INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

THE NATURALIZATION OF INTENTIONALITY

by

JONATHAN PETER WILCOCK

A dissertation submitted to the Graduate Faculty in Philosophy in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2001

This manuscript has been read and accepted for the Graduate Faculty in Philosophy in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

July 26, 01
Date

Chair of Examining Committee

July 26, 01
Date

Executive Officer

Professor David Rosenthal

Professor Michael Levin

Professor Michael Devitt

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

THE NATURALIZATION OF INTENTIONALITY

by

Jonathan Peter Wilcock

Adviser: Professor David Rosenthal

Quine's thesis of the inscrutability of reference provides grounds for doubting that the project of naturalizing the contents we ascribe in interpretation will be successful. Quine argues that it possible to give multiple adequate interpretations of a language on which some sub-sentential linguistic expressions are given incompatible interpretations. If so, such expressions fail to have determinate content. The strongest objections to the thesis consist in the exhibition of a sentence that will be assigned different truth-values by the standard and a deviant interpretation. It is shown that there is an interpretation of the apparatus of individuation, in particular the quantifier, which ensures that each whole sentence in a deviant interpretation has the same truth-value as the corresponding sentence in the standard interpretation.

A weaker basis for resisting Quine's thesis is to suppose that some further evidence is to be found that will tell between the standard and deviant interpretations. A consideration of proposed naturalistic theories of intentionality does not support this supposition. The most promising contemporary proposal is

teleological. It is shown that the classes that the most promising theory of teleology assigns as contents diverge from those that we assign in interpretation, so the teleological proposal does not provide an adequate basis for a theory of the contents we actually ascribe.

Another reason for rejecting Quine's thesis is because it is thought to imply some form of anti-realism. On my defense of the thesis, however, objects that are quantified over in the standard interpretation are quantified over in the deviant interpretation. Therefore anti-realism does not directly follow.

Acknowledgments

I thank my adviser, David Rosenthal, for his considerable assistance of both a philosophical and a practical nature. His contributions significantly increased not only the quality of this dissertation but also the probability of its completion. I also thank Michael Levin for providing me with a great deal of useful comments on my work and Michael Devitt for generously providing his time to serve on the committee. Thanks also to John Greenwood, both for acting as a reader and for his support over many years. This work has benefited from the influence of Galen Strawson, Brian Loar and Seth Crook.

Table of Contents

CHAPTER 1

INTENTIONALITY AND EXTERNALISM

Intentionality is the property certain things have of representing other things. The problem of intentionality is that of accounting for our having thoughts about anything. Some philosophers (most notably Brentano) have claimed that there is something puzzling, if not downright mysterious, about intentionality. In contrast to many philosophical problems, however, there is no paradox of intentionality which suggests this. It is worth considering what might lead someone to think this.

One may try to explain the intentionality of thoughts on the basis of non-mental representations, such as words and pictures. Given that a word is nothing but a set of marks on paper or a certain sound, there is nothing intrinsic to the word itself that makes it refer to anything. This may lead one to believe that words get their references from the intentions of their users to refer to various things. If so, the intentionality of the mental states of language users cannot be explained in the same way as the intentionality of their language, for in order to have intentions to refer to a thing it is necessary to already be able to think about it.

Those other non-mental representations, pictures, have no more intrinsic connection to their subjects than do words. A picture of the Eiffel tower in Paris, for example, does not derive its representational power from any resemblance it has to that structure. If one were to build an exact replica of the Eiffel tower and photograph it, one would have a picture identical to pictures of the actual Eiffel tower, without that picture being a representation of the Eiffel Tower. If mental states are physical objects like pictures and inscriptions of words, then they too have no more intrinsic connection to the things that they are about than non-mental representations.

This conclusion only indicates that intentionality is puzzling if there is some reason for supposing that mental states have an intrinsic connection to the things they are about, but it is far from obvious that this is the case. In fact, the most promising proposal for accounting for intentionality is one on which mental states are intentional owing to their causal relations to certain objects and other mental states. These sorts of views are externalist in that they account for our mental states possessing content in terms of the relations of those states to things external to the thinker

The Rationale for Externalism

Externalist accounts of intentionality have been largely inspired by externalist theories of reference for public languages. These theories of reference owe their popularity primarily to the work of Kripke and Putnam. In Naming and Necessity, Kripke argues that we should not understand proper names as referring in virtue of their association with definite descriptions but rather in virtue of their standing in certain causal relations to their referents. Kripke suggests that speakers stipulate that a word is to be used for a certain object (or kind of object) in an initial 'baptism'. The object is initially picked out either by ostension or description. Speakers may come to have all sorts of false beliefs about this object, but so long as they acquired the term in an appropriate way, there will be a causal chain of the right sort stretching back from their use of the term to the baptism, and hence to the referent.

In 'The Meaning of "Meaning"', Putnam argues that reference cannot be accounted for in terms of states internal to a thinker/speaker. The main basis of this claim is his Twin Earth thought experiment, on which the references of the terms of speakers on Earth and Twin Earth differ even though their psychological states, considered purely internally, are identical. Putnam concludes that the

environment makes an ineliminable contribution to the determination of reference, or, as he puts it, '"meanings" just ain't in the *head*!'. In his final 'reconstruction' of the notion of meaning in 'The meaning of "Meaning"', Putnam suggests that a description of the meaning of a predicate like 'water' has four elements: syntactic markers, semantic markers, a stereotype and the extension of the predicate.[1] The syntactic markers reflect the type of term that it is, e.g. a mass noun, the semantic markers reflect what type of thing is in the extension, e.g. a liquid, and the stereotype contains information about the extension that the community requires an individual to believe in order to be considered as competent users of the term.

Not all of these elements are relevant to the determination of reference. The stereotype, for example, may attribute properties to the extension that it does not have, and yet the term may still refer. The semantic markers do contribute to the determination of the extension of the term. This is because in order to extend the reference of a term from an initial sample of a substance, it must be specified in what respects other samples which fall under the term are similar to it. Putnam at one point sums up his account of reference by saying that he has uncovered a previously overlooked indexical component in the meaning of predicates like 'water'.[2] This way of putting matters is misleading, for it suggests that the referent of 'water' will change according to its circumstances of utterance, but this is not what Putnam intends. Rather, his point is that for something to fall into the extension of 'water', it must bear the same-liquid relation to what is called 'water' around here. This is not to offer a theory about how reference for 'water' is originally fixed, but to insist that there is only one kind of water (i.e. the kind that

---

1. Hilary Putnam, 'The meaning of "Meaning,"' in Mind, Language and Reality (Cambridge: Cambridge University Press, 1975), 269.

2. Ibid., 234.

we have here) and anything that isn't the same liquid as that which we call 'water', isn't water.

Putnam 'heartily' endorses Kripke's assertion that descriptions don't fix reference (or at least if they do, they are 'rigid' descriptions) and states that both he and Kripke are making the point that natural-kind terms are rigid designators.[3] There is a gap between this and the conclusion that Putnam supports Kripke's account of the determination of reference. In only slighter later work, Putnam says that in 'The Meaning of "Meaning"' he was offering only a theory of how reference is to be specified, rather than a theory of what reference is.[4] So when he explains, in 'The Meaning of "Meaning"', how one can give a definition of 'water' by pointing to a glass of water and saying 'This is water', this is an account of how one might explain to another what one refers to by 'water' and not an account of what makes it the case that 'water' has the reference that it does.[5]

Despite this reluctance of one of the originators of the causal theory of reference to commit to the theory, Kripke's and Putnam's examples are convincing that the environment and not the way in which the speaker conceives of the world plays the major role in determining reference for public language.[6] Clearly one cannot account for the semantic properties of mental states in terms of thinkers' conceptions, for the very issue is what it is that makes it the case that there are any concepts (and hence thinkers). Having drawn the conclusion that the environment supplies linguistic content, and if one rejects the view that mental content is

---

3. Ibid.

4. Hilary Putnam, Meaning and the Moral Sciences (London: Routledge and Kegan Paul, 1978), 58.

5. Putnam, 'The Meaning of "Meaning"', 230-232.

6. Kripke also expresses reluctance to commit to a theory of reference in Naming and Necessity, claiming instead to be offering only a 'picture'. There seems to be less substance to this than in Putnam's case, however.

inherited from linguistic, then it is enormously plausible that mental content is to be accounted for causally.

The Rationale for Internalism

The primary grounds for finding the externalist picture wanting are concerns with psychological explanation. Consideration of Putnam's twin earth case has suggested to some that, although people here on earth have different externally determined contents to those on twin earth, the way we would explain their actions is the same in each case. It is suggested that there is some aspect of content which these people share and with which we are concerned in psychological explanation[7]. There are really two main varieties of internalism: one which holds that in addition to externally determined content, there is another type of content, 'narrow content', and another variety which holds that at least some contents have no external elements in their determination. The most extreme form of this latter type of internalism holds that no content is externally determined.

Those who adhere to the former variety may do so without disputing the main element of externalism: that the only way a mental state may possess a content is owing to its relations (usually causal) to the extra-mental world. Their disagreement with the externalists consists in their assertion that there is some other property relevant to being in a particular contentful state (such as a having a belief). My concern in this essay is not with this dispute between the defenders of narrow content and the advocates of a purely broad content. Rather, my concern is with what determines that a state has any sort of content at all. For this purpose,

---

7. E.g. Fodor, 'Methodological Solipsism Considered as a Research Strategy in Psychology', Loar, 'Social Content and Psychological Content'. Loar also argues for this claim on the basis of his consideration of a variation on Kripke's 'Pierre' puzzle. In his presentation of Putnam's example, Dennett suggests the twins are psychological duplicates in virtue of being physical duplicates (The Intentional Stance, 127).

the views may be divided into those of the externalists, including those who mix externalism with some notion of narrow content, and those who take there to be some contents which are determined purely internally.

Internalists in this sense are at present considerably less numerous than externalists. It would be misleading to say that the reverse used to be the case, as it seems as though internalism is so prevalent in the history of philosophy that an alternative had rarely been considered. Descartes, for example, wishing to suspend belief in all that may be doubted, is willing to entertain the possibility that, so far as he knows, there is no world beyond his mind. Yet he does not consider the implications of this for the contents of his mind, assuming that he would nevertheless be able to think about spatial objects such as trees and houses even if there were no material world at all.[8] Locke and the Eighteenth Century empiricists follow Descartes in this. Showing that there are any material things at all becomes a concern for them, but not the issue as to whether they even have the concepts of material objects.[9] Whatever the basis of the historical popularity of the view, unlike externalism, there is no widely acknowledged contemporary source. The views of some of the leading contemporary internalists will be considered below.

One of the most widely discussed arguments against the possibility of constructing an externalist account of intentionality is Searle's Chinese room argument. Searle himself indicates that it is to be understood in this way:

> Suppose for example you had a perfect causal externalist account of the belief that water is wet. The account is given by stating a set of causal

---

8. Putnam, noting this consequence of extreme skepticism, and the fact that we can think about such material things, infers that such skepticism is false in Chapter 1 of Reason, Truth and History.

9. Hume is in some measure an exception to this, for in the Treatise he considers the question of how, given the empiricist assumptions of the theory of ideas, the idea of permanence is possible.

> relations in which a system stands to water and to wetness and these relations are entirely specified without any mental component. The problem is obvious: a system could have all of these relations and still not believe that water is wet. This is just an extension of the Chinese room argument . . . .[10]

The Chinese room argument is based on the following thought experiment: Suppose that someone who knows no Chinese is locked in a room with a set of instructions (in a language he does understand) for outputting strings of Chinese characters in response to strings of Chinese characters he receives. Unknown to the person in the room, the characters he receives are questions in Chinese, and the characters he outputs are answers, in Chinese, to these questions. If the instructions are good enough, and the person locked in the room executes them flawlessly, the answers received will be indistinguishable from those that would have been received from a native speaker of Chinese. Clearly, however, the person in the room does not understand Chinese. He does not even need to know that the characters he is manipulating according to the rules are Chinese, or belong to a language. Searle's point is that a computer is no better off than a person in the Chinese room. Just as the manipulation of symbols that are, for him, uninterpreted is not sufficient for understanding Chinese, no machine that operates by performing operations on formally specified elements can understand Chinese.

Whatever plausibility this argument possesses is, I believe, largely acquired from the fact that it appears to the person in the room that the symbols he is manipulating have no interpretation. Being locked in the room, he is unable to see what correlations there are between the symbols and the various things they

---

10. Searle, The Rediscovery of the Mind (Cambridge, MA: MIT Press, 1992), 51. Immediately subsequent to this passage, Searle claims that a symptom of this problem is that intentional notions are normative in that they set standards such as truth conditions which a system of nonintentional causal relations cannot do.

represent in the Chinese language. Of course, this does not mean that they are not there, but neither does it mean that the correlations are not actually sufficient for providing an interpretation. The most the Chinese room argument could show is that the semantic properties of mental states are not to be accounted for in purely syntactic terms. An externalist may concede this to Searle without compromising his externalism, for the externalist will maintain that the semantic properties are supplied by the causal relations of mental states to the extra-mental world, not by their syntactic properties alone.

Others have thought that conscious experience has some key role to play in intentionality. In his Direct Reference, Récanati suggests as a motivation for internalism what he calls 'the Cartesian intuition', which is the claim that when two individuals have qualitatively identical mental states then their states share a content. He gives as an example the distinction between veridical perception and hallucination. From the subjects' point of view, everything seems the same, and so for those who share the Cartesian intuition, these individuals will have thoughts similar in content, in some respect. In the twin cases, it is similarly said, everything seems the same from the subjects' perspectives and so by the Cartesian intuition they will share a content.[11] Récanati himself seeks to make this intuition compatible with externalism. He agrees that the twins share a (narrow) content, but attempts to account for this on externalist grounds. As the twin-earth story is

---

11. While the notion of things 'seeming the same' is fairly clearly applicable in veridical/non-veridical cases, there is less to it than may appear. If this is purely phenomenological, i.e. color patches in one's visual field and the like, then some explanation is owed as to how this is relevant to semantic properties. The externalist argument against such properties on their own having any semantic import is compelling. If instead, the similarity is a conceptual one, it is not at all clear that things do seem the same, despite the phenomenological similarity. This is particularly clear in the case of brains in vats. The notion of 'bracketing', which even Putnam appeals to, makes little sense here.

ordinarily told, the twins will fairly non-controversially share many concepts: *lake*, *river*, *sky*, *faucet*, etc.[12] Récanati suggests the twins' concepts of *water* and *twin-water* should be regarded as similar owing to their relation to these other shared concepts. Whether this sort of account could reconcile the Cartesian intuition and externalism would at least depend on ruling out the possibility of ungrounding all or sufficiently many concepts at once, as is done in philosophical thought-experiments such as the brain in the vat and the instant person (e.g. Davidson's swampman example). Récanati assumes that whenever there is a similarity in phenomenological history, this must have been brought about by interaction with similar types of material objects. Ordinarily we would expect this to be the case, but not necessarily.

Galen Strawson has recently maintained that experience is directly relevant to intentionality. He defends the claim that 'there is no deep problem or puzzle of intentionality that is genuinely distinct from the problem of accounting for experience, so far as the task of giving a naturalistic, materialistic account of the mind is concerned'.[13] He dubs this claim 'the no-problem thesis'. Strawson divides intentionality into the type had by a mental state when it is about something real and when it is 'about' something nonexistent (i.e. when the representation fails to have a referent). These kinds of intentionality may be further sub-divided into states about concrete objects (E/C or N/C) and states about abstract objects (E/A or N/A).

Strawson asks us to imagine two human beings who have qualitatively similar mental episodes, as-of thinking about a statue on Easter Island. X has a

---

12. This is not entirely non-controversial. If one thinks of concepts as functional roles, then my concept of a lake and my twin's concept of a lake will be different, owing to my concept's relation to the concept *water*, which is a concept my twin lacks.

13. Strawson, Mental Reality (Cambridge, MA: MIT Press, 1994), 178.

normal causal link to the statue, although he cannot remember his seeing it, whereas Y's experience is produced randomly with no causal link to the statue at all. Strawson agrees with the externalist that X's experience has E/C intentionality, whilst Y's fails to have it. Strawson accepts that there is nothing puzzling about how X's thought is truly about a statue on Easter Island and Y's isn't (although he adds the caveat that this is so only once one has subtracted the problem of accounting for conscious experience in physical terms). The difference consists in the way these experiences are caused, and this is just a relation of 'routine causation', of the type that paintings and photographs bear to their subjects. The difference between X's and Y's experiences is substantially the same difference that makes one painting a portrait of a particular individual whereas a qualitatively identical one is not, when the latter is a work of imagination and is not based on any real person at all.

One way to pose the problem of intentionality is to ask how it is that a mere act of thought can make a connection to an extra-mental object. The answer, endorsed by Strawson, is that it doesn't. It is correct to suppose that there is no deep puzzle about why X's state represents that particular statue rather than something else, and why Y's state fails to represent anything at all. However, what is not yet accounted for is what makes X's state a representation of a thing as being a statue, or a representation as-of a statue. This is a feature which, according to Strawson, Y's state shares.

X and Y are identical in that it seems to them that they are thinking about a particular thing. They both have 'the same experience of subjective conviction that a particular object is targeted in thought'[14]. I take Strawson to be quite right to suppose that nothing about the character of X's experience, or any neurological

---

14. Ibid., 181.

fact about X, makes it the case that his thought is about the statue 'it purports to be about'. What is less clear is what makes it 'purport' to be about something. The causal history is what makes the difference between X having a genuine representation and Y a failed representation, but this cannot account for the 'purported' intentionality of their experiences, for their experiences share this feature.

Strawson's claim is that whatever else is puzzling about the experience of X, its having E/C intentionality is not. It seems to me that a better way to put this is to say that it is not puzzling why X's intentionality is of the type E/C rather than N/C. This does not account for why X's mental state has intentional properties, but rather why it has the particular intentional property that it does rather than a different one. This leaves it open whether intentionality is just a matter of dispositions and causal history, such that an experienceless being could be in fully intentional states. Even if one does not think this, and holds that experience is necessary for full intentionality, as Strawson does, then although there may be no deep problem in explaining how X has one type of intentionality rather than the other type that Y has, it does not follow that there is no deep problem of explaining how their states come to be intentional at all.[15]

Having maintained that causal history is what marks the difference between a mental state with reference and a qualitatively identical one which does not refer, one might wonder why Strawson supposes there to be a puzzle about intentionality at all. In order to clarify what he believes is left problematic, Strawson asks us to consider a person who has just sprung into existence, but is otherwise a duplicate

---

15. What would make the one problem deeper than the other is that in the case of explaining why one state has the referent it does rather than none, although we are perhaps unable to exactly specify the causal relation involved, we have a good idea of what sort of explanation would be satisfactory.

of an ordinary person. Even if one supposes that such an instant person does not have E/C intentionality owing to a lack of regular causal connections with the rest of the world, some sort of intentionality is present in him, according to Strawson. He supports claim by pointing out that it is conceivable (presumably he also believes it possible) that one actually is an instant person. If the instant person does have intentional states, then the content of these states is, of course, independent of their causal origin.

