property *being a cow* and the property *being the cause of HORSE-tokenings* (see, e.g., Fodor, 1990, p. 101, 1994, Appendix B). The problem is that the property *being the cause of HORSE-tokenings* does not cause anything – at least it does not cause tokenings of HORSE (although it might cause

musing in philosophers – see Block, 1990). Compare: the cause of the fire, in fact, caused the fire; but surely the cause of the fire did not cause the fire *in virtue of its property of being the cause of fire*; rather, it did so in virtue of certain of its chemical properties (Rupert, 2006).

Fodor might recast AD in terms of causally efficacious properties, but it is not obvious that such a modified view yields reasonable content assignments in the human case.[12] As AD is typically presented, content assignments rest on the truth of counterfactuals of the form, "if it were the case that such-and-such nomic connections were broken, then it would be the case that other nomic connections are (or are not) affected in such-and-such a way."[13] Whether this yields the right answers in cases of human mental representations depends on which nomic connections could be broken such that, for example, horses no longer cause HORSE-tokenings. It is in virtue of horses' various observable properties that they cause the tokening of HORSE: horses cause HORSE-tokenings because they have horse-like manes, equine gaits, etc. Thus, one has to imagine the nearest possible worlds in which the connection between those properties and HORSE is broken (and similarly for the properties of cows that are operative when cows on dark nights cause HORSE-tokenings). It is simply not clear which mental representations – individuated nonsemantically – a human will token in response to a horse or to a cow when these animals' observable properties are nomically dissociated from the mental representation I have been referring to as HORSE.

A further factor exacerbates this problem: a subject's reactions in counterfactual circumstances also depends on the *contents* of her various mental states (which will, I take it, remain the same in the nearest possible worlds in which the connections in

question are broken) (Boghossian, 1989, pp. 539-40; 1991, pp. 78-83; Rupert, 2000). Since, however, Fodor presents AD as a naturalistic theory, AD should not assign content to a given mental representation by invoking the *contents* of other mental representations. The problem is that, with regard to the breaking of nomic connections that plausibly inhibit the relevant causal processes, the subject's mental or cognitive states might contribute to a pattern of tokenings which, given AD, entails an incorrect content assignment. If we attempt to specify the nomic connections to be broken independently of the contents of the subject's mental or cognitive states, we run the risk that the contents of the subject's other states will confound the expected effect. If, however, our choice of connections to break depends on the contents of the subject's mental states, AD does not handle this kind of case in a naturalistic way. It is one thing to say, as Fodor does, that content-laden states can mediate the content-conferring relation; it is another to say that the content-conferring relation can only be specified by adverting to the contents of subjects' other mental states.

2d. Views that appeal to actual history

As I explicated AD, it makes no appeal to the actual history of causal interactions into which a given kind of mental representation enters. Sometimes, however, Fodor considers the merits of an actual-history requirement (1990, pp. 120-127), i.e., a requirement that for *R* to represent *Q*, some *Q*s must actually have caused some *R*-tokenings in the subject in question. In the end, Fodor rejects the actual-history requirement (1994, Appendix B), but many other authors have thought it useful to embrace it in some form or other (Dretske, 1981, 1988, Prinz, 2002, Rupert, 1998/99,

1999, Ryder, 2004), and in what follows, I discuss a pair of such approaches – first my own, then that of Dan Ryder.

I am inclined to think that actual history does play a content-determining role, at least if we limit our attention to basic cases, i.e., to cases of syntactically atomic mental representations the content of which is not fixed by the content of any of our intentional states directed toward those representations.[14] Here is the fundamental principle behind an approach I have elsewhere dubbed the 'best test' theory of content, but which I should rather call the 'causal-developmental theory' (the prior label seemed to cause only confusion):

> If a subject S bears no content-fixing intentions toward *R*, and *R* is an atomic mental representation (i.e., not a compound of two or more other mental representations), then *R* has as its extension the members of natural kind or property *Q* if and only if members or instances of *Q* are more efficient in their causing of *R* in S than are the members of any other natural kind or property.[15]

We cash efficiency in terms of ordinal comparisons between past relative frequencies (PRFs) of certain causal interactions. Take a mental representation *R* in a given subject S. For each natural kind or property *Qi*, calculate *its* PRF relative to *R*: divide the number of times an instantiation of *Qi* has caused *S* to token *R* by the number of times an instantiation of *Qi* has caused *S* to token *any* mental representation whatever.[16] Then make an ordinal comparison of all *Qj* relative to that particular *R*; *R's* content is the *Qj*

with the highest PRF relative to *R* (Rupert, 1996, 1999; see Usher, 2001, for the terminology of ordinal comparisons). Furthermore, for the purposes of assigning contents to mental representations, only PRFs resulting from a substantial number of interactions matter: if a subject tokens TIGER upon causally interacting with her first hyena, hyenas have a maximal PRF relative to TIGER, but this does not entail that hyenas constitute the extension of TIGER. On the causal-developmental view, content is determined by a substantive developmental process, a result of the subject's physically grounded dispositions having been shaped by the subject's developmental interaction with the environment.

Consider some of the advantages of the causal-developmental view. First, it can be fruitfully integrated with a neural account of the way in which the realizers of mental representations become functioning, coherent units (Rupert, 1998/99, 2001) and take on causal roles appropriate to their content. Mental content appears as the PRFs reach relatively stable values; during the same time period, the neural vehicles of mental content are being stabilized by the infant's or toddler's interaction with the world. Given that this shaping up is partly reward-dependent, the process of content-fixation can help to explain the appearance of content-appropriate behavior. If the pattern of rewards had not influenced the stabilization of the relevant neural unit, there would be no such stable functioning unit. So, successful or useful behavior is a precondition for acquiring a stable neural unit. But, at the same time, the content-determining pattern of PRFs is what facilitates the appearance of said neural unit. If there were not a clear winner in the PRF competition – relative to a particular incipient neural unit – that unit would not cohere as a functional unit. And, it seems natural to expect that the behavior to be rewarded on a

regular basis would be the behavior appropriate to the PRF-winning kind; in which case, a human will typically come to possess a stable neural unit only under conditions where that unit represents (in keeping with the causal-developmental view) the same kind or property to which its rewarded actions (which shaped up the unit) are appropriate.

The causal-developmental view also avoids the problem of the learning period. There is no precise cut-off point when ordinal comparisons are to be made; at least in the case of primitive mental representations, stable ordinal comparisons are in place by the time we have any confidence in our mental-state attributions (or think that the states so attributed play a role in cognitive science).

The causal-developmental view also solves the notorious *qua* problem (Devitt and Sterelny, 1987, pp. 63-65, 72-79). Items causing our mental representations instantiate more than one property or kind. How can our causal interactions with those items affix a mental representation to one property or kind to the exclusion of the others? This problem arises particularly for theories that ground content in actual causal interactions between the subject and her environment: if the subject interacts with a dog, she also interacts with a mammal. Given that both properties or kinds caused the subject's tokening of, say, DOG, why should one kind be the determinate content of DOG? The solution is straightforward in some cases. Even though every dog is a mammal, the PRF of dogs relative to DOG is much higher than mammals relative to DOG. Almost all dogs that cause the tokening of any mental representation in S cause DOG-tokenings in S, but frequently when mammals cause the tokening of some mental representation or other, they do not cause DOG-tokenings. This solves the *qua* problem where the threat is "from above," as one might say; we see clearly how a mental representation can be assigned the

correct extension in a case where the two candidate extensions are more and less specific (the former contained in the latter) and in which the more specific extension should be assigned.

A *qua* problem also arises via competition from below. In some cases, we should like to assign a more general content, but where the PRF-based measure recommends a more specific content. In such cases, the causal-developmental theory offers a less elegant, but still independently motivated, solution: the subject directs intentions toward the mental representation in question, and this limits its content to an intended kind or property. Allowing this kind of content-fixation requires a content-fixing principle beyond the simple one stated above (recall that the principle stated above applies only to primitive mental representations – primitive in the sense that they are both atomic and are such that we do not direct intentions about their content toward them). There might be some particular kind of dog that has a higher PRF relative to DOG than dogs in general do; perhaps this is a stereotypical dog breed (e.g., the Golden Retriever). The causal-development theory handles such cases in two steps. First, the principle stated above explains how naturalistically respectable mental content enters the scene. Once in play, however, such content-properties can have influence of their own. The purely PRF-based principle is operative in the early stages of development, and it is not clear that in this case, we need solve the *qua* problem from below; instead, it may do no harm simply to apply the first principle with the result that the child's concepts refer to narrower kinds than we might have thought. (It is striking that children need great help and explicit guidance to learn highly inclusive concepts, such as ANIMAL and VEHICLE.) Second, children (and adults, for that matter) can employ existing content to alter or fix the

content of a new concept. The child needs only a rudimentary concept of a kind of thing in order to survey Golden Retrievers, German Shepherds and Poodles and think "DOG should represent whatever all those things have in common." Even if the child's DOG had represented only Golden Retrievers prior to that point, the child has used existing naturalistically determined content to assign new content to DOG (cf. some of the strategies described by Stanford and Kitcher, 2000).

The causal-developmental view faces various difficulties. Generally speaking, the view takes an empirical risk. It presupposes that PRFs alone suffice to determine content in a wide enough range of actual cases to provide the materials for the fixation of the content of further mental representations – as discussed above in connection with the *qua* problem. Humans think about a wide variety of things, however – hammers, beauty, politics – that do not seem to be natural kinds and that cannot easily be defined, or even, for practical purposes, uniquely characterized, in terms of natural kind concepts. The causal-developmental view cannot guarantee that PRF-determined content covers such a wide range of case.

On a related note, humans think about nonexistent kinds and individuals, such as unicorns and Santa Claus. On the causal-developmental view, the subject must directly interact with the kind members, property instances, or individuals represented *or* the subject's existing stock of content-possessing mental representations must somehow allow her to home in on such kinds, properties, or individuals. There is no causal interaction with nonexistent referents, though, so all emphasis is placed on the construction of representational content from already established content. The advocate of the causal-developmental view might hold that some of the properties or kinds

discussed earlier – *beauty* or *politics*, for instance – are natural properties, and thus that there can be direct causal interaction with their instances. This can hardly work in the case of non-existent individuals or uninstantiated kinds, though.

Lastly, consider the possibility that an unexpected kind have the highest PRF relative to a given mental representation. Perhaps, for instance, a particular neural kind – a stimulation of cells in visual cortex – has the highest PRF relative to a given mental representation, in a case where we would think that mental representation takes as its content a property or kind beyond the skin of the subject. There would seem no way to rule this out *a priori*, and it would be an embarrassment to the causal-developmental view were it have to be forced to say that the content of, for example HORSE, is a pattern of cells firing in visual cortex.[17]

Dan Ryder's SINBAD theory of content (Ryder, 2004) also appeals to developmental and learning history but focuses primarily on changes at the neural level. Ryder's general idea is that neurons come to have a certain structural properties best explained by their responsiveness to particular natural kinds; given that these structural properties ground the important cognitive functions of neurons (or groups thereof), it is reasonable to assign to them, as content, the natural kinds interaction with which explains the neurons' coming to have that cognitively useful structure.

Each neuron in the brain receives incoming signals through branch-like structures called 'dendrites'. Ryder holds that dendrites follow a 1/n rule, according to which each individual dendrite of a particular cell adjusts its firing strength so that, when the cell fires (sending outgoing activation down what is called its 'axon', thereby stimulating further neurons' dendrites), each dendrite contributes equally to the cell's firing. I.e., if *n*

is the number of dendrites attached to a neuron, each dendrite will "strive" to contribute one *n*-th of the amount of incoming stimulation on occasions when the neuron fires. To achieve this, a learning mechanism adjusts the contribution of each dendrite so as to bring that dendrite's contribution closer to that of each of the other dendrites that contribute to the firing of the cell in question.

On this view, historical considerations fix the representational content of an individual cell (or group of cells with similar profiles); call it *c*. Given that *c* has achieved its "goal" of making a one *n*-th contribution (or to the extent that it has achieved the goal), interaction with members of one natural kind rather than others will explain the cell's achieving of that "goal." Typically, the members of a given natural kind manifest numerous, correlated features. Assuming, then, that each of the correlated features of a kind is paired causally with the firing of one of *c*'s dendrites, the 1/n equilibrium will be achieved *because c*'s firing was caused by interactions with members of one kind *k1* (the kind manifesting the particular package of correlated features in question – see Boyd, 1990) rather than any other kind. The kind manifests a particular package of correlated features, and it is the regular appearance of these features together that explains the stabilization in the firing pattern of *c*'s dendrites. For this reason, *c* has *k1* as its content.[18]

SINBAD neurosemantics offers a clear account of misrepresentation. A cell or group of cells *R* represents a particular kind *Q* if and only if that cell's having reached 1/n equilibrium is best explained by causal interactions with *Q*s. If at some time after achieving equilibrium *R* is applied to something other than *Q*, *R* misrepresents that thing.