As some of the mental states of the instant person are both intentional and fail to have a reference, one might anticipate that Strawson's account of N/C content would apply to them. Strawson presents a sort of Humean theory of ideas to account for the N/C intentionality that some of our mental states ordinarily have: when we have N/C intentionality, such as in imagining the existence of a platinum coathanger, we take elements of our imagined objects from the realm of existing things, e.g. platinum and coathangers. So standard N/C intentionality involves E/C intentionality. This is a very close relation to Récanati's account of the externalist grounding of narrow content. One may hesitate to assign genuine N/C content to the instant person, as he lacks the relevant causal histories to ground his states in concepts with E/C intentionality. Récanati himself would take this view. Strawson's reply to this is that 'whatever [the instant person] has, it may be all that you have, and it deserves a decent name'.[16] Even so, if this type of state is genuinely intentional, some account of its intentional properties needs to be offered.

Strawson wants to make this sort of N/C intentionality a matter of experience. '[T]his richness is, in the end, just a matter of qualitative character. It is just a matter of experiential what-it's-like-ness for a subject from moment to

16. Ibid., 202.

moment'[17]. Consider a robot (well) designed to find a three-inch blue cube and a human being instructed to find such a cube. If one supposes that the robot fails to have genuine intentionality (as Strawson does), one might ask what it is that the conscious experience of the human being would add to the capacity of the machine to locate three-inch blue cubes. Strawson suggests as a 'powerful intuition' that the human being really has a three-inch blue cube in mind, whereas the machine doesn't really have anything in mind at all. Really having a three-inch blue cube in mind presumably amounts to consciously thinking of the thing that one is seeking as a three-inch blue cube.

The point of Strawson's argument is to defend the no-problem thesis: that if intentionality poses a deep problem for naturalistic materialism then it is part of the problem posed by experience. The claim is that once experience is put aside, there is no special problem of accounting for 'aboutness', for then one thing's being about another is just a matter of the two things standing in the right causal relations: 'The only deep puzzle for naturalistic materialism is the old puzzle. It is the scientific (rather than the philosophical) puzzle posed by the very existence of experience, given present-day science'[18]. In order to make these claims, Strawson must have shown that the N/C intentionality he attributes to beings that lack the necessary causal relations for both E/C intentionality and the sort of N/C intentionality that we ordinarily possess is entirely owing to their enjoying certain types of experience.

What is unclear is how it is that merely enjoying certain types of experience can give rise to intentionality. One gets the impression that Strawson wishes to put off this puzzle as part of the problem of explaining consciousness, which he does

---

17. Ibid., 196.
18. Ibid., 214.

not claim to be able to solve. However, that problem is primarily that of explaining why it is that beings in certain physical states are in certain conscious states, i.e. the 'old puzzle' of consciousness is the problem of explaining the existence of consciousness. The present problem is quite another, which is how it is that a conscious state comes to have intentionality, or what it is that makes a conscious state an intentional state. There is no logical guarantee, or even a good reason to believe, that a solution to the former problem will constitute a solution to the latter.

The Contribution of Experience

Although it seems obvious that we do have conscious states which are intentional, it is problematic to suppose that they are so essentially. Putnam, for example, has given the basic argument against this view.[19] Mental states can be no more essentially referential than anything else. Random splashes of paint that happen to resemble Winston Churchill do not constitute a picture of Winston Churchill. There does not seem to be any obvious reason to regard mental images as any better off in this respect than physical ones. Strawson would accept that if someone had only mental episodes merely as-of Churchill, those states would not possess E/C intentionality, which depends on causal relations, but would insist that they would still have some sort of N/C intentionality.

At times, Putnam seems to say no more than that representations do not have E/C intentionality without standing in the relevant causal relations. That representations do not intrinsically refer is clearly a conclusion he draws from consideration of the Twin Earth cases. If this is all that Putnam is saying, it leaves

19. Putnam, Reason, Truth and History (Cambridge: Cambridge University Press, 1981), chap. 1.

Strawson's position open, which is not that wide content, or E/C intentionality, is independent of external relations, but something else, narrow content or N/C intentionality, is.

Strawson is far from alone is supposing that there is such a thing as narrow content. There is considerable variation in the form that theories of narrow content take, however. The primary motivation for narrow content is from concerns for psychological explanation and usually involves the assumption that psychology is methodologically solipsistic. Fodor argues for methodological solipsism on the basis that the correct theory of the mind is computational. Although Fodor believes that some mental states are representations, the properties of mental states directly relevant to an explanation of mental processes are not semantic ones. This is because that explanation is computational, and computational processes, although defined over representations, are formal in that they do not depend upon the semantic properties of representations.

Taking a methodologically solipsistic attitude to psychology does not determine how mental states used in psychological explanations are to be individuated. It is possible to individuate them formally, as Stich has suggested. Stich argues that most state types will have tokens which differ markedly in the contents which we ordinarily ascribe to them, and that some state types will have tokens to which we are unable to ascribe content at all. It is his position that the semantic properties we ascribe to mental states play no role in a computational theory of the mind, and so Stich recommends the abandonment of intentional explanations of behavior. Although Stich believes in 'narrow psychology' he does not believe in narrow content.

Those who do advocate the utility of narrow content in psychological explanations generally accept that narrow contents are to be understood as functional or conceptual roles, that is the narrow content of a particular state is

determined by its connections to other mental states, sensory inputs and behavioral outputs. A less widely held position is that a narrow content is a function from a context to a truth condition. In explaining this latter position, Fodor says that there is some fact about the relation between him and Earth such that his 'water' thoughts are about H_2O but not XYZ, and some fact about the relation between Twin-Fodor and Twin Earth such that Twin-Fodor's 'water' thoughts are about XYZ but not H_2O. The obtaining of these facts constitutes the relevant contexts. Anything which is, he says, neurologically identical to Twin-Fodor (such as Fodor), and is in the same context (i.e. stands in the same relation to Twin Earth as the relation that makes Twin Fodor's 'water' thoughts about XYZ), will have 'water' thoughts that have the same truth conditions as the 'water' thoughts of Twin Fodor.

On this account, two particular mental states that share a narrow content do not necessarily share semantic properties. This is only the case if the narrow content determines the same truth conditions in the context of each mental state. What the mental states share in virtue of sharing a narrow content is the same potential truth conditions, relative to context. So this position does not involve an internalist notion of content. The conceptual role account of narrow content is further removed from the semantic properties of the mental state than the function account as in standard versions of the theory, the conceptual role does not play a role in the determination of truth conditions and reference. The point of the theory is that (folk) psychological explanation individuates mental states according to their conceptual roles, rather than their truth conditions.

The fact that conceptual roles do not play a role in determining truth conditions may lead one to question whether they properly constitute a kind of content at all. Loar considers this question, and answers it by defending a notion of content that does not involve truth conditions. On Loar's view, conceptual

roles, although unconnected with the referential properties of one's thoughts, do determine how it is that one conceives things. The conceptual role of a mental state does not determine its truth conditions, but it does determine what Loar calls 'realization conditions.'[20] The realization conditions of a state are the conditions under which it would be true if the thinker's conceptions were not misconceptions. Loar adapts Burge's arthritis case as an example. The example concerns an individual, Bert, who believes that he has arthritis in his thigh. Suppose that you do not know whether in Bert's linguistic community 'arthritis' refers to arthritis, or to a broader class of rheumatoid ailments, but you do know that Bert believes 'arthritis' refers to the broader class. Whether Bert's belief is true or false depends on the meaning of 'arthritis' in his community. Loar's point, however, is that given the knowledge of the realization conditions of Bert's belief, even in the absence of the knowledge of its truth conditions one can understand psychological explanations involving the belief.

Strawson's N/C intentionality has in common with narrow content in Loar's sense that states which possess it exhibit intentionality independent of their truth conditions. Loar says that in at least some cases a person on Earth shares a narrow content with their doppelganger on Twin Earth, meaning that they conceive of their world in a similar way. Putnam would agree that there are cases where two individuals differ in their broad contents but have similar narrow contents.

One such example is that of two communities, one of which speaks English and the other speaks a language identical to English, except that the word 'elm' refers to beeches and the word 'beech' refers to elms. Although experts who know how to distinguish elms from beeches will think of these types of tree in different

---

20. Loar, 'Social Content and Psychological Content,' in Contents of Thought, edited by Robert H. Grimm and Daniel D. Merrill (Tucson: University of Arizona Press, 1988).

ways, those who do not but 'borrow' the concepts from the experts may not think of these trees very differently. For such nonexperts, the only difference between their concept of an elm and their concept of a beech is that one is called 'elm' and the other is called 'beech'. If one considers a nonexpert from each community, their 'elm' and 'beech' concepts are subjectively the same, but differ in reference. In an example like this, it is uncontroversial that the members of different communities share many other related concepts, such as the concept tree.

This sort of way in which ascriptions of narrow content may be supported does not support the N/C intentionality that Strawson advocates. Putnam actually does make a stronger claim than an internalist like Strawson would accept. In his well-known discussion of brains in vats, Putnam suggests that when a being who has always been a brain in a vat thinks 'There is a tree in front of me', the content is different from what is thought by a qualitatively identical being who has always been embodied. This is because reference for the brain in a vat will be causally determined and so the referents of its expressions will be something other than the referents of our expressions. In 'vat-English', 'tree' does not refer to trees, but something rather strange such as electronic impulses or some feature of the program of the machine sustaining the envatted brain, to 'trees in the image' as Putnam says.

Putnam and Kripke's arguments for externalism found widespread acceptance relatively quickly, probably because even committed internalists shared the intuitions on which their examples were based. Someone strongly inclined towards internalist intuitions may seek to dispute the interpretation that Putnam gives of the mental states of brains in vats. In his elm/beech example, Putnam argues that nonexpert speakers in the linguistically deviant community refer to elms with the word 'beech', as they intend to refer to what the experts refer to, namely elms. In the case of the brain in a vat there is no community of users for

Putnam to appeal to in order to support his interpretation. Putnam's ground for the interpretation he ascribes to the states of brains in vats is the degree of indexicality in natural kind terms he claimed to have discovered in 'The Meaning of "Meaning".[21] By this he means that the reference is fixed indexically: water is whatever bears the same-liquid relation to some particular sample, i.e. *that* stuff.

It is possible to question Putnam's interpretation of brai9ns in vats on this basis. A thing is not picked out indexically by its being the thing 9that causally prompts somebody to point in its direction. This is shown by the possibility of deferred ostension. I may, for example, point to a photograph of the White House and say 'That's where the President lives'. So if a brain in a vat is somehow able to conceive of the world as being spatial, it may be that its primary intention to refer to something spatial with its word 'tree' would prevent it from ostensively picking out some feature of the program of the machine as the referent of that word.

Let us put aside this concern and suppose that in vat-English, 'tree' actually does refer to trees-in-the-image. Could one take the view that this is purely a matter of wide content or E/C intentionality, and that the narrow content of the mental states of the brains in vats is similar to that of embodied persons? If narrow content is understood in Loar's sense as a way of taking the world to be a certain way, this would mean that although the brain in a vat is really thinking about trees-in-the-image (as analogously members of the linguistically deviant community are really thinking about elms when using the expression 'beech'), the brain in a vat conceives of the world in the same way as an embodied person (as analogously nonexpert members of the linguistically deviant community conceive

---

21. Putnam, 'The Meaning of "Meaning"', 234.

of the world in the same way as nonexpert members of the English-speaking community).

This would mean that the brain in a vat conceives of the world as having tree-like things in it, which would be to say that it is radically wrong in the way in which it conceives of the things it actually refers to.[22] So when the brain in a vat thinks 'There is a tree in front of me', it takes the world to be such that there is a spatial object, several feet tall, several feet in front of it. It is wrong about this (let us suppose that its vat has not been placed in front of a tree). However, its thought will be true, for it is true that there is a tree-in-the-image in-front-in-the-image of it. This would be a very extreme case of the ordinary phenomenon of misconception of one's referents.

Putnam would not regard this as a possible state of affairs. He claims that 'no matter what sort of inner phenomena we allow as possible *expressions* of thought... it is not the phenomena themselves that constitute understanding, but rather the ability of the thinker to *employ* these phenomena....'[23] Although the mental life of the brain in a vat is qualitatively identical with that of an embodied person, in Putnam's view the brain in a vat does not conceive of the world as being in the same way as the embodied person. Loar would agree with Putnam on this. As his narrow contents are conceptual roles, the realization conditions he says they determine depend upon what concepts the thinker possesses. As concepts are

---

22. Misconception of one's referents is certainly routine and more radical cases seem possible. In 'The Meaning of "Meaning"' Putnam suggests that a 'stereotype' should be regarded as part of what has to be mastered in order to grasp the meaning of an expression. He considers somebody who points to a snowball and asks 'Is that a tiger?' Putnam's point is that such a person would not be considered to grasp the meaning of 'tiger'. Even if Putnam is right about this, the answer to the question is obviously 'No', suggesting that the speaker has managed to refer to tigers, even though having a very strange idea of what they might be.

23. Putnam, Reason Truth and History, 19-20.

determined externally, the brain in a vat conceives of its world as populated with trees-in-the-image. It is unable to get any take on the world as being other than in-the-image.

If Strawson wishes to disagree with this and maintain that the special N/C intentionality he ascribes to a being solely on the basis of its conscious experience, even if it lacks the necessary causal connections for any externally grounded concepts at all, is genuinely intentional in that it grounds something recognizably intentional like Loar's realization conditions, he has his work cut out. Strawson has insisted that all that experience adds to behavioral and functional facts about a being is itself, and this is supposed to make the difference between genuine intentionality and the merely 'as-if' intentionality that is sometimes ascribed to artifacts.

Strawson says that part of the difficulty in accepting this arises from the supposition that all experiential qualitative character must be something like sensory qualitative character. If one does make this assumption, then the idea appears ridiculous. Strawson urges us to realize that experience is 'as much cognitive as sensory'[24]. It seems that he means by this that in addition to visual experience, auditory experience and all the rest of the varieties of sensory experience, there is something else he calls 'understanding experience'. This is the claim that there is something it is like to understand a sentence, and this is not a matter of association of sounds of words with images or sensations of any sort. It is suggested, for example, that if one hears a sentence in a language one understands, one has a very different experience from a hearer who does not understand the language. The former person automatically interprets the words as signs whilst the latter hears them as mere noise. The suggestion is that this results

---

24. Strawson, 194.

in a difference in experience that is distinct from any sensations the two individuals might be enjoying.

Let us suppose that Strawson is right about this and that experience is richer than what is provided by sensation. How does this help determine content? Some reason needs to be given for believing understanding experiences to be better off than sensory experiences in reaching out to the world. Strawson would say that they are not, so far as determining E/C content goes, but he would claim that they secure (non-externally grounded) N/C content. If a state with only N/C and not E/C content has no external grounding the way that Récanati suggests, and Strawson accepts in the ordinary case of N/C content, there is no explanation of how it can constitute a way of taking the world to be. If it is not supposed to be a way of conceiving of the world, and plays no part in determining E/C content, it is not clear in what sense this property of mental states is a form of intentionality.

The theoretical utility of understanding experience, or experience in general, may be no more than that it provides a difference for internalists to make the distinctions their intuitions support. One who is intuitively uneasy about ascribing genuine intentionality to experienceless machines may simply point to the difference in experience in denying it of them. For one may intuitively say of the blue cube seeking robot described above that it does not *really* have blue cubes in mind, whereas we do have blue cubes in mind, for we have experiences of blue cubes. The world could be subtly or grossly different to the way it is, and it would make no difference to the machine: it would not misconceive of its environment. It may be thought that we, on the other hand, have experiences to provide us with misconceptions, but this would be a mistake. Without an explanation of how our experience may provide conceptual content it is unable to play this theoretical role.

If this is the role that experience is supposed to play in the internalist account, then it amounts to a confession that so far as taking the world to be a

certain way goes, experience has no role to play, but that we tend to find it very difficult not to think of our experience as playing this privileged role, so we deny genuine intentionality to beings who lack such experiences (and grant at least N/C intentionality to those beings who do possess them). Having a particular experience won't make your thoughts really about blue cubes unless that experience *qua* experience constitutes taking the world to contain blue cubes. There are good reasons to be suspicious of this claim and no reason has been provided to think otherwise, unless some very different account of experience is offered[25].

The Form of a Theory of Intentionality

The arguments of the externalists that it is the environment that plays an essential role in determining the full content of an intentional state are convincing. Were some way of accounting for the internalist intuitions to be found, it is very probable that any adequate internalist view would be some sort of hybrid that recognized the externalist point. In the absence of an internalist account, the obvious way to proceed is to attempt to construct an externalist theory of intentionality. There are great obstacles in the way of such a theory, however. The basic problem that such theories face is indeterminacy. The classic argument to this effect is provided by Quine in 'Ontological Relativity', where he argues that naturalistic facts fail to determine reference, or that reference is 'inscrutable'. I believe that Quine's argument is more resilient than is often thought, and some of the most powerful objections to it ultimately fail. The argument is sufficiently general, however, that one may hope that some naturalistic theory of reference may be found that when worked out in detail will provide a refutation of it.

---

25. Cf. Churchland on experience in Scientific Realism and the Plasticity of Mind (Cambridge: Cambridge University Press, 1979), chap. 2.

Causal theories of linguistic reference recognize that the mere fact of causation itself is not enough to confer a determinate content on a linguistic term. The speakers' conceptions or intentions are generally appealed to as a determining mechanism. Given that mental states are no better off than linguistic terms in this respect, the challenge for a theory of intentionality is to find something else to play this role.

CHAPTER 2

THE INSCRUTABILITY THESIS

The thesis of the inscrutability of reference is articulated by Quine in his essay 'Ontological Relativity'. Quine argues that to adopt a naturalistic perspective on language is to recognize that there are no semantic properties beyond those implicit in speaker's dispositions to behavior:

> In psychology one may or may not be a behaviorist, but in linguistics one has no choice. Each of us learns his language by observing other people's verbal behavior and having his own faltering verbal behavior observed and reinforced or corrected by others... There is nothing in linguistic meaning, then, beyond what is to be gleaned from overt behavior in overt circumstances.[1]

The argument for the inscrutability thesis is that the speech dispositions of speakers are compatible with multiple inconsistent assignments of referents to their linguistic terms. Given that there are no further facts that could determine reference, reference is indeterminate, or inscrutable as Quine puts it.

The argument for the inscrutability thesis resembles that originally advanced by Quine for the indeterminacy of translation in chapter two of Word and Object. One of the difficulties of stating the thesis of indeterminacy of translation is making clear in what sense the rival translation manuals for the language would be inconsistent. This can only be the claim that they ascribe different meanings to the terms in the language, which means that in order to evaluate the indeterminacy thesis one first has to settle on some reliable notion of what it is for two sentences to have the same meaning. The thesis of the inscrutability of reference does not

---

1. W. V. Quine, Word and Object (Cambridge, MA: MIT Press, 1960), 5.

suffer from this difficulty, owing to the relatively clearer status of reference compared to that of meaning: the referent of a term is that which it is true of.

One may see the truth condition of a sentence as being determined by the referential properties of its parts. A subject-predicate sentence of the form *a* is *F* is true if and only if the referent of *a* is in the extension of *F*. The truth conditions of complex sentences are determined by the truth conditions of their parts and the truth-functions represented by the logical connectives. Quine's argument for the inscrutability thesis rests upon the idea that a single sentence may have two different assignments of referents to its parts such that the resulting truth conditions are always jointly satisfied or jointly unsatisfied. The sentence will always have the same truth-value on each interpretation. This means that whenever a speaker is disposed to assent to or dissent from the one sentence, he will be disposed to assent to or dissent from the other. The difference in reference will make no difference to the dispositions to speech behavior. The simplest example Quine gives of this is the interpretation of the French expression 'ne... rien'. One may either interpret 'rien' as 'nothing' and construe 'ne' as pleonastic, or interpret 'rien' as 'anything' and 'ne' as 'not'. Either way, sentences containing this pair of expressions will be interpreted as having the same truth condition. Quine notes that this example may strike one as disappointing, for it appears that he has simply chosen unrealistically small units. Perhaps, one may suggest, 'ne... rien' is to be understood as a complete expression and the individual components do not bear any independent significance. Quine's response to this objection is to apply the same argument to units which are commonly acknowledged to have independent significance, such as predicates like 'rabbit'.

Quine's famous 'gavagai' example from Word and Object here returns as an example of the inscrutability of reference. The example may be presented as concerning a pair of linguists in the field, Ling1 and Ling2, attempting to translate

an unknown language. The natives tend to utter the word 'gavagai' around rabbits and in Ling1's interpretation, Int1, 'gavagai' is interpreted as referring to rabbits. Ling2 advances a different interpretation, Int2. Ling2 suggests that 'gavagai' refers not to rabbits but to undetached parts of rabbits. Quine's claim is that despite this divergence in interpretation, sentences containing 'gavagai' may be assigned the same truth-value in each interpretation by suitable adjustments in the apparatus of individuation. The claim is not merely that there would be no way to tell which interpretation was correct. It is assumed that Ling1 and Ling2 have access to all the data that is relevant to interpreting the language but may still produce incompatible interpretations. Quine infers from this that there is no fact of the matter as to which interpretation is correct. What plausibility this argument has depends upon it being the case that there is a rabbit present if and only if there is an undetached rabbit part present. This fact makes it conceivable that there are possible adjustments that may be made to ensure that each interpretation would assign the same truth value to the sentences in which 'gavagai'occurs.

For example, one might try to decide between interpreting 'gavagai' as dividing its reference over rabbits and over undetached rabbit parts by asking a speaker of the language 'Is this gavagai the same as that one?' while pointing to appropriate parts of the rabbit. Apart from the problem of the indeterminacy of ostension, an additional problem is that this objection assumes that some particular native expression is to be translated as 'is the same as.'[2] However, Quine observes, when we take ourselves to be asking in the native language 'Is this Gavagai the same as that one?' we may instead be asking 'Does this Gavagai belong with that one?' where the native expression that may be translated as

---

2. The problem of the indeterminacy of ostension is that when the interpreter points to a part of the rabbit, the gesture itself does not discriminate between indicating the rabbit part and the whole rabbit of which it is a part.

'belongs with' expresses an equivalence relation which would ordinarily be translated by us as identity.

The possibility of interpreting a certain native expression which appears to play the role of identity as a different equivalence relation also provides a solution to the problem of how to interpret numerical expressions in the native language, given the interpretation of 'gavagai' as referring to undetached rabbit parts. The problem is that when one asks the native how many gavagai are present, the answer given will correspond to how many rabbits are present. When there is only one rabbit present, the native will respond to the query by uttering an expression the linguist interprets as expressing the number one, and so on. The solution is to use a suitable alternative equivalence relation in the place of identity when interpreting numerical statements. For example, the statement 'There is exactly one rabbit' may be expressed quantificationally as follows (where 'R' is satisfied by rabbits):

$\exists x\,(Rx \wedge \forall y\,(Ry \supset y{=}x))$

By replacing the identity predicate in the above formula with an expression for 'belongs with', the alternative equivalence relation suggested by Quine, Ling2 may now offer an empirically adequate interpretation of the natives' numerical statements in the following way (where 'URP' is satisfied by undetached rabbit parts and '$=_r$'stands for the belongs-with relation):

$\exists x\,(URPx \wedge \forall y\,(URPy \supset y{=_r}x))$

This formula asserts that there is at least one undetached rabbit part that belongs with any other undetached rabbit part. This effectively amounts to the claim that all the undetached rabbit parts are part of the same rabbit, although rabbits are not quantified over.

For the belongs-with relation to function in this way, any two rabbit parts must belong together if and only if they are undetached parts of the same rabbit.

The relation needs to be restricted to the present time. If one were to amputate a leg from each of two rabbits and then reattach each severed leg to the rabbit other from which it came, one would have a situation where there were rabbit parts that belonged with parts from two different rabbits. In this case, the belongs-with relation would be non-transitive, making it useless as Ling2's alternative to identity.

Restriction to the present time will not help with another problem case for the relation, that of Siamese rabbits. Presumably, when Siamese rabbits are present there are two rabbits present. The obvious problem for the belongs-with relation is that the parts of Siamese rabbits are not all undetached parts of the same rabbit, but they are undetached parts of the same thing. However, this thing is not a single rabbit, but a pair of rabbits. So long as each part of the pair of undetached rabbits is an undetached part of one rabbit in the pair and not the other, it is not a problem that they are all undetached parts of the same object.[3]

There is a continuum of cases from an ordinary single rabbit to full Siamese rabbits, which are two entire rabbits joined together at a single point. In between these cases are such things as two-headed rabbits. Is a two-headed rabbit a single rabbit, or a pair of Siamese rabbits? If it is a pair of Siamese rabbits, to which head does the rest of the body belong? There will not always be obvious answers to such questions. That is to say that it will not always be clear when one undetached rabbit part belongs with another undetached rabbit part. However, this no more shows that there is something wrong with the belongs-with relation than it shows that there is something wrong with the predicate 'rabbit'. This is because it is not always clear whether something is a rabbit.

---

3. This phenomenon does show that this trick for numbers cannot be pulled for objects of a type which can share parts, if there are any.

If we define the belongs-with relation in the present context by saying that two things belong together if and only if they are presently undetached parts of the same rabbit, then there will be uncertainty as to what rabbit parts belong with what other parts only when there is uncertainty as to how many rabbits one is dealing with. In the case of the two-headed rabbit, for example, an English speaker may be unsure whether there are two rabbits present or only one. Every undetached part of the two-headed rabbit is an undetached rabbit part, but given the uncertainty as to how many rabbits are present, it is also uncertain whether *all* these undetached rabbit parts belong together or not. That is, it is uncertain which of the following sentences correctly describes the situation:

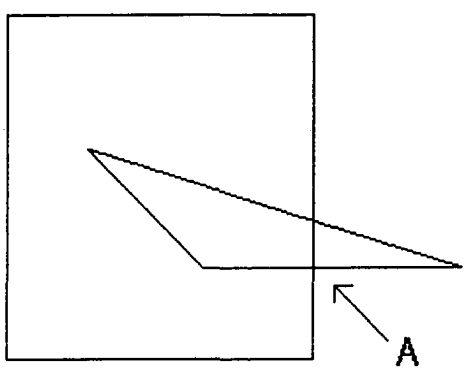
$\exists x\,(\mathrm{URP}x \wedge \forall y\,(\mathrm{URP}y \supset y=_{\mathrm{r}}x))$

$\exists x \exists y\,((\mathrm{URP}x \wedge \mathrm{URP}y) \wedge (\sim x=_{\mathrm{r}}y \wedge \forall z\,(\mathrm{URP}z \supset (z=_{\mathrm{r}}x \vee z=_{\mathrm{r}}y))))$

There need be no concern that as the belongs-with relation is defined in terms of the English predicate 'rabbit', a language that contains a term for this relation must also contain a predicate satisfied by rabbits. In Int2, the predicate satisfied by undetached rabbit parts is understood by us in terms of it being satisfied by something which is a part of a rabbit. The assumption is, I believe, that those who speak this language do not themselves understand this predicate in this way, for it is being assumed that their language does not contain a predicate satisfied by rabbits. Instead, for speakers of this language, the predicate satisfied by undetached rabbit parts is not defined in terms of other predicates in the language. Similarly, I assume that for speakers of the language as interpreted by Int2, the belongs-with relation is primitive. Defining the relation in terms of 'rabbit' assures us, who do not speak this language, that the relation exists and is no less well defined than the predicate 'rabbit'.

Fodor's Objection to the Inscrutability Thesis

In The Elm and the Expert Fodor advances a seemingly powerful argument against the inscrutability thesis. Fodor observes that although there are rabbits present if and only if there are undetached rabbit parts present, there may be some property which is compatible with being an undetached rabbit part but not compatible with being a rabbit. In his discussion of Quine's 'gavagai' example, Fodor assumes that the linguists are in possession of the semantics of the sentential connectives of the language being interpreted, noting that the claim of the inscrutability thesis is that the semantics of subsentential expressions is indeterminate, even assuming that the semantics of sentential expressions is fixed. The task of each linguist is to show that each sentence that a native informant, Inf, judges to be true comes out true on his interpretation. In order to avoid confusion over merely epistemic issues, it is assumed that Inf always correctly judges the truth-value of sentences in his language.



Fodor asks us to consider the situation depicted left. Ling1 suggests that Inf's utterances of 'square' should be interpreted as referring to squares. Ling2 suggests that Inf's utterances of 'square' should instead be interpreted as referring to undetached proper parts of squares. Whenever Ling1 points to a square and asks 'Square?' Inf will reply 'Yes' and whenever Ling1 points to something that is not a square and asks 'Square?' Inf will reply 'No'. Ling1's hypothesis that Inf is referring to squares by his term 'square' is not better confirmed by this than Ling2's hypothesis, however, for whenever Ling2 points to an undetached square part and asks 'Square?' Inf will again reply 'Yes' and will reply 'No' whenever Ling2 makes the same query while

pointing to something that is not an undetached square part. Inf is guaranteed to do this because there are squares present if and only if there are undetached proper parts of squares present.

Fodor's idea is that one can find a property such that a part of a square may instantiate it, but a square cannot. He suggests as a candidate the property of being a triangle. It is quite possible for something to be a part of a square and also be a triangle, but it is not possible for something to be both a square and a triangle. Fodor's expectation is that this fact will be reflected in speech dispositions, and so it may be assumed that no competent speaker of English will ever sincerely say 'There is something that is both a square and a triangle'. The problem for Quine is that there are circumstances under which a competent speaker of English would utter 'There is something that is both an undetached square part and a triangle'. When such circumstances obtain, Ling2 predicts that Inf will accept the sentence 'There is something that is both a square and a triangle' whereas Ling1 predicts that Inf will not accept this. So Ling2's interpretation does not leave all dispositions to speech behavior unchanged, and so reference is not as inscrutable as Quine has claimed.

In order for this argument to work, a situation must be exhibited in which Ling2 (but not Ling1) predicts that Inf will accept a sentence which asserts of one and the same thing that it is both a square and a triangle. Given that the point A is both part of a triangle and a part of a square, Ling2 will predict that Inf accepts 'A is a square' and 'A is a triangle'.[4] This fact does not yet show that Inf accepts a sentence which asserts that one and the same thing is both a square and a triangle,

---

4. The point A in the diagram is not a triangle, so Inf would not accept the sentence 'A is a triangle'. Fodor must be assuming that Ling2 takes Inf's term 'triangle' to refer to undetached triangle parts. The argument could be presented with equal success without this assumption, by making 'A' refer to a triangular region contained by a square.

for it has not yet been shown that the name 'A' unambiguously names the point A. Without this assumption, it cannot be assumed that Ling2's prediction is contrary to fact. The basis of the argument is that Ling2 will predict that Inf will accept a sentence that in fact Inf will not accept, owing to its contradictory nature. It is no contradiction, however, to assert both 'A is a square' and 'A is a triangle' when each occurrence of 'A' in the asserted sentences names a different object.

Having concluded that Ling2's alternative construal of Inf's ontology can be ruled out if it can be determined when two expressions of Inf's language name the same thing, Fodor goes on to argue that in certain circumstances this can be determined. In particular, this can be done if some structure in Inf's language can be determinately translated as predicate conjunction. As Fodor observes, the semantic function of the structure '*x* is *F* and *G*' is to apply both predicates to the same individual. It is not possible to determine that a particular structure in a language is predicate conjunction simply from observing which sentences are held true in that language, but Fodor persuasively argues that this can be determined if the inferences that speakers of the language are willing to draw from the sentences they hold true are also taken into account. Fodor suggests the following criterion:

"A predicate connective '*' is predicate conjunction if(f?):

Inf always takes sentences of the form 'A is F * G' to imply the corresponding sentence conjunction 'A is F and A is G';

and

whenever Inf is prepared to accept 'A is F * G', he is prepared to infer 'A is F * G' from 'A is F and A is G'."[5]

On this criterion, should Inf, when confronted with Fodor's diagram, accept the inference from 'A is a square' and 'A is a triangle' to 'A is a square and a

---

5. Jerry Fodor, The Elm and The Expert (Cambridge, MA: MIT Press, 1994), 71.

triangle' then Ling2, who predicts that Inf will accept 'A is a square' and 'A is a triangle' will be forced to conclude that Inf will accept 'A is a square and a triangle'. Whatever Inf refers to by 'A', this sentence will be false.

It is apparent that Fodor's argument may be extended beyond the geometrical example that he considers. Fodor's speculation is that whenever Quine or his followers suggest that an English expression *F* could be taken to have as its extension not the class of *F*s but rather some distinct class, the *G*s, it will be possible to find some property *H* such that being *H* is compatible with being *G* but not with being *F*. This clearly works for undetached square parts and squares, because a point (or a triangular region) can be an undetached part of a square, but cannot be a square. It is also works for undetached rabbit parts and rabbits, as nothing is both a rabbit's foot and a rabbit, but there are things that are both an undetached rabbit part and a rabbit's foot. So no competent English speaker will say 'There is something that is both a rabbit and a rabbit's foot' but there are circumstances under which someone who speaks the language interpreted by Int2 will assent to that sentence. So Fodor's argument, if successful, would show in these cases as well that overall dispositions to linguistic behavior are not preserved in the deviant interpretation.

The substance of Fodor's objection to the inscrutability thesis is that on the deviant interpretation we should expect the speaker of the language to assent to sentences that a speaker of English never would, such as 'There is something that is both a rabbit and a rabbit's foot'. This conclusion can be avoided if one adjusts suitable parts of the apparatus of individuation, as Quine suggests. In particular, an adjustment to the interpretation of Inf's locution 'there is' so that it no longer expresses the existential quantifier has this consequence. Fodor assumes that Inf's utterances of 'There is a rabbit' are interpreted by Ling2 as the statement that there is an undetached rabbit part. If one interprets Inf's utterance

(1) There is a rabbit

to effectively make the statement that there is something entirely composed of things that belong with a particular undetached rabbit part, and Inf's utterance

(2) There is a rabbit's foot

to effectively make the statement that there is something entirely composed of things that belong with a particular undetached rabbit's foot part, it does not follow that Inf would accept the sentence

(3) There is something that is both a rabbit and a rabbit's foot

Expressed quantificationally, Inf's utterance of (1) is to be interpreted as follows:

(1') $\exists x$ (URPx $\wedge$ $\exists y \forall z$ (z=$_{r}x$ $\leftrightarrow$ Pzy))

Let Inf's utterance of (2) be interpreted as follows (the predicate 'URFP' is satisfied by undetached parts of rabbits' feet, '$=_{rf}$' expresses the is-an-undetached-part-of-the-same-rabbit's-foot-as relation and the two-place predicate 'P' expresses the is-an-undetached-part-of relation):

(2') $\exists x$ (URFPx $\wedge$ $\exists y \forall z$ (z=$_{rf}x$ $\leftrightarrow$ Pzy))

In the presence of a rabbit, Inf would accept both (1) and (2), interpreted as (1') and (2'), but will not accept (3), for this is to be interpreted as follows:

(3') $\exists x$ ((URPx $\wedge$ URFPx) $\wedge$ $\exists y \forall z$ ((z=$_{r}x$ $\leftrightarrow$ Pzy) $\wedge$ (z=$_{rf}x$ $\leftrightarrow$ Pzy)))[6]

An utterance of (3) interpreted in this way can roughly be understood as making the statement that there is something which is both entirely composed of things which belong with a particular undetached rabbit part and entirely composed of things which belong with a particular undetached rabbit's foot part. That is to say, there is an entity that is both entirely composed of undetached rabbit parts and

---

6. There are other ways to handle predicate conjunction that are less advantageous for the line of argument that I am to suggest. However, I take it that it is open to Ling2 to stipulate how predicate conjunction works in Int2.

entirely composed of undetached rabbit's foot parts. There is no such entity, of course, so given the assumption that Inf will only accept true sentences, Inf will not accept (3) interpreted as (3'). Similar reasoning will show that Inf will not accept the sentence 'There is something that is both a square and a triangle' when confronted with Fodor's diagram because there is nothing that is both entirely composed of undetached square parts and entirely composed of undetached triangle parts.