It is not clear, however, that SINBAD neurosemantics offers an empirically plausible account of the content of human mental representations. The 1/n rule presupposes a kind of regularity in the distribution of features among individuals of a given kind. Perhaps if one chooses the right features, this could be made to work in certain cases: virtually all faces have two eyes, a nose, and a mouth. Nonetheless, these cases seem rare. Even in the case of faces, one might think that many optional features, e.g., having a beard, play a role in categorization (the features assigned low weights in a prototype or probabilistic view – see Smith and Medin, 1981).

In response to a worry of this sort, Ryder considers multivariate functions (2004, pp. 219-20), but his solution applies most clearly to kinds that exhibit a balanced feature-structure, e.g., where members of the kind are detected by the rule (A or B) and (C or D). Many kinds, however, do not exhibit such a neat structure of conjoined disjunctive conditions; instead, they make up a motley overlap of features, and thus, individual dendrites would require very complex, probabilistically weighted sensitivity to a range of features if the 1/n measure is to be realistic. This objection can only be answered by the tandem investigation of categorization behavior in humans and of the computational properties of human neural structures. Thus, it is a challenge as much as anything.

3. Teleological views

The teleological view of mental content – teleosemantics, for short – takes a variety of forms. The essential idea is that mental representation *R* represents whatever *R* has the *function* of representing. The idea of a function can be spelled out in many ways. To narrow the field, then, I focus on what is arguably the most influential of the lot, at least with regard to CTs: the aetiological notion of a function taken from evolutionary

biology.[19] On this view, a subject's current tokenings of *R* have the function of representing *Q* if and only if the capacity or mechanism the subject possesses for producing *R* was selected for because it produced structures that correlated with *Q*. That is to say, the presence of *R* tokenings in the subject now can be causally explained by the increase in fitness conferred on prior subjects who tokened representations of type *R, and* this increase in fitness is best explained by *R*'s relation (covariation is typically cited) to *Q* in cases involving those prior subjects. Misrepresentation occurs, then, when the subject applies *R* to something other than the property or kind *R* has the function of representing.

To illustrate the more general idea of an aetiological function – where the function of the item is determined by the causal interactions responsible for, or which explain, the appearance or continued presence of that item – consider a bicycle pedal. The function of a bike pedal is the way it contributes to the turning of the bicycle's wheel. The causal disposition of the pedal to drive the bicycle's wheel explains why the pedal is there at all. Teleological theories turn this idea into a theory of content by considering the possibility that the, at least indirectly, the aetiological function of some structure is that it correspond to the presence of a particular property or state of affairs in the environment. The two most influential proponents of this form of teleosemantics are Ruth Millikan and David Papineau, whose proposals are discussed below.[20]

3a. Millikan's view

Millikan develops this idea into a CT using the notion of a proper function: the intentional content of a mental representation *R* is determined by *R*'s proper function, where, generally speaking, the proper function of *x* is the effect of *x* the having of which

is responsible for the continued reproduction of members of the reproductively established family of which *x* is a member (Millikan, 1984, p. 28). The idea of a reproductively established family is important because Millikan wants to allow some structure present in an organism now to have its content determined by what structures of that same type contributed to the reproduction of the current organism's ancestors.[21] When we explain how a proper function has been performed in a way that leads to continued reproduction of members of a reproductively established family, we offer what Millikan calls a 'Normal explanation' of the performance of that proper function (Millikan, 1984, pp. 33-34). It is important to recognize that 'Normal' is not meant in the statistical sense: a Normal explanation is not simply an explanation of how the mechanism in question *usually* performs; a Normal explanation explains how the mechanism performs in cases where the performance actually contributes to reproductive success. (Millikan also employs the term 'Normal conditions' to refer to the conditions which must hold in order for a proper function to be successfully carried out in accordance with a Normal explanation.) Some devices or organisms have a proper function they successfully perform only rarely, statistically speaking. Consider sperm. Of all sperm which have existed, the percentage that have fertilized ova is tiny, yet it is the proper function of any individual sperm to fertilize an ovum. It is only because sperm have fertilized ova in a sufficient, albeit a comparatively small, number of cases, that sperm, as a type, has continued to exist.

On Millikan's view, the interdependence of the proper functions of distinct devices is central to the determination of mental content. The proper function of one mechanism or form of behavior can be determined by its role in helping another mechanism or form of

behavior to perform *its* proper function. Consider how this might occur in linguistic communication. The device that produces speech can perform its proper function of, say, conveying information to a listener only when the listener's relevant auditory and interpretive mechanisms perform their proper functions. What is more, relative to a particular spoken sentence, the particular kind of information it has the proper function of carrying may depend on what sort of listeners are in the environment such that *their* interpretation of the sentence causes speakers to continue producing sentences of that kind. Now think more generally of producing devices and their products – like the speakers and sentences of the preceding example – and the consuming devices and their particular acts of consumption – like the listener and her particular acts of interpretation in the preceding example. The proper function of a producing device *R* is determined by the way in which *R* assists a consuming device in the performance of the consuming device's own proper function.

Applying this idea to mental content, a mechanism that generates belief *R* is such that *R*'s mapping onto the world in a certain way is part of the Normal conditions for the operation of the consumer that makes further use of the products of the belief-producing mechanism. A typical case might involve the production of the belief THERE'S A BERRY. For the consuming devices (in this case the reasoning and motor control systems) to fulfill their proper functions, BERRY must map onto berries. If THERE'S A BERRY is applied in the absence of berries, then the subject has misrepresented the world.

3b. Objections to Millikan's view

Robert Cummins objects that when accounting for the continued reproduction of certain mental representations, the teleological view reverses the proper order of explanation (Cummins, 1996, p. 46). It may well be that the tokening of certain mental structures provided a selectional advantage, but they offered this advantage *because* they corresponded to, carried information about, or referred to the world in the appropriate way: *R*'s existing relation (say, correspondence) to something in the environment (e.g., a charging rhino) explains why the subject's tokening of *R* led to survival and reproduction. The intentional relation is already in place prior to selection. Selection does not confer content; things get selected for because they have useful content.

Millikan recognizes the value of various relations that an object might have stood in historically.[22] She says, "Picturing, indicating, and inference are equally involved in human representing, but as biological norms rather than as mere dispositions" (Millikan 1993, pp. 10-11). Evolution has selected for our use of all of these different relations for representational purposes, and so Millikan has identified a common thread. Nevertheless, one should wonder about the import of this emphasis on biological selection as a unifying explanation of intentionality. If there are a variety of ways to come into the intentional relation, we might hope to understand that variety of ways. It is not clear that their common element, having been selected for, does any of its own causal or explanatory work.