The suggested interpretation has the disadvantage of making the expression 'There is' ambiguous. This can be accounted for in a systematic way, however. When one asserts the existence of something by uttering the sentence 'There is something that is *F*' the usual assumption is that the 'There is' locution is to be understood as supplying the existential quantifier:

$\exists x\, \phi x$

Instead, one might suppose that 'There is' is doing something rather more complicated, as shown by the following schema:

(F) $\exists x\, (\phi x \wedge \exists y \forall z\, (z{=}_{\phi\uparrow} x \leftrightarrow \mathrm{P}zy))$

The expression '$=_{\phi\uparrow}$' in schema (F) stands for the equivalence relation that obtains between all the parts of an object whose parts satisfy ϕ. This expression stands for the is-a-part-of-the-same-thing-of-the-type-that-ϕs-are-parts-of-as relation. Naturally, this construction is designed to be coupled only with predicates that are satisfied by the parts of certain types of object.[7] On Int2, each predicate in the language is associated with a certain 'belongs-with' relation. Schema (F) is

---

7. There is no issue here (beyond that raised generally by the inscrutability thesis) about what the correct interpretation of our quantifier is. For the purposes of the present discussion, no special assumptions are made about the interpretation of our quantifier. The issue is the interpretation of a fragment of Inf's language. Int1 interprets Inf's 'There is' as our quantifier, my suggestion is that on Int2, this part of Inf's language should be interpreted as a more complex construction, which itself contains our quantifier.

designed to systematically show which equivalence relation is associated with each predicate.

Although Fodor's argument is not successful, it does show something troubling for the defender of the inscrutability thesis. Quine responds to the objection that one could point to different parts of the same rabbit while asking 'Is this the same *gavagai* as that?' by answering that 'same' could be translated as 'belongs with' so the question would then be translated as 'Does this *gavagai* belong with that?' Fodor's argument shows that there cannot simply be one belongs-with relation for all objects. This is because one object can be part of something else. A rabbit's foot can be part of a rabbit's leg, for example. All of the parts of a rabbit's leg belong together, but they also belong with all the other parts of the rabbit. If one attempted to use a single belongs-with relation, (1')-(3') above would be replaced with the following (where $=_b$ stands for the belongs-with relation):

(1'') $\exists x$ (URP$x \wedge \exists y \forall z$ ($z =_b x \leftrightarrow$ Pzy))

(2'') $\exists x$ (URFP$x \wedge \exists y \forall z$ ($z =_b x \leftrightarrow$ Pzy))

(3'') $\exists x$ ((URP$x \wedge$ URFPx) $\wedge \exists y \forall z$ (($z =_b x \leftrightarrow$P zy) $\wedge$ ($z =_b x \leftrightarrow$P zy)))

(3'') is true and so Fodor's argument would go through. The problem is that (2'') asserts the existence of an entity entirely composed of things which belong with the undetached rabbit's foot part, but this collection includes all of the undetached parts of the rabbit and not just those undetached parts of the rabbit which are also undetached parts of the rabbit's foot. So commitment to this way of overcoming Fodor's objection requires abandoning a universal belongs-with relation.

The universal belongs-with relation would have to be given up anyway. When the linguist points first to a rabbit's hind leg and then to a front leg of the same rabbit while asking 'Is this the same *gavagai* as that?', the native speaker is

supposed to give his assent, so the hind leg and front leg of the same rabbit belong together. However, when the linguist points first to the hind leg and then to the front leg of the same rabbit while asking 'Is this the same *gavagai*'s leg as that?', the native speaker would again give his assent, for the belongs-with relation holds between the hind leg and front leg of the rabbit. The problem is that for some identity statements one wants all the parts of an object to belong together, and for others one does not, for one does not want it to turn out that all the parts of an object are identical to each other.

Evans' Objections to the Inscrutability Thesis

Fodor claims to have been inspired in part by Evans' paper 'Identity and Predication'. In this paper Evans does not merely present a counterexample to the inscrutability thesis, but does so on the basis of a characteristically insightful philosophical point. Nevertheless, the same interpretation of 'There is' that provides a defense of the inscrutability thesis against Fodor's argument also serves as an effective defense against those of Evans.

Evans asks us to consider a language which contains two classes of terms, *F* terms and *G* terms. The *G* terms are stimulus-synonymous with one-word sentences such as 'A rabbit' and the *F* terms are stimulus-synonymous with one-word sentences such as 'White' or 'Furry'. The *F* terms and the *G* terms may be coupled together, or with a negation operator. For the speakers of this language to assent to the conjoined predicates (*F G*) it is not sufficient for *F* and *G* to both be assented to, nor that there is some region of space in which both *F*ness and *G*ness are instantiated. For example, it is not sufficient for speakers to assent to 'White rabbit' that they assent to 'White' and 'Rabbit' separately, nor that there is some part of a rabbit that is white. What is sufficient is that the *F*ness be instantiated in some particular way in an object whose presence prompts assent to the queried *G*

terms. Evans puts this by saying that competence in this language requires mastering the identity conditions of the relevant objects:

> To say that an expression has a particular divided reference makes sense only in the context of the explanation of compound sentences. To decide that a term divides its reference over rabbits is to decide that the sentences in which it occurs involve predication of rabbits. And to decide that a set of sentences involves predication of rabbits is to identify the way those sentences' assent conditions are generated from their parts as depending on the identity conditions of rabbits, and so systematic mastery of those sentences requires mastery of the identity conditions of rabbits.[8]

These considerations present the objection to Quine's inscrutability thesis that sentences of English that may naively be taken as speaking of rabbits depend upon the identity conditions of rabbits and not upon the identity conditions of rabbit parts. Evans makes this argument in the following way.

If one supposes that 'rabbit' divides its reference over rabbit parts, then one must amend the satisfaction conditions of the expressions with which it may be coupled. This is because, for example, a white foot of an otherwise brown rabbit will elicit dissent from the query 'White rabbit?' even though a white part of a rabbit is present. This is to say that if the sentence 'White rabbit' has been interpreted to mean something equivalent to the English sentence 'White rabbit-part', the assent conditions of the sentence 'white rabbit' on the new interpretation would be different from its assent conditions on the old. It would provide no defense of the inscrutability thesis to suppose that 'white' is satisfied by things which are a part of something which is white, for the foot of a brown rabbit with a

---

8. Gareth Evans, 'Identity and Predication,' in Collected Papers (Oxford: Oxford University Press, 1985), 39.

white leg is a part of a white thing, but speakers of English will still dissent from the query 'White rabbit?' in the presence of such a creature. Evans suggests that the only way for the defender of the inscrutability thesis to overcome this problem is to take 'white' to be satisfied by things which are parts of a white rabbit. Of course, the problem with this is that it does not yield the correct assertability conditions for other sentences containing 'white'. An English speaker does not assent to the query 'White carpet?', for example, only in the presence of a carpet which is part of a white rabbit.

This objection to Quine is not so crushing as it may appear, as Hookway has suggested a solution to this problem for the defender of the inscrutability thesis.[9] Evans argues that in order to yield the correct interpretation of compound statements when 'rabbit' is interpreted as referring to rabbits, it is necessary to give the following satisfaction condition for the term 'white':

$\forall x$ (x satisfies 'white' if and only if x is a part of a white rabbit)

However, this satisfaction condition for 'white' bars the use of the term in conjunction with expressions other than 'rabbit'. Hookway suggests as a different satisfaction condition for 'white':

$\forall$x (x satisfies 'white' if, and only if, either

(i) 'white' occurs together with 'rabbit' and x is a part of a white animal

or

(ii) 'white' does not occur together with 'rabbit' and x is white)

The *ad hoc* nature of this maneuver makes this interpretation impractical but this is no objection to its tenability as an interpretation for Quine's purposes. Still, we may produce a more systematic way of overcoming Evans' objection to the

---

9. Christopher Hookway, Quine: Language, Experience and Reality (Cambridge: Polity Press, 1988), 154-155.

inscrutability thesis by utilizing the schema that served to undermine Fodor's argument (the predicate 'W' is satisfied by things that are white):

(4) $\exists x$ (URPx $\wedge$ $\exists y \forall z$ ((z=$_{r}x$ $\leftrightarrow$ Pzy) $\wedge$ Wy))

The value of the variable y in (4) is a rabbit. This means that there need be no amendment to the satisfaction conditions for predicates that may be combined with 'rabbit', such as 'white'.

Having presented the objection to the inscrutability thesis discussed above, Evans goes on to say 'However, let us waive this difficulty; for another more interesting difficulty emerges.'[10] This more interesting difficulty is that as a consequence of whatever adjustments one has to make in the satisfaction conditions of the terms which may couple with 'rabbit' in order to preserve dispositions to speech behavior, different parts of the same rabbit will be indistinguishable by the predicates of the language. Evans continues:

> [B]y Quine's absolute and objective criterion of what two-place predicate to count as the identity predicate, the predicate the theory attempts to treat as 'is a part of the same rabbit as' has the substitutivity property of, and therefore, for Quine, is, the identity predicate. It turns out to be no more in accordance with Quine's principles than it is with mine to discern predication of rabbit parts in this discourse.[11]

The 'absolute and objective' criterion of identity of which Evans speaks may be found in Quine's 'Reply to Professor Marcus'.[12] In that paper, Quine says that an open sentence 'ϕxy' is to be read as 'x=y', i.e. stating an identity, if the requirements of (i) strong reflexivity and (ii) substitutivity are met:

---

10. Evans, 43.

11. Ibid.

12. W. V. Quine, 'Reply to Professor Marcus,' in The Ways of Paradox (Cambridge, MA: Harvard University Press, 1966).

(i) $\forall x\, \phi xx$

(ii) $\forall x \forall y\, (\phi xy \wedge ...x... \supset ...y...)$

As Int2 requires interpreting some binary predicate of the language as the is-a-part-of-the-same-rabbit-as relation (i.e. the belongs-with relation for rabbits), Evans' criticism, if correct, is devastating. It is not correct, however, as on the interpretation utilizing schema (F), the is-a-part-of-the-same-rabbit-as relation will not meet the substitutivity requirement. Evans believes that on any adequate interpretation this relation will meet condition (ii), so that if any two things are part of the same rabbit, whatever predicates in the language are satisfied by the one part will be satisfied by the other. This will fail on my suggestion because the satisfaction conditions for the other predicates are not amended. In the case of the brown rabbit with a white foot, there will be a rabbit part (the white foot) that satisfies the predicate 'white' and a rabbit part that does not. Evans appears to assume that whatever compensatory adjustments have to be made to accommodate taking 'rabbit' to refer to undetached rabbit parts must come in the other predicates of the language. I dare say he assumes this because he has made the assumption that when one so interprets 'rabbit' one must interpret the sentence 'There is a rabbit' to state:

(5) $\exists x$ URPx

Given this assumption, when one comes to interpret 'There is a white rabbit' all there is left to adjust are the satisfaction conditions of the predicate 'white'. It seems probable, particularly given Quine's remarks about the apparatus of individuation, that one would be more successful in looking to the quantifier to make these adjustments.

Consideration of Evans' argument does reveal, however, that schema (F) is unsatisfactory for the purposes of Int2. This is because it does not account, for

example, for Inf's disposition to assent to the following sentence in the presence of, say, a single rabbit which has a white foot but is otherwise brown.[13]

(RP) 'There is a rabbit part that is white and a rabbit part that is non-white'

If one attempts to interpret Inf's predicate 'rabbit part' as 'undetached part of a rabbit part' (in order to conjoin it with Inf's existential quantifier, interpreted according to schema (F)), one faces the problem that all the parts of a rabbit part are themselves rabbit parts, and any two parts of a rabbit may be parts of the same rabbit part. This means that not only do all the parts of rabbit part satisfy the 'belongs with' relation for rabbit parts, but so do all the rest of the parts of the rabbit. This is to interpret (RP) as the following ('UPRP' is satisfied by undetached parts of rabbit parts and '$=_{rp}$' stands for the is-a-part-of-the-same-rabbit-part-as relation):

(RP') $\exists x$ (UPRP$x \wedge \exists y \forall z$ (($z=_{rp}x \leftrightarrow Pzy$) $\wedge$ Wy)) $\wedge \exists x$ (UPRP$x \wedge \exists y \forall z$ (($z=_{rp}x \leftrightarrow Pzy$) $\wedge \sim$Wy))

The value of each occurrence of the variable y in (RP') is a rabbit. Moreover, it is the same rabbit in each occurrence. This means that Inf will not assent to (RP'), as this sentence asserts that the same rabbit is both white and non-white, although Inf will assent to (RP).

In order to remedy this deficiency in Int2, instead of a predicate F of Inf's language being associated with a single equivalence relation, F is associated with a class of equivalence relations. For each object whose parts satisfy F, there is an equivalence relation that obtains between those parts and those parts only. The class to be associated with F is the class of all such relations, R_F. Instead of schema (F), Int2 interprets Inf's quantifier with the following schema:

($\exists$) $\exists x$ ($\phi x \wedge \exists y \exists R \forall z$ ($R \in R_{\phi\uparrow} \wedge (Rzx \leftrightarrow Pzy)$))

---

13. This was pointed out to me by Michael Levin.

The expression '$R_{\phi\uparrow}$' in schema ($\exists$) names the class of equivalence relations that obtain amongst the parts of each object whose parts satisfy ϕ. Using this schema, Inf's utterance of (1) is interpreted as the following in Int2:

(1''') $\exists x$ (URPx $\wedge$ $\exists y \exists R \forall z$ ($R \in \mathrm{R_r} \wedge (Rzx \leftrightarrow \mathrm{P}zy)$))

Although $\mathrm{R_r}$ contains each equivalence relation that obtains between the parts of each rabbit, as each undetached rabbit part is a part of only one rabbit, there is only a single relation that can be the value of the variable *R*. This ensures that the value of the variable *y* is the rabbit that *x* is an undetached part of.

Using schema ($\exists$), Inf's utterance of 'There is a white rabbit' is interpreted in Int2 as the following:

(4') $\exists x$ (URPx $\wedge$ $\exists y \exists R \forall z$ ($R \in \mathrm{R_r} \wedge (Rzx \leftrightarrow \mathrm{P}zy)$) $\wedge$ Wy)

Inf's utterance of (RP) is interpreted in Int2 as follows:

(RP'') $\exists x$ (UPRPx $\wedge$ $\exists y \exists R \forall z$ ($R \in \mathrm{R_{rp}} \wedge (Rzx \leftrightarrow \mathrm{P}zy)$) $\wedge$ Wy) $\wedge$ $\exists x$ (UPRPx $\wedge$ $\exists y \exists R \forall z$ ($R \in \mathrm{R_{rp}} \wedge (Rzx \leftrightarrow \mathrm{P}zy)$) $\wedge$ ~Wy)

In (RP'') the value of the variable *y* is a rabbit part. $\mathrm{R_{rp}}$ contains each equivalence relation that obtains between the parts of each rabbit part. There are many such relations that can serve as the value of the variable *R*, as each undetached part of a rabbit part may be a part of many different rabbit parts. A rabbit's foot, for example, is both part of a rabbit's leg, and part of the lower half of a rabbit.

As the value of the first occurrence of the variable *x* in (RP''), a particular undetached part of a rabbit part, is amongst the relata of many of the relations in $\mathrm{R_{rp}}$, when there is just a brown rabbit with a white foot present, there will be one relation in $\mathrm{R_{rp}}$ such that the whole that is entirely composed of the things that bear this relation to each other (i.e. a particular part of the rabbit) is white. The value of the second occurrence of the variable *x* in (RP'') is also amongst the relata of many of the relations in $\mathrm{R_{rp}}$. There will be one such relation that the whole that is entirely composed of the things that bear this relation to each other is non-white.

(RP'') makes assertions of rabbit parts, but does not specify precisely which rabbit parts are in question. Should a speaker of the language interpreted by Int2 wish to refer to some more specific rabbit part, some other predicate may be used to do so. For example, it may be asserted that there is a white rabbit's foot and a non-white rabbit's foot as follows:

$\exists x$ (URFP$x \wedge \exists y \exists R \forall z$ ($R \in \mathrm{R_{rf}} \wedge (Rzx \leftrightarrow \mathrm{P}zy)) \wedge \mathrm{W}y) \wedge \exists x$ (URFP$x \wedge \exists y \exists R \forall z$ ($R \in \mathrm{R_{rf}} \wedge (Rzx \leftrightarrow \mathrm{P}zy)) \wedge \sim \mathrm{W}y$)

Were it really the case that on Int2 different parts of the same rabbit were indistinguishable by the predicates of the language, the interpretation would violate Evans' principle that a necessary condition of determining that a sentence involves predication of objects of type *A* is that mastery of the sentence requires mastery of the identity conditions of *A*s. As Evans explains:

> We have associated with 'rabbit' criteria of identity that dispose it to behave in such a way that a sentence of the form 'A rabbit (here) is ϕ and a rabbit (here) is $\sim\phi$' may be true given a certain distribution of evidence in a scene in which just one rabbit is present; yet this disposition can never be exercised, since the language does not contain a suitable candidate for ϕ.[14]

Evans' point here is that Int2 attributes semantic resources to the term 'rabbit' which it is impossible to use, given the interpretation of the rest of the predicates of the language. The interpretation of 'rabbit' as referring to rabbit parts would only be useful were one able to assert a sentence such as the following in the presence of a single rabbit:

(6) There is a rabbit that is white and a rabbit that is non-white.

As Evans believes that Int2 makes this impossible, it is not the case that mastery of sentences containing 'rabbit' interpreted by Int2 requires mastery of the identity

---

14. Evans, 43-44.

conditions of rabbit parts, for the language does not even permit one to say that any rabbit part is in any way different to any other.

Given the behavioral facts, any adequate interpretation must not imply that speakers of the language will assent to (6) when uttered in the presence of a single rabbit. Evans imagines that this will be effected by the predicate 'white' not distinguishing between parts of the same rabbit, so that if one part of a rabbit satisfies 'white', all other parts of the same rabbit do. On my proposal, speakers will dissent from (6) when a single rabbit is present because of what 'There is a rabbit' means. Given the interpretation of 'there is' by schema (∃), it is no longer the case that the interpretation of 'rabbit' as referring to rabbit parts should result in sentences of the form 'A rabbit (here) is ϕ and a rabbit (here) is $\sim\phi$' coming out true when there is only one rabbit present.

It would certainly be superfluous to introduce a predicate which, when combined with the quantificational apparatus of the language, allows one to talk about rabbit parts when the rest of the language prevents one from using it in this way. However, this is not what the predicate is used to do when an utterance of the sentence 'There is a white rabbit' is interpreted as (4'). When coupled with the quantificational apparatus in the language, this 'rabbit' predicate is used to talk about rabbits, but this is achieved without there being a term in the language that divides its reference over rabbits.

The Relevance of the Inscrutability Thesis

Quine's argument presents a challenge as it concludes that ambiguities exist in the most unlikely cases. Fodor and Evans have suggested a disambiguation. I suggest a reambiguation. The consequences of my suggestion are more limited than those sometimes drawn from Quine's argument, however. Much of the interest in the

inscrutability thesis stems from its ontological implications. These implications are absent from my defense of the thesis.