3c. Papineau's view

David Papineau has also developed an influential teleological theory of mental content. Although Papineau appeals to historical selection in much the same way Millikan does, he assigns a privileged role to the content of desires. This is the distinctive contribution

of Papineau's view, so let us consider it in some detail. In Papineau's words, "[T]he biological aim of desires is not...to produce true beliefs, but the biological aim of beliefs is standardly to satisfy desires" (1993, pp. 62-63); and as a result, "[T]he teleological approach I am defending here is committed to explaining belief contents in terms of desire contents" (1998, p. 8; cf. 1984, p. 555-56).

According to Papineau, desires have intentional content by representing what they are desires *for*; these are sometimes called 'satisfaction conditions' (analogous to truth-conditions as these were introduced at the outset). The intentional content of a desire consists in those conditions the desire's bringing about of which accounts for selection and preservation of that kind of desire. The state we would normally describe as a desire for food has the satisfaction conditions *that I get food* because, in the past, states of that type caused the subject's acquisition of food, and for this very reason states of that type continued to be reproduced. A given belief's content then derives from the content of desires in the following way: the belief represents whatever conditions in the world are such that, if they obtain, the desire is guaranteed to succeed. For example, a given belief might have the content *there's food in the tree* because when it, together with a desire for food, causes action, the food's being in the tree is just the condition that has to hold for the action to satisfy the operative desire.

It is not clear, however, that this framework delivers sufficiently determinate mental content. Consider first a case involving a human desire, the one we think should be characterized as the desire for food. Plausibly enough, this desire (or the mechanism producing it) was selected for because it got food into the subject's digestive system. The problem of indeterminacy arises, however, as a result of two kinds of further effect of the

past desires for food. In the case of any particular desire for food, it had what we might call the 'upstream' effects: the reaching for the food or the moving of the arm in a specific way got the food into the system. It also had what we might call 'downstream' effects: digestive processing, the continuation of life, and the enhancement of reproductive chances. The desire had all of these effects and at least in propitious circumstances, all of these effects enhanced the subject's reproductive chances. On what basis can Papineau's theory decide the issue in favor of one particular effect, the getting of food into the subject's system, as the content of the desire?

Papineau proposes a distinct solution for each of the two kinds of effect (1998, pp. 11-13). To address downstream effects, Papineau appropriates Karen Neander's (1995) claim that malfunction diagnoses function. The function of a given desire (or desire-producing mechanism) is to do *F* only if, when that desire does not cause *F*, the desire has malfunctioned. Consider, then, that if the stomach does not digest food consumed, it would be wrong to say that the desire for food (or the mechanism producing that desire) has malfunctioned. Thus, causing digestion cannot be the function of the desire for food, and the desire does not have digestion as its satisfaction condition.

Concerning the upstream case, Papineau points out the extent to which such effects vary from case to case. Different such effects occur on different occasions depending on which beliefs appear in combination with the desire for food; which belief is active depends, for example, on where the food is located. "So if we want to identify effects which it is the function of the *desire* to produce," Papineau says, "we need to go far enough along the concertina [the chain of effects under consideration] to reach results which do not depend on which beliefs the desire happens to be interacting with" (1998, p.

12; also see 1984, p. 564). Given that the movement of the arm in a particular way results from a context-specific belief, the movement of the arm in that way is not the satisfaction condition of the desire for food.[23]

Papineau's approach to the determination of the content of desires may well succeed; at least, I shall not focus on any objections aimed particularly at that part of his theory. Instead, let us reconsider Papineau's attempt to derive the truth-conditions of beliefs from the satisfaction-conditions of desires, for this seems to yield the wrong truth-conditions even assuming the determinacy of desire content. Suppose that our subject Joe has a desire for ice cream. What belief would guarantee the satisfaction of the desire? Could it be a belief with the content *I should reach to such and such spot and move my arm at such and such angle of rotation is such and such direction*? It would seem not; for if Papineau genuinely wants truth-conditions to be those conditions that *guarantee* success in satisfying the apposite desire, the belief in question must represent all conditions required to hold in order that success be met: such conditions include extreme conditions such as that no lightning bolt will strike the ice cream and that no assassin will gun Joe down on his reaching for the cone, but also more mundane conditions, for instance, that the cone will not tip over while Joe is reaching for it and that no competitor will reach first. Presumably, though, it would be mistaken to assign the conjunction of all of these conditions as the content of the belief in question. This constitutes an outstanding problem for Papineau's teleological CT.

One noteworthy, and perhaps under-explored, version of teleosemantics is ahistorical; it emphasizes current maintenance and reproduction over historical selection (see Schlosser, 1998). On this view, the function of an existing mental structure is determined

by what it contributes to the ongoing maintenance of the system of which it is a part or by its contribution to its own continued maintenance. It is not so clear that this approach avoids the Cummins-style objection raised above to Millikan's theory: a structure or system remains in place *because* it corresponds to or carries information about the world, one might well worry, not vice versa.

4. Combined views

A number of CTs combine elements of the teleological and causal-informational views and do so to a greater extent than the theories discussed above.

4a. Stampe

Dennis Stampe (1977/79) offers what was perhaps the first causal theory clearly meant to apply to mental representations. All representation is causal, Stampe claims, including mental, or psychological, representation (*ibid.*, pp. 81-82). Stampe holds that for *R* to represent that an object has a certain property, that object's having the property in question must cause *R* to have certain of its properties; furthermore, and the object's having the property in question must have its effect on *R* in the appropriate way. According to Stampe, the appropriate-*ness* of a way is rooted primarily in the preservation of isomorphism (*ibid.*, p. 85), i.e., in the idea that the relations between the elements in the representing structure mirror the relations between the elements in the thing represented. *R* must be isomorphic to what it represents and *R*'s having its structure must have been caused by the analogous structure in what *R* represents. (Compare: the elements of a photograph relate to each other in the same way that the elements of the photographed scene related to each other at the time the photograph was taken.)