It should be noted that the value of the variable *y* in (1') is a rabbit. That Int2 interprets native sentences as quantifying over rabbits is a marked departure from Quine's presentation of the inscrutability thesis and its attendant doctrine of 'ontological relativity'. In Reason, Truth and History, Putnam appeals to the inscrutability thesis as showing the failure of metaphysical realism, the view that there is a single correct description of reality. The inscrutability thesis that I have defended here does not lend support to Putnam's 'internal realism', the view that there are 'equally coherent but incompatible conceptual schemes'.[15] This is because it leaves the ontological commitments of the language unchanged.

If one accepts my proposal, it may appear that the inscrutability thesis faces a similar problem to that faced by the thesis of the indeterminacy of translation. When there is a difference in ontological commitments ascribed to a language by different interpretations, this assures us that these interpretations are genuinely different. With no difference in ontological commitments whether these alternative interpretations show reference is inscrutable at all may be open to question.

The adequacy of (1') as an interpretation of 'There is a rabbit' remains interesting for those concerned with the determination of reference for subsentential expressions, as it only contains a predicate that divides its reference over rabbit parts, rather than rabbits. The leading naturalistic theories of reference assume that reference is determinate for subsentential expressions and so the inscrutability thesis in the form that I have defended it is very much a concern for these theorists. It is clearly possible for there to be a language whose speakers are

---

15. Hilary Putnam, Reason, Truth and History, 73.

able to refer to rabbits even though there is no predicate of the language that divides its reference over rabbits. Speakers of such a language could, for example, pick rabbits out demonstratively or by description. I have argued that given Quine's assumptions about interpretation, it is possible to interpret our own language in this way.

As Searle notes, 'If Quine is right, the [inscrutability] thesis has vast ramifications for the philosophy of language and mind.'[16] In order to avoid these ramifications, many have questioned Quine's assumptions about the constraints on interpretation. One of the premises of Quine's argument for the inscrutability thesis is that a naturalistic account of language recognizes no semantic properties beyond those implicit in speaker's dispositions to behavior. According to Quine, whatever semantic properties there are will be determined by what responses to stimuli speakers are disposed to put forward. This is not to deny that there are any internal mechanisms which mediate the response to a given stimulus, only to deny that they are relevant to constructing an empirical theory of reference.

Searle describes this assumption as 'linguistic behaviorism with a vengeance'. He suggests that the best way to understand the argument for the inscrutability thesis is as a *reductio ad absurdum* of linguistic behaviorism: '[I]f all there were to meaning were patterns of stimulus and response, then it would be impossible to discriminate meanings, which are in fact discriminable.'[17] The main reason that Searle believes this is that he claims to know, in his own case, that when he utters 'rabbit' he intends to refer to rabbits and not undetached rabbit-parts. Searle's route to this conclusion is based on Chomsky's objection to Quine.

---

16. John Searle, 'Indeterminacy, Empiricism, and the First Person,' The Journal of Philosophy (1987), 123.

17. Ibid., p. 125.

Chomsky objects against Quine that the thesis of the indeterminacy of translation, and hence the inscrutability thesis, is no more than a case of underdetermination of a hypothesis by empirical evidence.[18] This sort of underdetermination does not imply indeterminacy. It may simply be the case that a particular theory is underdetermined by all the available evidence, as not all possible evidence has been considered. It may be that a particular theory is underdetermined by all possible evidence for us. The idea here is the realist one that there may be facts that it is physically impossible for us to acquire the evidence for. On Chomsky's view, linguistic theory is no worse off than physical theory with respect to indeterminacy. Quine responds to this objection as follows:

> Though linguistics is of course part of the theory of nature, the indeterminacy of translation is not just inherited as a special case of the underdetermination of our theory of nature. It is parallel but additional. Thus, adopt for now my fully realistic attitude towards electrons and muons and curved space-time, thus falling in with the current theory of the world despite knowing that it is in principle methodologically under-determined. Consider from this realistic point of view, the totality of truths of nature, known and unknown, observable and unobservable, past and future. The point about indeterminacy of translation is that it withstands even all this truth, the whole truth about nature.[19]

As Louise Antony points out, the phrase 'totality of truths of nature' implies that Quine takes all the natural facts there are to be physical facts, i.e. facts admitting

---

18. Noam Chomsky, 'Quine's Empirical Assumptions.' In Words and Objections: Essays on the Work of W. V. Quine, edited by Davidson and Hintikka (Dordrecht: Reidel Publishing Company, 1975).

19. W. V. Quine, 'Reply to Chomsky.' In Davidson and Hintikka, 303.

description in the language of physics.[20] She criticizes this assumption, arguing that in order to adequately account for human behavior, 'special sciences' not reducible to physics may be required. Linguistics and cognitive science in general enjoy a high degree of scientific legitimacy, says Antony, not least because the posits of cognitive sciences are justified in the same way that those of the physical sciences are, owing to their utility in explaining the phenomena. The limitation to a behaviorist psychology imposed by Quine, she says, is analogous to a restriction that physical theories posit only observable objects.

Searle says of the Chomsky objection that it misses Quine's point altogether. This is because, he says, 'Quine assumes from the very start the nonexistence of (objectively real) meanings in any psychological sense.'[21] In Searle's view, instead of showing that it is not possible to give an 'intentionalistic theory of meaning' as is commonly supposed, it is only on the assumption that there are no 'intentionalistic meanings' that the argument for the inscrutability thesis can succeed.[22] This is, he says, to beg the question against mentalism. Once the question-begging assumption is jettisoned, Chomsky's objection is valid.

Jerry Katz shares this diagnosis of what has gone wrong with Quine's argument. Quine is arguing that a number of semantic notions essential to 'traditional' philosophical thinking about language, such as sense, synonymy and analyticity lack any sort of scientific legitimacy. The inscrutability thesis adds the traditional notion of reference to this list. Katz does not accuse Quine of assuming an indefensible behaviorism, however:

---

20. Louise Antony, 'Naturalized Epistemology and Language.' In Naturalistic Epistemology, edited by Shimony and Nails.

21. Searle, 129.

22. Some, such as Jerry Katz, would say that what the indeterminacy thesis rules out is an intensionalist theory of meaning.

> The behaviorism he has in mind here [that linguistic meaning is determined by speakers' dispositions to behavior] is not the fierce reductive doctrine of days gone by, but merely a way of putting the study of language on a par with other sciences by requiring the linguist's theoretical constructions to be justified on the basis of objective evidence... Since it merely faces [linguists] with the task of arriving at a theory of language on the basis of the overt behavior of its speakers in overt circumstances, Quine's behaviorism is a behaviorism we can live with.[23]

Katz claims that the argument for the thesis of the indeterminacy of translation does not follow, even if one accepts what Searle calls linguistic behaviorism. The key issue for Katz is whether actual translation is relevantly similar to Quine's situation of radical translation, in which Katz grants that indeterminacy obtains. He identifies the reason that there is indeterminacy of translation in the situation of radical translation as the absence of what he calls 'independent controls'. The issue then, is whether independent controls are also lacking in actual translation.

Quine takes himself to have shown that there is no scientifically acceptable synonymy relation.[24] According to Katz, Quine then goes on to use this as a premise in the argument for the indeterminacy thesis in Word and Object. The key argument is that synonymy cannot be defined on the basis of linguistic theory. Suppose, for example, that one attempts to define synonymy on the basis of substitution criteria. If one chooses an intensional context to do this, such as ⌜Necessarily, all x are F⌝, with the understanding that any two expressions are synonymous which yield the same truth-value as values of the variable x, then the procedure is circular. This is because the context itself has to be specified in terms

---

23. Katz, The Metaphysics of Meaning (Cambridge, MA: MIT Press, 1990), 179.

24. W. V. Quine, 'Two Dogmas of Empiricism,' in From a Logical Point of View (Cambridge, MA: Harvard University Press, 1953).

of synonymy or a related concept. If one chooses an extensional context in which to carry out the substitution, truth will no longer do as the feature left invariant by substitution. This is because the preservation of truth-value in extensional contexts does not discriminate synonymous from nonsynonymous expressions. Instead, one must say that the substitution preserves necessary truth or analyticity. These concepts are related to synonymy and so the procedure is again circular.

Katz accepts this argument, but does not accept the conclusion that there is no scientifically acceptable synonymy relation. This is because Katz recognizes another form of definition besides substitution criteria, that he calls 'theoretical definition'. The important of feature of theoretical definitions is that, unlike definitions on the basis of substitution criteria, they permit expressions to be defined using concepts related to the concept expressed by the expression being defined. The possibility of giving a theoretical definition of synonymy provides Katz with independent controls present in actual translation, which he believes show that actual translation is determinate.

These independent controls take the form of the linguists asking the informant about the sense properties of expressions in his native language. Katz suggests that in order to support his interpretation, Ling1 may ask Inf whether 'gavagai', when modified by the native expression identified as meaning 'undetached' in Int1, is redundant like 'unmarried bachelor', or contradictory when modified by the native expression identified as meaning 'detached' in Int1, like 'married bachelor'. Katz also suggests that Ling1 may ask whether a gavagai bears the same relation to some other object as a finger bears to the hand of which it is a part. The possibility of receiving useful answers to questions such as these does not refute the inscrutability thesis. Using schema (∃), the expression 'gavagai' when employed in complete sentences in the native language is used to refer to rabbits. This being so, even though 'gavagai' is true of undetached rabbit-

parts, Inf will not accept the redundancy of the statement that there is an undetached gavagai, or that the statement that there is a detached gavagai is contradictory. Neither will Inf accept that there is some object *X* such that some gavagai bears the same relation to *X* as a finger bears to the hand of which it is a part. As both Int1 and Int2 predict the same responses on the part of Inf to Katz's questions, the inscrutability thesis remains intact.[25]

Unlike Katz, Searle believes that the flaw in Quine's argument is to demand that all semantic facts be publicly available, so that when an interpreter is unable to make a semantic distinction on the basis of public evidence there is no genuine distinction at all. Searle's view is that the argument for the inscrutability thesis amounts to no more than a case for the underdetermination of hypotheses about meaning on the basis of publicly observable facts.[26] The common-sense objection to the publicity requirement is that I know in my own case that I am referring to rabbits rather than undetached rabbit-parts by 'rabbit'. That some observer is unable to conclusively infer this from my public behavior does not persuade common-sense that this is false. All this means is that from the third-person perspective, hypotheses about the contents of other minds is underdetermined. Davidson considers this objection to the inscrutability thesis, but says only that 'one should stand firm against this thought' and repeats the claim that all semantic features are public.[27]

---

25. This is not to say that Katz is unable show that indeterminacy of translation fails. Failing to show produce data to prefer Int1's assignment of reference to 'gavagai' over Int2's does not show that there are no data to prefer whatever sense 'gavagai' has in Int1 over whatever sense it has in Int2.

26. Searle makes this claim about Davidson's argument for the inscrutability thesis in his 'The Inscrutability of Reference.' The point applies equally to Quine's version of the argument, as this publicity requirement is the basis of Quine's linguistic behaviorism.

27. Davidson, 235.

Searle insists that this claim is false and common-sense is quite right that we know what we mean from the first-person perspective. What each of us knows, he says, about the semantic properties of our own mental states (and so what we mean by our public utterances) goes beyond our knowledge of the conditions under which we would hold our sentences true. So, for someone else to genuinely understand another it is necessary for them to know more than the circumstances under which the person they seek to understand would hold his sentences true. If so, it is unremarkable that the circumstances under which someone holds his sentences true fails to determine an interpretation of his language.

In real life, says Searle, we are able to understand what others mean despite this being underdetermined by the observable evidence. This is because we share what he calls the Network of shared assumptions and the Background of nonrepresentational mental capacities. It is the Background, he says, that makes it 'quite out of the question' that when someone speaking English says 'There is a rabbit' that one is equally justified in taking him to be speaking of rabbits or undetached rabbit-parts. Of course, in the ordinary course of events, the alternative interpretations suggested by Quine are out of the question. It would be most unusual for one of the alternative interpretations to even occur to someone, and even if one was suggested, it would be rejected out of hand without being given any serious consideration. There must be something that makes people behave in this way, and it is not because they have considered the facts about speakers' dispositions.

That people actually do behave in this way does not mean that Searle is correct to say that there are facts about meaning and reference beyond what is implicit in dispositions to speech behavior. Whatever it is that actually guides us in interpretation need not be based on genuine evidence for the semantic properties of others' mental states and public utterances, nor need it be related to whatever

makes it the case that our mental states and utterances have the semantic properties that they do. If our mental states have semantic properties there must be something that determines what these are. Given Searle's other commitments, it is hard to know what it is that he believes to determine the semantic facts.

Searle is not just claiming that we can know what the semantic properties of our mental states are independently of knowing what our dispositions to speech behavior are. He is saying that the semantic properties of our mental states are such that they do not reveal themselves in our dispositions to speech behavior. This is a much more implausible claim. Antony is forced to make a similar claim as a consequence of her criticism of Quine. She notes that it would be unwarranted for Quine to stipulate that all the facts relevant to the determination of semantic properties must be stated in the language of physics, thus ruling out the posits of an unreduced cognitive science. Antony is not arguing that the posits of cognitive science will enable one to show how Quine's alternative interpretations will subtly alter speech dispositions and so determine reference. Instead, she must agree with Searle that it is possible to explain why 'rabbit' refers to rabbits and not undetached rabbit-parts, even though there is no difference in speakers' dispositions to speech behavior on either interpretation.

This is implausible because of the nature of the theories put forward to explain the intentionality of the states posited by cognitive science. The most promising of these theories account for the content of mental states on the basis of the referent causing a token of the state. Essentially, the content of mental states is explained by the circumstances under which they are tokened. One would expect, then, that where there is no difference in the circumstances under which a mental state is tokened, that there is no difference in content.

However, as Searle cautions, 'One of the peculiar features of this entire discussion is the speed with which breath-taking conclusions are drawn on the

basis of a few sketchy remarks and underdescribed examples.'[28] In order to see whether he is correct that what we know of our own mental states from the first-person perspective assures us that reference is determinate, the most promising theories of mental content must be closely examined.

28. Searle, 133.

CHAPTER 3

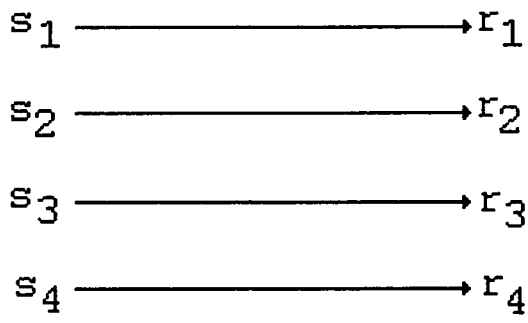
INFORMATION AND INTENTIONALITY

Probably the best way to respond to the broad challenge to the possibility of a theory of intentionality posed by Quine's inscrutability thesis is to articulate an account which shows how reference can be naturalistically determined. One of the most significant elements of recent theories of reference is the notion of information. This provides the basis for an articulation of a causal theory of reference which does not presuppose any intentional notions. The central text dealing with the application of information to the problem of intentionality is Dretske's Knowledge and the Flow of Information.

Dretske bases his account on the mathematical theory of information, or communication theory. This theory is concerned with the quantity of information produced by particular events. The information produced by an event is measured by the number of possibilities which that event eliminates. For example, suppose one amongst eight persons is to be selected for some task. When the person is selected, the number of possible outcomes is reduced from eight to one. In communication theory, the information produced by an event is measured in binary digits (bits)of information. This is a measure of the number of decisions between two equally likely alternatives, that is binary decisions, it would take to match the reduction in possibilities brought about by the event. The reduction of eight possibilities to one takes three such decisions and hence the selection of one person from eight generates 3 bits of information.

Although the transmission of information ordinarily depends upon causation, and I have suggested that an information-based theory provides an articulation of the causal theory of reference, Dretske maintains that transmission of information is possible without causal processes. Suppose, for example, that there is a variable

at the source that can assume one of four different values : s_1, s_2, s_3 or s_4. There is also a variable at the receiver that can assume four difference values: r_1, r_2, r_3 or r_4. If s_2 occurs at the source, let us suppose that this causes r_2 to occur at the receiver. s_2's occurrence generates 2 bits of information (the reduction of four possibilities to one). Although s_2 causes r_2, this does not mean that r_2 also carries 2 bits of information about s. The causal relation between s_2 and r_2 is sufficient to create a signal, but there may be some equivocation in this signal. In order to determine the extent of the equivocation, the other relations between the source and receiver must be examined.

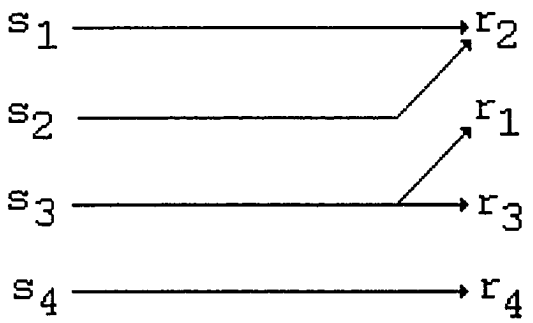


A 1:1 correspondence between source and signal

Suppose that s_1 uniquely causes r_1, that s_3 uniquely causes r_3 and that s_4 uniquely causes s_4. In this situation, should s_2 occur, there will be no equivocation in the occurrence of r_2 and r_2 will carry 2 bits of information about s. If instead, there are other causes of r_2, say, s_1, the occurrence of r_2 would then contain equivocation.

Assume that s_1 and s_2 are equally likely causes of r_2, and that there are no other causes of r_2. As the equivocation of r_2 means that instead of reducing four possibilities to one, four possibilities are reduced to two, r_2 now carries only 1 bit of information about s instead of the 2 bits that are generated at s. This means that the amount of r_2's equivocation is 1 bit. If the other possibilities at r (i.e. r_1, r_3 and

r_4) are not equivocal, the average amount of equivocation at r will, of course, be less.



r_2 is equivocal in this system

Dretske proposes to calculate the average equivocation of a signal as follows. First, the equivocation of each possible event that may occur at the receiver, r_i, is calculated. For the equivocation of a particular signal, $E(r_i)$, one sums the probability of each possibility at s given r_i multiplied by the information generated, given r_i, by the realization of that possibility. The calculation of the equivocation of a univocal signal, such as r_4 in the system pictured above, should of course result in zero. The probability of each of s_1, s_2, and s_3, given r_4, is 0. The probability of s_4, given r_4, is 1. Therefore, given r_4, the information generated by s_4 is 0. So $E(r_4) = 0$.

The formula for the equivocation of r_2, $E(r_2)$, is given below. $P(s_i / r_i)$ is the conditional probability of s_i, given r_i. When there are a range of possible events, $s_1, s_2, \ldots, s_n$, not all of which are equally likely, the amount of information generated by a particular event, s_i, is given by $- \log_2 p(s_i)$.

$$E(r_2) = -\Sigma P(s_i / r_2) * \log_2 P(s_i / r_2)$$

As The probabilities of each of s_3 and s_4, given r_2, are zero, in this case the formula may be simplified to the following:

$$E(r_2) = -[P(s_1/ r_2) * \log_2 P(s_1/ r_2) + P(s_2/ r_2) * \log_2 P(s_2/ r_2)]$$

As the probability of s_1 given r_2 is one-half, as is the probability of s_2 given r_2, the formula becomes

$$E(r_2) = -[0.5 * \log_2 0.5 + 0.5 * \log_2 0.5] = 1$$

The average equivocation at r is calculated by summing the equivocation of each signal multiplied by the probability of its occurrence:

$$E(r) = P(r_1) * E(r_1) + P(r_2) * E(r_2) + P(r_3) * E(r_3) + P(r_4) * E(r_4)$$

If none of r_1, r_3 or r_4 are equivocal, then the average equivocation at r would be 0.5.

In the above diagram both s_1 and s_2 are shown as causing r_2, hence r_2 is equivocal. s_4 causes r_4 and s_3 brings about either r_1 or r_3 by, let us suppose, an indeterministic process. r_4 is clearly non-equivocal, and so are r_1 and r_3. If either r_1 or r_3 occurs, the full 2 bits of information about the source is carried: that of the four possibilities at s, it was s_3 that occurred. Dretske points out that although in this system there is a tight causal connection between s_2 and r_2 and a relatively loose connection between r_1 and s_3, r_1 carries more information about s than r_2 does. This is because an effect does not embody information about its cause, if indistinguishable results could have been the product of a different cause. The results of an indeterministic process may carry information about the events which brought it about, however. Dretske likens this sort of case to a doorbell with a loose wire in its circuit. Even though the doorbell rarely rings when it is pushed, when it does ring, as it occasionally will, the ringing bell carries the information that the button is being pushed. Although the average amount of information carried by this signal is very low, the system will sometimes carry as much information as a doorbell in perfect working order. As Dretske does not consider indeterministic processes to be causal, he maintains that a causal relation between events A and B is not necessary for B to carry information about A.

This claim is not as surprising is it first sounds when one bears in mind that it depends upon information being carried by indeterministic processes. Dretske does give an example of an informational relation between two events which does not depend on any direct interaction between the events at all. He observes that 'From a theoretical point of view... the communication channel may be thought of as simply the set of dependency relations between s [the information source] and r [the receiver].'[1] He gives the example of a television transmitter, *A*, and two television sets tuned to the transmitted channel, *B* and *C*. There is a direct causal link between *A* and *B* and *A* and *C*. *B* carries information about the transmitter, *A*, as does *C*. *B* also carries information about *C*, and *C* information about *B*, despite the absence of any causal or physical connection between *B* and *C*. In fact, if *C* is further away from the transmitter than *B*, says Dretske, *B* carries information about what *will* happen at *C*. Supposing that no physical signal can travel backwards in time from *C* to *B*, this demonstrates the distinction between information and causation.

Although a causal relation between two events is not necessary for an informational relation between them, the theory of information may still provide the basis for a causal theory of reference if a causal relation is sufficient for an informational relation. That *A* causes *B* does not suffice for *B* carrying the information that *A* has occurred. Whether a particular instance of causation results in information being transmitted depends upon what other potential causes of the signal there are, i.e. whether the signal is equivocal. The possibility of equivocation in the signal is likened by Dretske to the story of the shepherd who shouted 'wolf' in situations other than those involving wolves threatening his

---

1. Dretske, Knowledge and the Flow of Information (Cambridge, MA: MIT Press, 1981), 38.

sheep. With this history, when he shouts 'wolf' in the presence of a threatening wolf, his shout no longer conveys the information that a wolf is present. The problem that equivocation in the signal poses to a theory of reference based on information, or more generally causation, is known as the problem of error, or the disjunction problem. The problem is that virtually every real signal will sometimes be made in error. Even highly competent speakers of a language are not immune from misidentifying an object and applying to it a term which does not properly apply to it. Under certain circumstances, for example, cats can be mistaken for dogs. On a pure informational theory of content, the extension of the term 'dog' would then include at least some cats (making the extension effectively the class of dogs-or-cats, hence suggesting the name 'disjunction problem'). The task of providing a satisfactory solution to the disjunction problem will take up much of the subsequent discussion.

Dretske notes that communication theory is not itself concerned with semantics. It is concerned with the amount of information that is carried by a signal, not what that information is, or what the content of the signal is. For example, when one amongst eight people is chosen, communication theory tells us that the event of choosing generates 3 bits of information, but this is true no matter which individual is chosen by the process. Furthermore, the orthodox concern of communication theory is not the specific amount of information carried by particular signals, but the average amount of information carried by a channel of communication.

Consider, for example, a chess board on one square of which is placed a marker. There are 64 possible positions of the marker. Let us suppose that it is equally likely that the marker will be placed on any particular square. When the marker is placed on a square the 64 possibilities are reduced to 1 and so 6 bits of information are generated by this event. Suppose that I wish to communicate this

information, but have at my disposal a communication channel that is only capable of transmitting a single binary digit (e.g. the channel only permits the transmission of one of two symbols, '0' and '1', one time only). Although I only transmit one binary digit on my channel, I can occasionally use it to transmit more than one bit of information. If, for example, I send '1' when the marker has been placed on the square in the bottom right-hand corner of the chess board and '0' otherwise, I transmit 6 bits of information when I send '1'. My signal '0' reduces 64 possibilities to 63, which means that only approximately 0.02 bits of information is transmitted about the location of the marker. As the marker is equally likely to be placed on each of the squares on the board, the average amount of information transmitted by this system will be 0.11 bits.[2]

The interesting thing about this example is that given the nature of the channel, no more than 1 bit of information may be transmitted by it, on average. This does not prevent individual messages from being significantly more informative, so long as there are likely to be many more individual messages that bring the average information carried down to at most 1 bit. This may suggest the irrelevance of communication theory to semantics. That theory deals in the average quantity of information transmitted by a channel, while the information carried by a particular signal, to say nothing of its semantic value, is unconstrained by the channel's capacity. As Dretske himself points out, there is no sense to be made of the notion of the 'average meaning' of a number of messages.

Even in the light of communication theory, there simply is no way to assign a numerical value to the meaning or content of a signal.[3] The quantity of

---

2. -Log2 (63/64) = approximately 0.02. (63 * 0.02 + 6)/64 = approximately 0.11. As one would expect, Dretske's formula for the equivocation of a signal yields a value of approximately 5.98 for the signal '0'.

3. Ibid., 48.

information that a signal carries, which in communication theory is an index of the number of possibilities the signal eliminates, is not to be identified with its meaning. The information that an utterance of a sentence conveys changes according to context, for example, although its conventional meaning does not. So, if I utter the sentence 'The Democratic candidate did not win the election' in a context where there were only two candidates, one of whom must win, I reduce two possibilities to one. If I utter the same sentence in a race with three possible winners, I reduce three possibilities to two and hence convey less information.[4] The semantic meaning of the sentence is the same in both contexts, however. The statement of necessary truths (if there are any) would convey zero information, as there are no other possibilities. There are few, however, who would maintain that the candidates for necessary truths, such as elementary arithmetical truths like 2 + 2 = 4, have no meaning.

Although communication theory is concerned primarily with the average information transmitted across a channel, it does provide a measure of the amount of information generated by a particular event and the amount of information a signal carries about such an event (this is the information generated by the event minus the equivocation of the signal). At the very least, communication theory licenses Dretske to assume that there are such quantities in developing a semantic theory based on information.

---

4. Assuming that each candidate in each race has the same probability of victory, when there are two candidates my statement carries $-\log_2(1/2) = 1$ bit of information. When there are three candidates my statement carries $-\log_2(2/3)$ = approximately 0.6 bits of information. This example assumes that I pronounce infallibly on electoral matters.

Dretske's Semantic Theory

Dretske proposes a fundamental principle concerning the transmission of information that he names the 'Xerox principle':

If *A* carries the information that *B*, and *B* carries the information that *C*, then *A* carries the information that *C*.

As Dretske notes, if there are to be communication systems with anything other than minimal complexity, the Xerox principle must be satisfied. The transmission of a message by radio, for example, involves several links from the sound waves going into a microphone to the sound waves coming out of a speaker. If information is not preserved over these links, the sound waves produced by the speaker cannot carry the information the sound waves going into the microphone did.

Dretske gives three conditions that a definition of information must satisfy. For a signal to count as carrying the information that *s* is *F*:

(A) The signal carries as much information about *s* as would be generated by *s*'s being *F*.

(B) *s* is *F*

(C) The quantity of information the signal carries about *s* is (or includes) that quantity generated by *s*'s being *F*.

Condition (A), named 'the communication condition' by Dretske seems harmless enough (if one remembers the relativity of amount of information generated) and follows from his Xerox principle. Dretske notes that (A) is insufficient to guarantee that the information that a signal carries about *s*, though great enough to be the information that *s* is *F*, actually is the information that *s* is *F*. A necessary condition of any signal carrying this information is that *s* actually is *F* (or at least was when one follows the signal back to its source). Although conditions (A) and (B) are individually necessary for a signal to carry the

information that *s* is *F*, they are clearly not jointly sufficient. For example, suppose that *s* is a red square, that its being red generates 2 bits of information and also that its being square generates 2 bits of information. A signal could carry the information that *s* is red without carrying the information that *s* is square.

To remedy this deficiency, Dretske proposes condition (C). Condition (C) looks somewhat odd, as the *quantity* of information may presumably remain the same although the content carried varies. Dretske himself admits this: 'It is not clear... what it could mean to say that one quantity (amount of information the signal carries) is (or includes) another quantity (amount of information generated) when this is meant to imply something more than a numerical comparison.'[5] Clearly, what Dretske intends is that the information carried by the signal should be the information that *s* is *F*, but just saying that would be entirely unilluminating.

Dretske proposes the following as his definition of information content:

A signal *r* carries the information that *s* is *F* = The conditional probability of *s*'s being *F*, given *r* (and *k*), is 1 (but given *k* alone, less than 1).[6]

The variable *k* stands for whatever the 'receiver' already knows about the source. So, for example, if someone knows that *s* is not blue, and a signal indicates that *s* is either red or blue, this signal effectively carries the information for that person that *s* is red. The applicability of *k* is restricted to instances where the receiver is a system which is capable of storing past information, such as a being with knowledge.

Dretske says that this is the only definition that will satisfy conditions (A),. (B) and (C). Clearly (A) and (B) are satisfied, as the occurrence of *r* implies that *s* is *F*. Dretske also says that his definition satisfies (C), as 'whatever *other*

---

5. Ibid., p. 64.
6. Ibid., p. 65.

quantities of information the signal may carry about *s*, our definition assures us that the signal includes the *right* quantity (the quantity associated with *s*'s being *F*)...'[7] As pointed out above, the talk of 'quantity' here is misplaced. What Dretske really means to say is that whatever other information *r* may carry about *s*, his definition ensures that *r* carries the information that *s* is *F*. This is odd, as the point of specifying conditions (A), (B) and (C) is that whatever satisfies them is a satisfactory account of information content. It seems, however, that in order to know whether (C) is satisfied, we already need to know that the definition is successful.

This actually isn't a very serious problem. Condition (C) was introduced in order to rule out a definition of information content that would not distinguish, for example, between the information that a red square is red and the information that it is square, when both these facts about the square generate the same quantity of information. If Dretske's definition is successful in distinguishing between these different pieces of information, then his definition has overcome the problem which (C) was supposed to screen theories of information for. Even though (C) is not a very helpful condition for this purpose, there is no reason to think his definition inadequate on these grounds.

Dretske notes that his definition provides an account of what one might call a signal's *de re* informational content only. This means that on this definition, the information that *s* is *F* and the information that *t* is *F* will count as being the same piece of information, so long as $s = t$. There are well-known philosophical puzzles concerning the phenomenon that someone might find, say, the statement that George Orwell is the author of 'Animal Farm' less informative to them than the statement that Eric Blair is the author of 'Animal Farm', despite the fact that

---

7. Ibid., p. 66.

George Orwell is the same person as Eric Blair. So, if a signal carries the information that George Orwell is the author of 'Animal Farm', it may not carry the information that the author of that work was named 'George Orwell', or even that the work is titled 'Animal Farm'. Although this feature of the definition of information content that Dretske proposes appears to limit the scope of the resulting theory to only *de re* rather than *de dicto* contents, it would be accomplishment enough to provide a successful account of *de re* content. Moreover, the success of an analysis of *de dicto* content in terms of *de re* content should not be ruled out at present.[8]

Importantly, Dretske recognizes that, in his sense, there is no such thing as 'the' information that a signal carries. If a signal, *r*, carries the information that an object, *s*, is a square, then it will also carry the information that *s* is a rectangle. The reason for this is that *r* carries the information that *s* is *G* when the conditional probability of *s*'s being *G*, given *r*, is 1. As the probability of a square being a rectangle is 1, when the probability of *s* being square, given *r*, is 1, the probability of *s*'s being a rectangle, given *r*, is also 1. On Dretske's definition, this is sufficient for *r* carrying the information that *s* is a rectangle. As no squares are circles, *r* will also carry the information that *s* is a non-circle. So, if *r* carries any information at all about *s*, it will carry indefinitely many pieces of information about *s*.

A signal, *r*, which carries information about one thing may also carry information about other things in virtue of that fact. If *r* carries the information that *s* is *G*, and the conditional probability of *t*'s being *H*, given that *s* is *G*, is 1 then *r* will carry the information that *t* is *H*. Dretske illustrates this with the connection between a rise in temperature in the expansion of a sample of mercury.

---

8. Nathan Salmon, for example, attempts such an analysis in his Frege's Puzzle.

If the probability that the temperature has risen, given the expansion of the sample, is 1, then any signal which carries the information that the sample has expanded will also carry the information that the temperature has risen.

Dretske describes these sorts of situations as ones where information is nested in a state of affairs:

The information that *t* is *G* is nested in *s*'s being *F* = *s*'s being *F* carries the information that *t* is *G*.

Dretske distinguishes between information that is, as he puts it, 'analytically nested' in a state of affairs and one that is 'nomically nested'. The example of the square being also a rectangle is an example of analytic nesting, whereas the example of the expansion of mercury according to a rise in temperature is an example of nomic nesting. These cases where information is nested in a state of affairs are to be contrasted with the case of simple correlation. Suppose that everything that is *F* is also *G*. If a signal carries the information that *s* is *F*, it might not carry the information that *s* is *G*. Whether it does or not depends upon the reason for the correlation. If *F* and *G* are correlated according to some 'law of nature or principle of logic', says Dretske, then the probability of *s*'s being *G*, given that it is *F*, is 1 and so the signal would carry the additional information that *s* is *G*.[9] If the correlation between *F* and *G* is coincidental, then the probability of *s*'s being *G*, given that it is *F*, will be less than 1.

Dretske provides an example in order to illustrate his claim. Suppose that all of Herman's children have the measles, and that Alice is one of Herman's children. Dretske says of this example

> Presumably the fact that Herman's children... happened to contract the measles at the same time does not make the probability of their having the

9. Dretske, 74.

> measles, given their common parentage, 1. Since this is so, a signal can carry the information that Alice is one of Herman's children without carrying the information that she has the measles despite the fact that all Herman's children have the measles.[10]

Although this claim seems reasonable enough, consider the following example. Suppose that every coin in my pocket at time *t* is a dime. Although no law of nature or principle of logic supports this correlation, the probability of a coin being a dime, given that it is a coin in my pocket at time *t*, is 1. The only significant difference between this example and Dretske's is the reference to time. Dretske is correct to say that the fact that all *F*s are *G*s does not necessitate that any signal which carries the information that *s* is *F* must carry the information that *s* is *G*, but he is wrong to suggest that there must be some nomic or logical connection between *F* and *G* for the information that *s* is *F* to be sufficient for the information that *s* is *G*.

Dretske hopes to show that the intensionality exhibited by states with semantic properties, such as beliefs, will be explained by the reliance of the transmission of information on nomic processes. Clearly, when there is a nomic relation between two properties, so that when *s* is *F*, *s* is also *G*, this does not imply that whoever believes that *s* is *F* also believes that *s* is *G*. If it is physically (or at least biologically) necessary, say, that all creatures with a heart are creatures with kidneys, it is still possible for someone ignorant of the law to believe that a certain being has a heart while failing to believe that it has a kidney. Dretske aims to get around this problem with his definition of semantic content.

The crucial element in Dretske's definition of semantic content is his distinction between a signal carrying information in what he calls 'analog' form

---

10. Ibid.

and its carrying information in what he calls 'digital' form. A structure, *S*, carries the information that *t* is *F* in analog form if and only if *S* carries the information that *t* is *F* and it carries the information that *t* is *G*, where the fact that *t* is *F* is nested (either nomically or analytically) in *t*'s being *G* (but not vice versa). For example, if *S* carries the information that *t* is a square, it will also carry the information that *t* is a rectangle, a parallelogram and a quadrilateral. As these latter pieces of information are nested within the information that *t* is a square, but that *t* is a square is not nested in any of them, they are carried by *S* in analog form.

This fact about the way information is carried raises the problem of indeterminacy. If one wishes to nominate a piece of information that *S* carries in analog form as its content, the problem is that it will very probably carry many other pieces of information in analog form as well. For this reason, Dretske suggests that the semantic content of a structure is that information it carries in digital form. *S* carries the information that *t* is *F* in digital form if and only if *S* carries the information that *t* is *F* and it does not carry any other piece of information, *t* is *G*, which is such that the information that *t* is *F* is nested in *t*'s being *G* (but not vice versa). Dretske explains the notion of *S* carrying the information that *t* is *F* in digital form as *S* carrying no more specific information about *t* than that it is *F*. So if *S* carries the information that *t* is a square, and that *t* is a rectangle, the information that *t* is a rectangle cannot be carried in digital form by *S*, as the information that *t* is a square is more specific. Similarly, if *t* carries the information about *t* that it is a red square, *S* cannot carry the information that *t* is a square in digital form.