To head off one kind of indeterminacy, Stampe adds a teleological twist. Take the tokening of th mental representation AARDVARK. The tokening of AARDVARK occurs, in most cases, as the result of a series of causes, i.e., a chain of causes and effects. The problem arises that AARDVARK may be isomorphic to more than one element in that causal chain; furthermore, in the case of each such element, that element may well have caused AARDVARK to have the relevant structure (i.e., the structure it has such that it is isomorphic to the state of affairs in question). For example, if AARDVARK is a mental image, it may be isomorphic to a certain pattern of retinal firing that caused AARDVARK, but the image might also be isomorphic to the painting that caused the pattern of retinal firing in question. How is it that one of these elements constitutes AARDVARK's referent and the others do not? Here Stampe adverts to the function of the kind of representation involved (*ibid.*, 83-84, 91, 93-94); it is presumably the function of certain perceptual representations to be isomorphic to objects in the immediate environment, not to a structure of retinal firings (see Cummins, 1996, for a more elaborate version of this kind of view).

Teleology also plays a role in Stampe's explanation of misrepresentation. The content of *R* in a given case is determined by a reasonable hypothesis about what would cause *R* were conditions normal – in Stampe's terminology, were "fidelity conditions" to hold (*ibid.*, 88-89). Which conditions count as fidelity conditions is itself determined by the function of the representation in question (*ibid.*, 90).

Stampe's theory seems underdeveloped in at least two respects. First, the emphasis on isomorphism limits the theory's plausible application to certain sorts of complex representations. An atomic representation has no elements and thus is likely to be

isomorphic to too many things or too few, depending on what it should represent. If an atomic representation is supposed to refer to something unstructured, then we might worry that too many structure-less things will have caused the tokening of atomic representation *R*, and it seems unlikely that teleology can home in on just the right one. If the function of atomic representation *R* is to be isomorphic to something complex, then *R* is bound not to refer to anything; *R* has no parts that can be related to each other in the way that the relevant parts of the thing to be represented relate to each other. An attempt to apply Stampe's view to sentence-like mental representations will likely face similar problems.

Second, the talk of reasonable hypotheses should be made more precise. Stampe's decidedly epistemic and intentional language must be cashed if his CT is to provide a naturalization of content. Let us turn, then, to Dretske's more recent offering, a quasi-teleological revision of his informational view.

4b. Dretske's quasi-teleological view

In Dretske's book, *Explaining Behavior* (Dretske, 1988), information continues to play a central role, although it is called 'indication' (*ibid.*, pp. 56-58). Here, however, Dretske also appeals to functions. Consider an indicating structure *C* in an animal. Such a structure can be thought of as a mere detector: when it lights up, it has detected the presence of whatever's presence is guaranteed by that structure's lighting up. By individual learning, such an indicator can acquire a function within the cognitive system. *Because* its indicating some feature *F* on a particular occasion led to successful behavior, *C* can acquire the function of causing whatever it caused (e.g., a particular kind of movement *M*) that contributed to the behavior in question. In Dretske's words, "Once *C*

is recruited as a cause of *M* – and recruited as a cause of *M* because of what it indicates about *F* – *C* acquires, thereby, the function of indicating *F*" (Dretske 1988, p. 84). The content of, e.g., WARMTH is whatever WARMTH indicated when its connection to a movement was reinforced and is such that WARMTH's indication of that property explains why WARMTH acquired a new causal role (via reinforcement) in the cognitive system of which it is a part: in the case of WARMTH, such a role might be to cause movement in the direction of the property indicated. Misrepresentation occurs when *R* is later applied to something other than that the indication of which explains why *R* acquired its role in the cognitive system – if, e.g., WARMTH is later tokened when the organism comes into contact with a cold item.

4d. Objections

In some important respects, Dretske's quasi-teleological view is no more realistic than his purely informational theory. Human learning does not begin with perfect indication of whatever it is response to which causes desirable results (Slater, 1994). This is not to say that Dretske's general picture misfires. The following story may hold true often enough. The subject tokens a mental structure *R* that is in some way sensitive to *P*s. A series of such events then leads, via reinforcement, to *R*'s causing of *P*-appropriate behavior. Nevertheless, this story neither requires indication nor does it typically involve indication. The child, for example, can learn how to respond appropriately to birds, even if her BIRD-tokenings do not indicate birds – i.e., even if BIRD would have been caused by a bat had she walked onto the porch five minutes earlier than she actually did. Furthermore, the connection between the behavior in question (child says, "bird") and its reinforcement (father says, "yes, bird; good work, Molly") can be explained specifically

in terms of a bird's having caused the child's actual BIRD-tokening, not by a bird-or-bat's having caused it.

Given its requirement of actual learning in the subject, Dretske's view seems to preclude innate representations (Fodor, 1990, p. 41, Cummins, 1991, pp. 104-6), which comes as a surprise to those who, for varying reasons, think humans possess at least some innate representations.

Peter Godfrey-Smith (1991, 1992) offers one of the most influential critiques of Dretske's quasi-teleological view. This critique has potentially sweeping implications, as it appears to undermine any causal-informational theory that privileges one kind of statistical measure over others in the process of content determination. Godfrey-Smith argues that Dretske focuses too narrowly on indication – the measure such that if the concept is activated in the head, the property represented is guaranteed to be the cause of such activation (cf. Field, 1990, p. 108). Godfrey-Smith shows that, depending on the prevailing conditions in the environment and the value of successful behavior, a relation other than indication might better support successful behavior. If mates are not particularly hard to find but the cost of mating is very high, then it is useful to have a mate-indicator of the Dretskean sort: one that lights up *only* in the presence of mates. If, in contrast, mates are hard to come across and false attempts at mating are not particularly costly, then a different sort of measure becomes valuable: one such that if there is a mate in the environment, it lights up, even though it might light up in response many things other than mates. More generally, Godfrey-Smith emphasizes the contingent nature of the kind of statistical relation that will be of use to an organism. Indication might be most beneficial in some contexts and, for example, a high PRF in

others. Thus, any theory of content committed to one statistical measure is likely to assign contents incorrectly in at least some cases.

Godfrey-Smith's concerns are correct, in principle. In human systems in particular, though, there may be a constant value to one kind of statistical correlation. I suspect that, given (a) the associative mechanisms governing the strengthening and weakening of neural connections, (b) the typical structure of neural competition, and (c) the typical distribution of kinds in the environment, the statistical relations supporting adaptive human responses do not vary as much as Godfrey-Smith suggests. More importantly, if a theory of mental content solves numerous extant problems, we might do well simply to reject the assumption underlying Godfrey-Smith's critique: that contents are determined by whatever relation is most adaptive in the context in which subjects employ their mental representations. For example, even if indication is not always adaptively optimal, it is possible to motivate its theoretical role on grounds independent of adaptive optimality (for further discussion, see Rupert, 1999, pp. 332-39).

5. Two external objections

In closing, consider a pair of "external" objections to CTs, i.e., objections to the entire project of developing CTs. It is often thought that mental content has a normative dimension: when something has meaning, one *ought* to apply it in a certain way (Boghossian, 1989). The first external objection holds that no merely causal story, teleological or otherwise, can explain why we *ought to* apply a mental representation to one kind of thing but not a different kind.

The normativity objection seems to rest on a misunderstanding of the naturalistic project. First, the naturalistic project does not aim to capture *all* aspects of our intuitive

notion of thought-content. From the naturalistic standpoint, it is no surprise if no aspect of human psychology answers precisely to our intuitive notion of thought-content (or belief-content, or any related everyday notion). In fact, assuming that content – the actual thing we refer to when we use the term 'content' – is a natural phenomenon, it would be a surprise if our everyday concepts were to get things right in all respects. Commonsense views have turned out to be mistaken about a great many natural phenomena. On the naturalist's view, a certain property of mental representations has many of the features traditionally associated with the meanings of our thoughts. Mental representations have extensions, and these extensions contribute to truth-conditions of our thoughts (or of the sentences that express them). What is more, mental representations have more abstract content, which helps to determine the extension at various times in history. Such facts justify the identification of this as *mental content*, even if our best CT is inconsistent with some other intuitively appealing claims that have been made about meaning or about mental content – e.g., that it is normative in some special way that goes beyond that entailed by the possession of truth-conditions or correct conditions of application as these are characterized by a CT.

The second objection rests on the observation that, although a certain pattern of causal relations might be diagnostic of the presence of mental content, that pattern does not *constitute the nature of mental content*; the pattern of causal relations does not account for the nature of content "across possible worlds" as one might say (some version of this concern seems to be at work in recent critical discussions of CTs – see Bridges, 2006, Speaks, 2006). After all, content properties might play a certain causal role in the actual world, i.e., they might enter into a pattern of causal relations that is, in the actual world,

diagnostic of their presence; but what reason is there to think that all possible mental states with content play that same causal role?[24]

Naturalists should be wary of the presupposition behind the second objection: the assumption that mental content has a constitutive nature. Talk of constitutive natures carries modal commitments about which the most prominent naturalists, W.V. Quine, for instance, have long been skeptical (Quine 1960). Furthermore, natural scientists typically have little truck with constitutive natures, if such things carry weighty implications about thought experiments, essences, or all possible worlds. Thought experiments might help to generate hypotheses or explain theories in the sciences, but the theories in question answer primarily to events that occur in the actual world; such theories are not falsified by the results of thought-experiments, especially those involving events that differ greatly from the kinds observed empirically (e.g., a working cognitive scientist would, I suspect, balk at the idea that a theory of mental processing must answer to philosophers' intuitions about what would be true if rocks could talk). Thus, the naturalistically minded advocate of CTs can, on a principled basis, resist the demand that she deliver a theory of the constitutive nature of mental content; in contrast, she can stand in a boat with natural scientists, which is, presumably, where naturalistically minded philosophers mean to stand; in doing so, she can characterize content the same way physicists characterize fundamental particles: in terms of their causal interactions. Although this response might seem glib to some critics of naturalism, it seems dead-on if one's goal is to fit intentionality into the natural order.

Works Cited

Bennett, K. (2007) “Mental Causation,” *Blackwell’s Philosophy Compass*

Block, Ned. (1980) “Introduction: What is Functionalism,” in N. Block (ed.) *Readings in the Philosophy of Psychology*, Volume one (Cambridge: Harvard University Press), pp. 171-84

Block, N. (1986) “Advertisement for a Semantics for Psychology,” in P. French, T. Uehling, and H. Wettstein (eds.), *Midwest Studies in Philosophy*, Vol. 10: *Studies in the Philosophy of Mind* (Minneapolis: University of Minnesota Press), pp. 615-78

Block, N. (1990) “Can the Mind Change the World?” in G. Boolos (ed.), *Meaning and Method: Essays in Honor of Hilary Putnam*, (Cambridge: Cambridge University Press), pp. 137-170

Boghossian, P. A. (1989) “The Rule-Following Considerations,” *Mind* 98: 507-49

Boghossian, P. A. (1991) “Naturalizing Content,” in Loewer and Rey (1991), 65-86

Boyd, R. (1990) “Realism, Anti-foundationalism and the Enthusiasm for Natural Kinds,” *Philosophical Studies* 61: 127-48

Bridges, J. (2006) “Does Informational Semantics Commit Euthyphro’s Fallacy,” *Noûs* 40: 522-47

Cohen, J. (2004) “Information and Content,” in L. Floridi (ed.), *Blackwell Guide to the Philosophy of Information and Computing*, (New York: Blackwell), pp. 215-27

Cummins, R. (1989) *Meaning and Mental Representation*, Cambridge: MIT Press

Cummins, R. (1991) “The Role of Mental Meaning in Psychological Explanation,” in B. McLaughlin (ed.) *Dretske and His Critics* (Cambridge: Blackwell), pp. 102-17

Cummins, R. (1996) *Representations, Targets, and Attitudes*, Cambridge: MIT Press

Davidson, D. (1987) “Knowing One's Own Mind,” in the *Proceedings and Addresses of the American Philosophical Association* 60, 441-58

Davies, P. S. (1994) “Troubles for Direct Proper Functions,” *Noûs* 28: 363-81

Dennett, D. C. (1991) *Consciousness Explained*, Boston: Little, Brown and Company

Devitt, M., and Sterelny, K. (1987) *Language and Reality*, Cambridge: MIT Press

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

Dretske, F. (1988) *Explaining Behavior: Reasons in a World of Causes*, Cambridge: MIT Press

Dupré, J. (1981) “Natural Kinds and Biological Taxa,” *Philosophical Review* 90: 66-90

Elder, C. (1994) “Proper Functions Defended,” *Analysis* 54: 167-71

Field, H. (1990) “ ‘Narrow’ Aspects of Intentionality and the Information-Theoretic Approach to Content,” in E. Villanueva (ed.), *Information, Semantics & Epistemology*, Oxford: Blackwell, pp. 102-16

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

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

Fodor, J. A. (1994) *The Elm and the Expert: Mentalese and Its Semantics*, Cambridge: MIT Press

Fodor, J. A. (1998) *Concepts: Where Cognitive Science Went Wrong*, Oxford: Oxford University Press

Godfrey-Smith, P. (1991) “Signal, Decision, Action,” *Journal of Philosophy* 88: 709-22