It is not possible to simply identify the semantic content of a structure with what information it carries in digital form, however. Although *S*'s carrying the information that *t* is *F* in digital form means that there is no more specific information that *S* carries about *t*, *S* may carry the information that some other

thing, *r*, is *G*. The information that *t* is *F* may be nested within the information that *r* is *G*. In this case, the problem of indeterminacy which the distinction between digital and analog information transmission was supposed to solve reappears. As it is possible for information carried in digital form to be nested within other pieces of information, it is possible for a structure to carry more than one piece of information in digital form.

In order to solve this problem, Dretske identifies the semantic content of a structure as follows with the information that it carries in what he calls 'completely digitalized form'. His defines completely digitalized information as follows:

Structure *S* carries the information that *t* is *F* as its semantic content if and only if

(a) *S* carries the information that *t* is *F*

(b) *S* carries no other piece of information, *r* is *G*, which is such that the information that *t* is *F* is nested (nomically or analytically) in *r*'s being *G*.

As Dretske points out, this definition has the consequence that whatever information a structure has as its semantic content will be carried by it in digital form, but not all the information it carries in digital form need be its semantic content.

There is an obvious deficiency with the identification of the semantic content of a state with the information it carries in completely digitalized form. On this identification, it is impossible for there to be false semantic contents, for on Dretske's definitions, a necessary condition of a signal carrying the information that *s* is *F* is that *s* actually is *F*. This fact forms the basis of the disjunction problem. Those who would seek to base semantics on information must find a way of distinguishing between the information that a particular state actually carries and the information that is its semantic content.

Dretske, of course, is aware of this problem. His attempt to remedy this deficiency, and thus avoid the disjunction problem, appeals to what he calls 'the learning situation'. During this period, the system is exposed to signals, some of which carry the information that certain things are *F*, and others which carry the information that certain things are not *F*. Dretske's idea is that at the beginning of the process, the system receives this information in only analog form. The system learns, according to Dretske, when it develops an internal state which carries the information that something is *F* in digital form. Such a state is a semantic state, on Dretske's account. Dretske believes this solves the disjunction problem:

> Once this structure is developed, it acquires a life of its own, so to speak, and is capable of conferring on its subsequent tokens... *its* semantic content (the content it acquired during [the learning situation] *whether or not these subsequent tokens actually have this as their informational content.* In short, the structure type acquires its meaning from the sort of information that led to its development as a cognitive structure.[11]

Dretske gives as an example of this process the teaching of a child to recognize birds. The child is shown a number of examples of robins 'in close range and in such a way that their distinctive markings and silhouette are clearly visible'.[12] This sort of care is necessary, as unless the signals sent in training actually carry the information that there is a robin present, it will not be possible to develop a state that digitalizes this information. If the trainee was shown robins that are too far away for the human visual system to distinguish them from other birds, the signals received by the trainee would not carry the information that the

---

11. Ibid., 193. In some ways, this is a proto-teleological account.
12. Ibid., 195.

displayed bird is a robin, as they would not eliminate the possibility that some other bird was being displayed.

The training consists in the child being taught to say 'robin' in the presence of robins and 'not robin' in the presence of birds which are being used as a contrast to robins (Dretske suggests that 'A few bluejays' be 'thrown in' for this purpose). Once the training has been concluded, if the child points to a sparrow and says 'robin' whether the child has said something false in his or her idiolect will depend upon exactly what semantic content was formed during training. If the child developed a concept sensitive to the presence of robins only, then the child will have said something false. Given the poverty of the training, however, it is unlikely that this is the concept that will have been formed. The child may have only been responding to the color of the birds displayed, so that he or she applies the term 'robin' to things that are not blue (as the contrasting items were bluejays). In this case, in the child's idiolect, 'robin' is true of sparrows, on Dretske's account.

There are a number of problems with Dretske's suggestion. By fixing the content of a state according to the content it possesses at a particular point in its history, it is difficult to see how the theory could accommodate the phenomenon of conceptual change. The theory is implausible as an account of what Dretske suggests it is, of how human beings come to acquire concepts. This is because it seems to provide no acknowledgement of what Putnam has called 'the division of linguistic labor'. As Dretske himself admits, when a child is unable to perfectly discriminate robins from non-robins, when the child calls a sparrow a 'robin', what the child says is false. Dretske suggests that what the child believes is a different matter, but if a child (or anyone) is able to 'borrow' linguistic reference, why not mental content? If the child is able to talk about robins, it seems unreasonable to suppose that he or she does not get to have beliefs about robins until his or her

ornithological recognition skills improve.[13] Although it may not provide a plausible account of how humans acquire new concepts or linguistic terms, the theory may still serve as an explanation of how intentionality is possible. It faces further problems, however.

Fodor has criticized Dretske's distinction between the learning situation and subsequent events as unprincipled, suggesting at one point that perhaps a whistle blows to signal the end of training.[14] This is unfair to Dretske, as he states that the learning period for a particular state ends when that state comes to completely digitalize some piece of information. This feature of Dretske's position also provides a reply to Fodor's more substantive objection. Suppose that the system has been conditioned in training to produce tokens of a given state-type in the presence of objects of type *A*. If the state is subsequently tokened in response to the presence of objects of type *B*, Fodor's interpretation of Dretske's claim is that *B*s are not part of the semantic content of the state, as no *B* caused the state to be tokened in training. Fodor's objection is that if, contrary to fact, a *B* had been presented to the system during training, it would have caused the state to be tokened. Fodor's reasoning for this is that as *B*s are sufficient for the state to be tokened after training, they must be sufficient for the state to be tokened during training. So, Fodor concludes, the training process has established not a nomological dependence of tokens of the state upon instances of *A*, but a nomological dependence of tokens of the state upon instances of *A*-or-*B*.

---

13. This is not to say that all of such a child's 'robin' thoughts and beliefs would be about robins. See, for example: Brian Loar, 'Personal References,' in Information, Semantics and Epistemology, edited by E. Villanueva (Oxford: Blackwell, 1990).

14. Fodor, Psychosemantics Cambridge, (MA: MIT Press, 1987), 104. Loewer also accuses Dretske of failing to provide any principled account of the distinction between the learning period and subsequent periods in his 'From Information to Intentionality,' Synthese 70 (1987).

I believe that this objection rests upon a misreading of Dretske. Dretske's position is that the semantic content of a state is determined not by what causes it during training, but by what information it fully digitalizes at the end of training. If the state has fully digitalized the information that there is an *A* present, it will not be caused by *B*s at that time. If, at the end of the learning period, the state is still causally sensitive to the presence of *B*s, then it has not fully digitalized the information that there is an *A* present and that will not be its semantic content. Instead, it may have digitalized the information that an *A*-or-*B* is present. Dretske's solution to the disjunction problem accounts for changes in the information actually carried by a state after the learning period as errors, but it will not permit one to regard any potential cause of the state at the moment the learning period ends as an error.

For Fodor to misread Dretske in this way is entirely reasonable, as the position that Fodor attributes to Dretske, although flawed in precisely the way Fodor points out, is more plausible than the position Dretske actually holds. This is because Dretske does not provide a way of distinguishing, amongst the actual and potential causes of a given state, which ones result in veridical tokens and which ones result in errors. Instead, he points to a moment in the history of a state type and says that all the actual and potential causes of the state at that moment will result in veridical tokens. It is surely unrealistic to suppose that there has ever been such a moment in the history of any actual content-bearing state, in which case Dretske's theory is unable to explain the content that those states possess.

CHAPTER 4

THE DISJUNCTION PROBLEM AND ASYMMETRIC DEPENDENCE

Fodor has long been articulating a view he calls 'the representational theory of the mind'. On this view, thinking or believing consists in the tokening of mental representations. Fodor has argued that these representations are a system of symbols he calls 'the language of thought'. A large part of Fodor's program is explaining how a semantics may be provided for the language of thought: '[W]e would have largely solved the naturalization problem for a propositional-attitude psychology if we were able to say, in nonintentional and nonsemantic idiom, what it is for a primitive symbol of mentalese to have a certain interpretation in a certain context.'[1] He adds that he does not know how to do this, but that it is plausible to suppose that the interpretation of (primitive, nonlogical) symbols in the language of thought is determined by some of their causal relations, in the manner that Putnam has suggested that reference is determined for natural kind terms. So if tokens of 'water' are to be interpreted differently on Earth and on Twin Earth this is because there is some appropriate causal relation that such tokens bear to water on Earth and XYZ on Twin Earth. Fodor primarily intends to give a theory which applies to mental representations. Should his account fail, or any naturalized account fail, in its application to language, he would not accept this as a criticism.[2]

Fodor first articulates what he calls the 'Crude Causal Theory of Content'. On this crude theory, a symbol token is to denote what causes it, and a symbol type is to denote what reliably causes tokens of that type. What Fodor means by the

---

1. Jerry Fodor, Psychosemantics (Cambridge, MA: MIT Press, 1987), 98.

2. Fodor notes (ibid., 100) that a causal theory is much more likely to be successful for a theory of thought than of language, because the entertaining of thoughts is frequently more involuntary than the uttering of linguistic expressions.

reliable causation demand is that the causal dependence is to be counterfactual supporting: 'either instances of the property actually do cause tokenings of the symbol, or instances of the property *would* cause tokenings of the symbol *were they to occur*, or both.'[3] Fodor adds that he takes it to be necessary and sufficient for there to be such reliable causation that there is a nomological relation between two properties of events, for example the property of being an instance of the property dog and the property of being a tokening of the symbol 'dog'.

The crude causal theory amounts to the claim that a symbol expresses a property if it is nomologically necessary that all and only instances of the property cause tokenings of the symbol. There are problems with both parts of this formulation. Firstly, not all dogs cause 'dog' tokens. This isn't merely because dogs exist where there are no perceivers to perceive them, but because perceivers may sometimes fail to recognize a dog *as* a dog. Secondly, not only dogs cause 'dog' tokens. Some cats seen on dark nights, for example, may look sufficiently like dogs that they are mistaken for dogs and so also cause 'dog' tokens.

Probably the easier problem to deal with is that which affects the 'all' clause. Fodor correctly rejects a simple appeal to counterfactuals here. The simple appeal is that even though not all dogs do cause tokens of 'dog', they would if circumstances were right. The difficult part is saying what it is for circumstances to be right. After all, there are nondogs that cause 'dog' tokens under the 'right' circumstances. In order to deal with this problem Fodor develops an account of what are effectively observational concepts, which he terms 'psychophysical concepts'.

Psychophysics, Fodor informs us, is the science that tells us 'how the content of an organism's belief box varies with the values of certain physical parameters in

---

3. Ibid., 98.

its local environment.'[4] So, for example, a psychophysical theory might say that there are circumstances under which a human observer will think that a particular object in his environment is red. The theory will specify how much of a colored surface must be visible to the observer, what the color properties of the surface must be, and how illuminated the surface must be. When the observer faces the surface under these conditions, with eyes open and a normally functioning visual system, the theory says he will think of the surface that it is red.[5]

For psychophysical concepts, it is possible to specify the circumstances under which objects which fall into the extension of the concept will cause a token of the relevant mental representation to occur in an observer. This appears offer a solution to the problem with the 'all' condition of the crude causal theory. The theory would not demand, for example, that all dogs cause the mental representation 'dog', only that all dogs perceived under psychophysically optimal circumstances cause 'dog' tokens. Even this is too strong, however. Even if there are any psychophysical concepts, not all concepts can be psychophysical. Psychophysical concepts are ones for which there is no distinction between seeing something that instantiates the concept and, as it were, seeing it as something that instantiates the concept. In the example of the red-colored surface, a psychophysical theory says that under certain conditions this surface will appear red to a normal observer. If one identifies the concept 'red' as a psychophysical concept on this basis, it is because one believes that the appearance of red that is psychophysically guaranteed is sufficient for the tokening of the concept 'red'.

---

4. Ibid., 113.

5. This is weaker than saying that the observer will believe that the surface is red. The observer may, for example, believe his spectrum to have been recently inverted and so fail to believe that what appears red to him really is so.

Not all concepts are like this. The concept 'dog', for example, is not. Suppose that all dogs are such that they have a characteristically dog-like appearance. It would then be possible for a psychophysical theory to specify the circumstances under which the perception of a dog is sufficient for the tokening of a psychophysical concept. This will not be the concept 'dog', however, but something like the concept 'dog-appearing'. Even if the appearance of redness is sufficient for the tokening of the concept 'red', the appearance of dogness is not sufficient for the tokening of the concept 'dog'.

Fodor hopes to be able to exploit the existence of psychophysical concepts such as these. His idea is that even though not all dogs, under psychophysically optimal conditions, will cause tokens of the concept 'dog' in observers, they will cause the relevant psychophysical concept, which in turn will cause the concept 'dog' in the observer who has internalized an appropriate theory. The relevant psychophysical concept may be highly disjunctive. For example, although most dogs have a characteristic look that enables people with the concept 'dog' to recognize them as such, there might be a rare breed of dog the members of which don't look enough like dogs for them to immediately appear to be dogs to an observer. A typical observer might think of such a creature as a dog only after investigating its genealogical history, for example. So the psychophysical concept upon which the tokening of the concept 'dog' depends would probably not be the stereotype of a dog.

Even so, it would still be possible for a psychophysical concept peculiar to a particular type of object to ultimately fail to produce a token of a concept with that object in its extension, i.e. it is still the case that not all objects of type *F*, even in psychophysically optimal circumstances, will cause tokens of the concept of an *F*. This is because the causal link from the psychophysical concept that all *F*s cause to

the tokening of the mental representation might fail. One reason for this would be that the internalized theory of the observer was incomplete.

Although it might be possible to overcome this problem, it is not necessary to do so. This is because the basic assumptions of informational semantics do not make it a necessary condition of a symbol *S*'s representing objects of type *F*, that all *F*s cause tokens of *S*. The basic assumption is that a symbol *S* carries the information that *p* when the probability that *p*, given *S*, is 1. This will be the case when the only thing that causes *S* is *p*. It does not have to be the case that the probability of *S*, given *p*, is 1, even under some favored circumstances.

There are a number of possible versions of an informational theory, which we may call the actual-history version, the local-instantiation version and the pure-informational version. In the actual-history version, for a type of symbol, *S*, to refer to objects of type *X*, it is necessary that some tokens of *S* have actually been caused by *X*s. In the local-instantiation version it is necessary that the property of being an *X* is instantiated in the vicinity of the representing system, even if no *X*s have actually caused *S*. In the pure-informational version, neither of these things is required, only that the probability of the presence of an *X*, given *S*, is 1. Dretske, for example, presents a pure-informational version of the theory.

Effectively, an informational semantics can get by with the only the 'only' condition. Probably for this reason, difficulties with the 'only' condition of causal theories have been much more discussed than problems with the 'all' condition. The foremost of these former problems is the problem of error, or the disjunction problem. The basic idea behind the causal theory of content is that a symbol represents what causes it. If matters are left at that, all symbols are always veridical, for it is impossible to have a symbol be caused by something that it does not represent. If cats seen on dark nights sometimes cause the tokening of 'dog' then 'dog' cannot mean *dog*, but must instead mean *dog or cat on a dark night.*

However, when cats on dark nights cause tokens of 'dog' it is usually because they are mistaken for dogs. Perhaps it is this phenomenon that led Dretske to find it plausible to propose that the semantic content of a symbol at the end of the learning period is to be taken to be the semantic content of the symbol thereafter. This is certainly the idea that underlies Fodor's solution to the disjunction problem. As Fodor puts it, there is 'independently a semantic relation' between a symbol and its referent.[6] What Fodor means by this is that the error cases are somehow parasitic on the existence of a genuine reference relation.

Fodor hopes to account for this 'independent semantic relation' in terms of counterfactual properties of the causal relation between symbol and referent, which he expresses in terms of possible worlds as follows:

(1) Dogs cause tokens of 'dog' in the actual world.

(2) Cats on a dark night do not cause tokens of 'dog' in nearby worlds in which dogs don't cause tokens of 'dog'.

(3) Dogs do cause tokens of 'dog' in nearby worlds in which cats on a dark night don't cause tokens of 'dog'.

Faced with conditions like these it is very tempting to begin constructing counterexample cases which meet the conditions but we do not regard as genuine cases of reference. For example, the account must afford some means of distinguishing between proximate and distal causes, otherwise we would have no grounds for ruling out the patterns of retinal stimulation caused by dogs as the reference of 'dog'.[7] This undertaking can be cut short by noting Putnam's observation that (3) is false. Rephrased as a counterfactual conditional, (3) states:

---

6. Ibid., 107.

7. Fodor presents a reply to this objection in his A Theory of Content and Other Essays (Cambridge, MA: MIT Press, 1991), 108-110.

(4) If cats on a dark night did not cause tokens of 'dog' then dogs would cause tokens of 'dog'.

On the possible worlds analysis of counterfactuals, we look to the closest worlds in which the antecedent is true and see whether the consequent is true in those worlds. Putnam's observation is that the closest worlds in which cats on a dark night don't cause tokens of 'dog' are ones in which 'dog' means something else, and so dogs don't cause tokens of 'dog' in those worlds either. The alternative is to suppose that the closest worlds in which cats don't cause tokens of 'dog' are worlds in which the relevant misidentifications do not occur. Putnam's point is that these worlds of perfect observers are more distant than worlds in which we speak a slightly different language. This judgment appears to be correct.[8]

Fodor starts from the reasonable observation that errors can only occur where there are semantic relations to exploit. Given that the aim of the project is to account for semantic properties in terms of the nonsemantic, Fodor cannot be content with observing that cats cause 'dog' tokens (in part) because 'dog' means dog whereas it's not the case that dogs cause 'dog' tokens because 'dog' means cat. Having rejected Dretske's attempt to account for the existence of a semantic relation owing to the temporal priority of certain correlations, Fodor tries to capture the same phenomenon in terms of counterfactuals.

Assuming that error cases are exploiting an independent semantic relation, and further assuming an (actual cause) informational semantics (condition (1)), error cases will only occur when genuine cases occur, so if there were to be no genuine cases there would be no error cases. This justifies Fodor's condition (2).

---

8. Hilary Putnam, Renewing Philosophy (Cambridge, MA: Harvard University Press, 1992).

It is also true that genuine cases are not exploiting an independent semantic relation to the error cases. In fact we are assuming that there is no semantic relation there at all for them to exploit. The apparent consequence of the causal theory that there is a semantic relation between the symbol and the error cases is the problem we are trying to solve. However, it does not follow from the lack of a semantic dependence of the genuine cases on the error cases that in counterfactual situations without the error cases the genuine cases will still occur. We would only be entitled to draw that conclusion were there to be no dependence of any kind of the genuine cases on the error cases.