Godfrey-Smith, P. (1992) “Indication and Adaptation,” *Synthese* 92: 283-312

Harman, G. (1982) “Conceptual Role Semantics,” *Notre Dame Journal of Formal Logic* 23: 242-56

Lettvin, J. Y., Maturana, H. R., McCulloch, W. S., and Pitts, W. H. (1959) “What the Frog’s Eye Tells the Frog's Brain,” *Proceedings of the IRE* 47:1940-1959

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

McGinn, C. (1989) *Mental Content*, Oxford: Blackwell

Millikan, R. G. (1984) *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press

Millikan, R. G. (1993) *White Queen Psychology and Other Essays for Alice*, Cambridge: MIT Press

Neander, K. (1995) “Misrepresenting and Malfunctioning,” *Philosophical Studies*, 79, 109-41

Papineau, D. (1984) “Representation and Explanation,” *Philosophy of Science* 51: 550-72

Papineau, D. (1993) *Philosophical Naturalism*, Cambridge, MA: Blackwell

Papineau, D. (1998) “Teleosemantics and Indeterminacy,” *Australasia Journal of Philosophy* 76: 1-14

Prinz, J. J. (2002) *Furnishing the Mind: Concepts and Their Perceptual Basis*, Cambridge: MIT Press

Quine, W. V. O. (1960) *Word and Object*, Cambridge: MIT Press

Rupert, R. D. (1996) “The Best Test Theory of Extension,” Ph.D. Dissertation, University of Illinois at Chicago

Rupert, R. D. (1998/99) “On the Relationship between Naturalistic Semantics and Individuation Criteria for Terms in a Language of Thought,” *Synthese* 117: 95-131

Rupert, R. D. (1999) “The Best Test Theory of Extension: First Principle(s),” *Mind & Language* 14: 321-55

Rupert, R. D. (2000) “Dispositions Indisposed: Semantic Atomism and Fodor’s Theory of Content,” *Journal of Philosophy* 98: 499-530

Rupert, R. D. (2001) “Coining Terms in the Language of Thought: Innateness, Emergence, and the Lot of Cummins’s Argument against the Causal Theory of Mental Content,” *Journal of Philosophy* 98: 499-530

Rupert, R. D. (2006) “Functionalism, Mental Causation, and the Problem of Metaphysically Necessary Effects,” *Noûs* 40: 256-83

Ryder, D. (2004) “SINBAD Neurosemantics: A Theory of Mental Representation,” *Mind & Language* 19: 211-40

Schlosser, G. (1998) “Self-Re-Production and Functionality: A Systems-Theoretical Approach to Teleological Explanation,” *Synthese* 116: 303-54

Slater, C. (1994) “Discrimination without Indication: Why Dretske Can’t Lean on Learning,” *Mind & Language* 9: 163-80

Smith, E. E., and Medin, D. L. (1981) *Categories and Concepts*. Cambridge: Harvard University Press

Speaks, J. (2006) “Is Mental Content Prior to Linguistic Meaning?” *Noûs* 40: 428-67

Stampe, D. W. (1977/79) “Toward a Causal Theory of Linguistic Representation,” in *Contemporary Perspectives in the Philosophy of Language*, P. French, T.

Uehling, Jr., and H. Wettstein (eds.) (Minneapolis: University of Minnesota Press), pp. 81-102 (page numbers refer to the 1979 edition)

Stanford, P. K., and Kitcher, P. (2000) "Refining the Causal Theory of Reference for Natural Kind Terms," *Philosophical Studies* 97: 99-129.

Usher, M. (2001) "A Statistical Referential Theory of Content: Using Information Theory to Account for Misrepresentation," *Mind & Language* 16: 311-34

---

[1] When a phenomenon has been successfully explained as part of this natural order, the phenomenon – intentional content, for example – is said to have been "naturalized."

[2] Unless otherwise indicated, terms set in all capitals refer to mental representations as psychological particulars (or their types). CAT refers the psychologically real structure we would normally take to refer to, or be about, cats. For some purposes, such units should by typed according to nonsemantic individuation conditions, i.e., conditions that determine when two mental representations are identical (or of the same type), but which make no reference to the *contents* of the units in question. This allows that two subjects could have mental representations with the same content even though the structures – often referred to as 'vehicles' – carrying that content differ (see, e.g., Rupert 1998/99, 2001).

[3] More subtle distinctions can be made, for instance, between current extension, timeless extension, and even the extension across all possible worlds; these differences – and the further puzzles to which they give rise – are here set aside.

[4] As I occasionally do below; 'represents' or 'refers to' frequently sounds much better than the clunky 'has the referential content that' or 'has the representational content that'.

[5] Those writing about CTs do not always explain what tokening amounts to. The straightforward – but largely misleading – view would present the phenomenon in terms of conscious experience: to apply, activate, or token a concept is to become conscious of the concept's meaning. To avoid circularity, this way of talking should be set aside; it presupposes quite a bit of machinery that CTs are meant to help explain. Better, then, to think in terms of a cognitive scientific model: perhaps tokening is the retrieval into working memory of a symbol structure stored in long-term memory; if a neuroscientific approach is preferred, tokening might consist in increased rates of synchronized firings of the set of neurons that realizes the mental representation in question.

[6] Compare Fodor's discussion of the Crude Causal Theory which he introduces as a first approximation to a CT (1987, p. 99).

[7] We might, for example, also want a CT to explain how mental content can be causally efficacious. Presumably, the *content* of human mental states causally affects human behavior, and a CT should render content's causal role intelligible. I leave my house whistling a hungry tune and end up at the local fast-food restaurant. Why do I wind up there rather than at the barber? Well, it is partly because I left home *thinking about* the local fast-food restaurant not the local barber shop. The question of how mental content can be causally efficacious raises a host of thorny issues, including general questions about the nature of causation. See Bennett (2007), for a survey of many of the relevant complications. Note that of the CTs to be discussed below, Dretske's second theory (Dretske, 1988) constitutes the most explicit and sustained attempt to satisfy this criterion.

Lastly, the scientifically minded philosopher might think that a satisfactory CT must meet the demands of successful cognitive science; in other words, a philosophical theory of mental content had better deliver something that can play the role asked of it by successful sciences of the mind (Cummins, 1989). It is, however, too big a project to spell out the achievements of cognitive science and to measure competing CTs against them; thus, this criterion shall be largely ignored in what follows,

[8] As is standard practice, I refer to Dretske's theory as a *causal* theory – a CT – but Dretske himself distances his talk about information from certain conceptions of causality (see, e.g., Dretske, 1981, pp. 31-39, 158).

[9] Three qualifications must be attached to all talk of guarantees and of conditional probabilities that equal one. First, this talk must be relativized to fixed channel conditions (or more broadly to an assumed context of transmission). Second, among these background conditions, we must include the state of the receiver; and when the receiver is a human, these conditions include the human's relevant knowledge, e.g., her knowledge about the source. Third, the state of a receiver does not carry information about states of a source when the probability of the states in question equals one regardless of the state of the receiver: it is irrelevant to our measurement of information carried by receiver state *R1* that the conditional probability of two plus two's equaling four given *R1* is one; after all, the probability of two plus two's equaling four is one no matter what state the receiver is in.

[10] Dretske's appeal to a learning period is one species of the general strategy of appealing to perfect covariation under optimal or ideal conditions (cf. Fodor, 1987, pp. 112ff). Such conditions are, however, notoriously difficult to specify without appealing to the mental states of the subject (or to the content of the concept itself – Cohen, 2004).

[11] Fodor has a variety of other reasons for wanting to couch AD in terms of laws; see, e.g., Fodor 1990, p. 100-03.

[12] I am treating AD as a theory of content for humans' mental representations. This runs counter to Fodor's own characterization of his project: he claims only to be offering sufficient conditions for intentionality, in the attempt to show that it is wrong to claim that intentionality is irreducible (Fodor, 1990, p. 96). If AD is supposed to state only a sufficient condition for the determination of mental content, then its failing to assign the correct contents to humans' mental representations is no strike against the theory. From the naturalist's standpoint, however, we can evaluate Fodor's proposal only by applying it to the sole example of a full-blown intentional system we know of, i.e., the human being. If Fodor's theory does not yield correct answers in the case of humans, it is not clear what would justify a naturalist's claim that the holding of AD suffices for mental content.

[13] Sometimes Fodor leans toward a view according to which the nomic connections are metaphysically basic, i.e., not to be reduced to any other way of talking; in which case, he can stipulate that certain nomic connections are dependent on the others. This may be unobjectionable, so long as one takes AD to state only sufficient conditions for intentionality, without regard for the human case in particular; but see note 10, above.

[14] This circumvents Fodor's worry about productivity – 1994, pp. 90-91.

[15] The causal-developmental view can be naturally extended to the case of individuals, though I limit the discussion in the text to kinds and properties. Similar comments apply to most of the views canvassed here, e.g., AD.

[16] A pair of complications: When no *Q* has ever caused S to token any mental representation at all, the rule as stated says that the efficiency rate of *Q*s relative to *R* in S = 0/0. The efficiency rate in such cases equals 0 (by stipulation, but in keeping with the spirit of the causal-developmental approach). Also, I do not mean to suggest that an individual item causes the tokening of only one mental representation per causal interaction with S. Frequently, if not always, the subject's reaction to an object she encounters includes a multitude of conscious and subconscious associations. Thus, I propose to complicate the calculation of success rates in the following ways: (1) any causal interaction that includes a *Q*'s causing S to token *R* (among, perhaps, other mental representations) is counted in the numerator, (2) any causal interaction that includes a *Q*'s causing S to token *R* (even if the subject tokens other terms as well) is counted only once in the denominator, and (3) in the event that a member of K causes the tokening of one or more mental representations other than *R*, this event is counted only once in the denominator. This way of counting has, among its other effects, the result that the tokening of a mental representation other than *R* in response to a *Q* does not lower the efficiency rate of *Q* relative to *R* in S, so long as S also tokened *R* on that occasion.

[17] Some of these criticisms of the causal-developmental view, as well as possible responses to them, raise general issues in metaphysics that have implications for most, perhaps all CTs. For instance, the plausibility of a CT may well depend on the range of natural kinds, or more broadly, mind-independent properties, that, in fact, exist (cf. Dupré, 1981). A paucity of such properties would seem to spell trouble, but so would a surfeit; for if there are far more objective properties that we would have expected, this

increases the likelihood that a CT might, together with the empirical facts, entail strange and intuitively unacceptable content assignments.

[18] Ryder's argument is actually a bit more complex: he adds the teleological claim that the neural system has the function of making predictions and that the correlation discussed in the text is the right one to focus on when giving a theory of content because this correlation explains why cells contribute successfully to the neural system's predictive capacities (2004, p. 229).

[19] This is also the notion of a function most closely associated with CTs. Other notions of a function have been discussed in connection with mental content, however. It has sometimes been suggested that a theory of content be grounded in computer science's conception of logico-mathematical functions (Cummins, 1989). Similarly, some conceptual role theories emphasize functions in the sense identified with widely known functionalist positions in philosophy of mind (Harman, 1982) or tied specifically to inferential-role aspect of functional profiles (see Block, 1986); see Block (1980) for general discussion of functionalism in philosophy of mind.

[20] See also McGinn, 1989.

[21] The precise definition of "reproductively established family" is fairly technical (see *ibid.*, pp. 27-28). For concerns about the viability of Millikan's definition of a reproductively established family, see Davies 1994. For a defense of Millikan's definition, see Elder 1994.

[22] Or might come to stand in developmentally; some mechanisms get selected for because they produce new tokens over the course of a subject's lifetime that map in some particular way onto the subject's environment.

[23] Teleological theorists often add a further point, which might – independently or in conjunction with Papineau's strategy – help to winnow contents. There is a sense in which evolutionary selection does not "care" whether the frog represents flies as flies or as small dark things (Lettvin, Maturana, McCulloch, and Pitts, 1959); either representational content suffices for the frog to stay alive and reproduce. Nevertheless, if one emphasizes evolutionary *explanation*, a distinction between small dark things and flies quickly emerges, so long as there are some small dark things in the frog's environment that are not flies. Flies exhibit properties (pertaining to their chemical constitution) that explain the frog's continued existence on the basis of the frog's having consumed a fly. Nothing about small dark things as a kind of thing explains why frogs who snap at them flourish; to the extent that there is available such an explanation, it invokes the chemical properties flies have in virtue of their being flies.

[24] This objection is sometimes discussed in connection with Donald Davidson's example of the hypothetical swampman (Davidson, 1987), a being that materializes suddenly as Davidson goes out of existence but which is a perfect duplicate of Davidson. Swamp-Davidson has no history, the concern runs, but it certainly has mental states with content (many philosophers, although not Davidson, are inclined to think). No theory appealing only to actual historical relations can respect this judgment, for swamp-Davidson has no history, selectional or otherwise. For detailed discussion, see the swampman symposium in *Mind & Language*, 1996; for present purposes, note especially Dennett's contribution.