Fodor seems to think that there is no such dependence: '[T]he fact that cows cause one to say 'horse' depends on the fact that horses do; but the fact that horses cause one to say 'horse' does not depend on the fact that cows do.'[9] There is some sort of dependence, however. This is because the objects that prompt the errors are such that they are likely to be mistaken for the genuine article. On the account Fodor gives in his treatment of the first problem with the crude causal theory, part of the causal chain resulting in the tokening of a mental symbol is the production of a psychophysical effect on a perceiver, which in turn results in a tokening of a symbol via an internalized theory. Cows sometimes produce sufficiently similar psychophysical effects in perceivers that the internalized theory produces the same response as it would to a clearly perceived horse. In this sense, the fact that horses cause one to say 'horse' does depend on the fact that (some) cows do, because the causal process that results in a 'horse' token is the same for (certain) cows and horses, given the possibility of perceiving cows under the right circumstances for them to produce sufficiently similar psychophysical effects to horses. Fodor's

9. Fodor, Psychosemantics, 108.

asymmetric dependence account is supposed to capture, in non-semantic terms, a plausible claim made in semantic terms, but it in fact fails to do this.

It may appear that Fodor has simply chosen the wrong counterfactual and the theory may be saved by substituting something more suitable. Putnam's point is that the worlds in which there are only perfect observers are more distant than the worlds in which our language is different. Let us suppose this is so. Why not then limit our attention to those worlds in which there are only perfect observers? One might try the following as a substitute for condition (3):

(5) If there were no misperceptions (so cats did not cause tokens of 'dog') then dogs would cause tokens of 'dog'.

The problem with this formulation is that it assumes that in the cases where cats cause 'dog' tokens there is misperception. The point of giving this set of counterfactuals is to provide a distinction between veridical and non-veridical cases and so this distinction cannot be assumed in the conditions themselves. Even if there were some way to describe the relevant worlds in non-semantic terms, some case would have to be made as to why these worlds, understood in those non-semantic terms, were relevant.

The intuition behind Putnam's objection is that worlds in which there are no mistakes in observation are very distant. Fodor has remarked that he does not put very great weight on either the presentation in terms of counterfactuals or the analysis of counterfactuals in terms of possible worlds. Instead his preferred formulation is in terms of nomic relations:

> [W]hat the story about asymmetric dependence comes down to is that "cow" means *cow* if (*i*) there is a nomic relation between the property of being a cow and the property of being a cause of "cow" tokens; and (*ii*) if there are nomic relations between other properties and the property of being a cause of

"cow" tokens, then the latter nomic relations depend asymmetrically upon the former.[10]

Fodor gets this formulation from the assumption that if the generalization that *X*s cause *Y*s supports counterfactuals then there is a law that relates the property of being *X* to the property of being a cause of *Y*s. As he puts it, 'counterfactual supporting causal generalizations are (either identical to or) backed by causal laws, and laws are relations among properties.'[11]

In a footnote, Fodor addresses what is essentially Putnam's point.[12] Fodor claims that he is not committed to the view that there are worlds close to the actual world in which mistakes in observation could not be made: 'In fact, what I hold is only that if "cow" means *cow* and not *horse* then it must be nomologically possible to tell any cow from a horse; which doesn't sound all *that* wild after all.'[13] The element of the theory that both Putnam and Wagner have criticized results, says Fodor, only if one adds to his claims about nomological relations the assumption that if *P* is nomologically possible, then there is a (possible) world in which it is the case that *P*. It must be assumed that Fodor rejects this assumption.

It must be agreed that if Fodor is committed only to what he claims here, then his commitments are not at all wild. The question is: How are we now to understand his theory?

On ontological grounds, Fodor thinks that truths about nomic relations among properties are not to be analyzed in terms of counterfactuals:

---

10. Fodor, A Theory of Content, 93.

11. Ibid.

12. Ibid., 95, footnote 10. The objection is attributed to Steven Wagner, 'Theories of Mental Representation' (unpublished). Wagner objects that Fodor's theory has 'the *wildly* implausible consequence that there are worlds remotely like ours in which cows could not be mistaken for horses.'

13. Ibid., 133.

> I suspect, in particular, that some of the troubles we're about to survey stem not from there being anything wrong with the proposal that content rests on asymmetrical dependences among nomological relations, but rather from there being everything wrong with the assumption that claims about nomological relations need counterfactual/possible world translations.[14]

He uses the gas laws as an example. These laws make such assumptions as that molecules are perfectly elastic and containers infinitely impermeable. As Fodor points out, the satisfaction of these conditions is physically impossible and so it is hard to say what would follow in the counterfactual circumstance that they were true. Despite our ignorance concerning what would be the case were there to be ideal gasses, the gas laws are perfectly respectable. According to Fodor, it is the laws themselves which tell us what counterfactuals are true: '[I]f there were ideal gasses, then, ceteris paribus, their volume would vary inversely with the pressure upon them. And that counterfactual the theory itself tells us is true.'[15] In a similar vein he says of his theory:

> We can know that there are asymmetric dependences among nomic relations between properties without knowing much about which counterfactuals these asymmetric dependences make true. All we need to know is that if the nomic relation between P1 and P2 is asymmetrically dependent on the nomic relation between P3 and P4, then, ceteris paribus, breaking the relation between P3 and P4 would break the relation between P1 and P2. And that counterfactual the theory itself tells us is true.[16]

These remarks are puzzling, for Fodor's assertion that two nomic relations are asymmetrically dependent has been understood in terms of certain

---

14. Ibid., 95.
15. Ibid.
16. Ibid.

counterfactuals being true (which according to the objection are false). Fodor must provide some way of understanding the asymmetric dependence of one nomic relation upon another without using counterfactuals. The problem he faces is that this non-counterfactual expression of asymmetric dependence does not imply the relevant counterfactuals, contrary to what he suggests.

What Fodor seems to have in mind is that because the generalization that dogs cause 'dog' tokens supports counterfactuals, it's a law that whatever has the property of being a dog also has the property of being a cause of 'dog' tokens (in short, that it's a law that dogs cause 'dog' tokens). Other properties are nomically related to the property of being a cause of 'dog' tokens in this way (see the discussion of psychophysical properties above). It is also a law, let us suppose, that whatever has the property of being a cat on a dark night also has the property of being a cause of 'dog' tokens. Fodor's theory is that this latter nomic relation only exists because the former one does.

According to Fodor, the claim that cats cause 'dog' tokens is asymmetrically dependent on the law that dogs cause 'dog' tokens is to be understood as the claim that it is a law that cats don't cause 'dog' tokens unless dogs do, and there is no law that dogs don't cause 'dog' tokens unless cats do. In other terms:

[not (cats cause 'dog' tokens) unless (dogs cause 'dog' tokens)] is nomologically necessary,

and

[not (dogs cause 'dog' tokens) unless (cats cause 'dog' tokens)] is not nomologically necessary.[17]

Equivalently:

[if not (dogs cause 'dog' tokens) then not (cats cause 'dog' tokens)] is a law

---

17. Fodor, A Theory of Content, 113; idem, Psychosemantics, 109-110.

[if not (cats cause 'dog' tokens) then not (dogs cause 'dog' tokens)] is not a law.

Ideally, this theory is supposed to tell us that the following counterfactuals are true:

(4) If cats on a dark night did not cause tokens of 'dog' then dogs would cause tokens of 'dog'.

(6) If dogs did not cause tokens of 'dog' then cats on a dark night would not cause tokens of 'dog'.

(6) is the counterfactual version of Fodor's condition (2). There is widespread agreement that this counterfactual is true, and Fodor's asymmetric dependence claim implies its truth, for if it is a law that cats don't cause 'dog' tokens unless dogs do, were dogs not to cause 'dog' tokens, cats would not do so either. As it is generally agreed that possible worlds in which laws in the actual world hold are closer than worlds in which they do not, on Fodor's theory the closest possible worlds in which dogs don't cause 'dog' tokens are also worlds in which cats don't.

That Fodor's theory implies (4) is less certain. It does not follow from the fact that it is not a law that dogs don't cause 'dog' tokens unless cats do, that if cats on a dark night were not to cause tokens of 'dog' then dogs would still cause tokens of 'dog'. It is not a law, for example, that if George W. Bush does not win the 2000 U.S. Presidential election that Ralph Nader does not win. In other words, it is nomologically possible that Bush loses and Nader wins. It manifestly does not follow from this that if Bush were to have lost the election then Nader would have won. Given the facts about the support enjoyed by the various candidates in the race, rather than the closest possible worlds in which Bush loses being worlds in which Nader wins, instead, the closest possible worlds in which Bush loses are worlds in which Nader also loses.

In saying that it is not a law that dogs don't cause 'dog' tokens unless cats do, Fodor is saying no more than that it is nomologically possible that both dogs cause 'dog' tokens and cats do not. This is clearly insufficient to support the counterfactual (4). A nomic relation between the antecedent of a counterfactual conditional and its consequent may provide evidence for judging its truth-value, but the absence of one surely does not. Neither does the nomological possibility that dogs cause 'dog' tokens and cats do not permit one to infer that the closest possible worlds in which cats don't cause 'dog' tokens are worlds in which dogs do. Putnam's objection, recall, is that although it is nomologically possible that dogs cause 'dog' tokens and cats do not, these worlds are more distant than those in which neither cats nor dogs cause 'dog' tokens.

Fodor suggests we can understand his theory perfectly well without worrying what would really happen were the antecedents of the counterfactuals it implies to be satisfied. So far as the gas laws are concerned, for example:

> God only knows what would happen if molecules and containers actually met the conditions specified by the ideal gas laws (molecules are perfectly elastic; containers are infinitely permeable; etc.); for all *I* know, if any of these things were true, the world would come to an end. After all, the satisfaction of these conditions is, presumably, *physically* impossible, and who knows what would happen in physically impossible worlds?
>
> But it's not required, in order that the ideal gas laws should be in scientific good repute, that we know anything like all of what would happen if there really were ideal gasses. All that's required is that we know (e.g.) that if there were ideal gasses, then ceteris paribus, their volume would vary

> inversely with the pressure upon them. And *that* counterfactual *the theory itself tells us is true.*[18]

Fodor goes on to draw an analogy between the gas laws and his own theory by saying that 'All we need to know' (presumably for the success of the theory) is that if one nomic relation, *A*, is asymmetrically dependent on another, *B*, then breaking relation *B* would break relation *A*. As Fodor notes, that counterfactual is implied by his theory. Yet if the dependence is asymmetric, we must also know that breaking the relation *A* would not break the relation *B*. On Fodor's theory, there is no law that says that when relation *A* is broken, *B* is broken, but there is no law that says that it isn't, either.

To return to the analogy with the 2000 U.S. Presidential election, as a matter of fact Bush won the election and Gore lost. It is (was) not a law, however, that if Bush does not win the election, then Gore does not lose it (i.e. that if Bush loses, Gore wins), for there were other candidates in the race and it was not nomologically impossible (although it probably was psephologically impossible) for one of them to have won. The actual facts and the absence of a law connecting a Bush defeat to a Gore victory does not imply that if Bush were to have lost the election then Gore would still have lost. In all probability, were Bush to have lost, Gore would not still have lost. By analogy, the fact that both dogs and cats actually cause tokens of 'dog' and the absence of a law that connects the failure of cats to cause 'dog' tokens with the failure of dogs to cause 'dog' tokens does not imply the counterfactual (4).

Fodor often shows reluctance to accept the analysis of counterfactual conditionals in terms of possible worlds. In so far as this analysis does not support the counterfactuals he uses in formulating his theory, his reluctance is justified. As

---

18. Fodor, A Theory of Content, 94-95.

the nomic relations specified by the theory leave open the truth-value of one of the counterfactuals he asserts in articulating the theory, Fodor should go further and resist the temptation to explain asymmetric dependence in terms of counterfactuals at all. His theory provides no reason for supposing that the relevant counterfactuals are true.

This is not to say that Fodor's theory is wrong, just that it should not be presented in terms of counterfactuals. This leaves Fodor's condition (2) in the example above being replaced by the claim:

(7) it is nomologically impossible that cats cause 'dog' tokens and dogs don't.

Condition (3) is replaced by the claim:

(8) it is nomologically possible that dogs cause 'dog' tokens and cats do not.

Although it was condition (3) that caused the problems when the theory was stated as counterfactuals, when the theory takes the form of statements of nomological possibility, it is the analogue of condition (2), (7), that is suspect. Few would deny (8), but (7) is in need of defense.

Ned Block argues that (7) is only plausible owing to an equivocation. If what is meant by "'dog'" is the phonological/orthographic sequence #d^o^g# then there is no reason at all to believe (7). Amongst the possible worlds where 'dog' refers to cats will be a world in which cats cause 'dog' tokens and dogs don't. If (8) is true, symmetry suggests that this world is nomologically possible. Some reason, at least, should be given for supposing that this is not the case. The claim that as a matter of nomologically necessity, tokens of #d^o^g# are caused by dogs or nothing at all is at best unsupported and at worst simply false.

An alternative is to propose that it is nomologically necessary that either #d^o^g# tokens are caused by dogs, or if something else causes #d^o^g# tokens then #d^o^g# tokens refer to something other than dogs (i.e. if tokens of #d^o^g#

are not caused by dogs then either nothing causes #d^o^g# tokens or they refer to something else). This is clearly unacceptable for the project of providing a naturalistic semantic theory, for the theory appeals to semantic notions. Neither is it acceptable for the purposes of this project, for the same reason, to insist that tokens of #d^o^g# are not tokens of the same type as the word 'dog', owing to its status as a word. So, Block concludes, the asymmetric dependence theory is either false or circular.[19]

In reply to this argument Fodor invokes a *ceteris paribus* clause. His position is that 'the asymmetric dependence proposal is that *all else being equal*, breaking cow → "cow" breaks X → "cow" for all X.'[20] So he is claiming that the law that cats don't cause 'dog' tokens unless dogs do, is only a law *ceteris paribus*. According to Fodor, in the possible worlds imagined by Block, all else isn't equal. The example that Fodor reports Block as using is the possible world in which 'cow' refers to trees. In such a world, not only does one suppose that the causal connection between cows and 'cow' tokens is broken but one additionally supposes that a causal connection between trees and 'cows' is in force. It is this additional supposition which Fodor holds to be in violation of his *ceteris paribus* clause.

Fodor grants that there are nomologically possible worlds where there are no misperceptions. So, for example, there are possible worlds where only dogs cause 'dog' tokens and cats never do. The asymmetrical claim is that there are no nomologically possible worlds where no dogs cause 'dog' tokens and only, say, cats do. As Block points out, purely as a matter of nomological possibility there are such worlds. Fodor intends to rule these worlds out of consideration on the

---

19. Fodor reports this argument from a conversation with Block in ibid., 111-112.

20. Ibid., 112. Naturally, the reference to cows is merely an application of the proposal to a particular case.

basis of his *ceteris paribus* clause on the grounds that in these worlds one is not merely supposing that the dog → 'dog' connection is broken, but in addition that there is some other causal connection in place. I believe this is an illegitimate use of the *ceteris paribus* clause. The basic reason for this is that just as there are nomologically possible worlds where there are no misperceptions, there are nomologically possible worlds where there are only 'misperceptions'.[21] This does not suppose that any new, previously unspecified, causal connections are assumed.

As I presented Block's objection above, the counterexample world is one in which it is cats, not trees, that cause 'dog' tokens while dogs do not. The difference is that, it has been assumed, cats (under certain not too unusual circumstances) are reliable causes of 'dog' tokens in the actual world. One does not have to introduce causal chains that do not exist in the actual world in order to provide counterexamples to Fodor's claim that if a term *X* refers to things of type *Y* then, *ceteris paribus*, it is nomologically impossible for tokens of *X* to be caused by only non-*Y*s. As tokens of *X* are caused by some non-*Y* things, the *Z*s, in the actual world, one need only consider a world in which *X* is caused solely by *Z*s for a counterexample.

In my presentation of Block's argument, I suggested as a counterexample to Fodor's claim of nomic impossibility that amongst the possible worlds where 'dog' refers to cats would be a world in which only cats cause tokens of 'dog' and no dogs do. It would be open to Fodor to object that this counterexample not only supposes that the dog → 'dog' causal connection is absent but further supposes an additional semantic relation between cats and 'dog' which accounts for the

---

21. It may be that in such a world the 'misperceptions' would actually be veridical, if they possessed any content at all. As Fodor's theory must be stated in nonsemantic terms, however, whatever facts are deemed to be in violation of the *ceteris paribus* clause cannot be described in semantic terms.

presence of the cat → 'dog' causal connection. This element of the counterexample is unnecessary and may be dispensed with. There is no need to specially account for the cat → 'dog' causal connection, as it already exists in the actual world.

Fodor's main point, of course, is that the cat → 'dog' causal connection is dependent upon the dog → 'dog' causal connection, and so breaking the latter connection would also break the former. So, one can understand why Fodor would demand some special account of how there gets to be a cat → 'dog' causal connection but no dog → 'dog' causal connection in some possible world. This demand need not be met. Fodor is correct to say that if dogs did not cause tokens of 'dog' then cats would not either. In terms of possibilia, the closest possible worlds where dogs do not cause 'dog' tokens, are worlds in which cats do not either. This is because in those worlds, 'dog' does not mean dog. So a tendency to occasionally confuse cats with dogs does not lead to the tokening of 'dog' in the presence of cats in those worlds. Although Fodor is entitled to infer the truth of counterfactual (6) from his claim that it is nomically impossible that cats cause 'dog' tokens while dogs do not, he is not entitled to infer the nomic claim, even *ceteris paribus*, from the truth of the counterfactual. If all that Fodor has to support the assertion that it is nomically impossible, *ceteris paribus*, that cats cause tokens of 'dog' but dogs do not, is that counterfactual (6) is true, then to genuinely have an asymmetry, he would have to be able to show that counterfactual (4) is true. I have argued that he is unable to do this.

Lynne Rudder Baker presents an objection to the asymmetric dependence theory which I believe is rather illuminating as to the reason for its failure. She considers an example where someone has come into contact with robot-cats only,

although there are many real cats in the local environment.[22] When the person encounters a real cat for the first time, the cat causes in him a token of the same type of mental state as the robot-cats do. Baker suggests three possibilities for interpreting the cat-caused token:

(a) The token refers to cats only. Baker rejects this because the disposition to apply 'cat' to cats asymmetrically depends on the disposition to apply it to robot-cats, and so according to Fodor's conditions, 'cat' does not refer to cats.

(b) The token refers to robot-cats only. Baker rejects this because it is an accident that the person learned the token from only robot-cats. If he had encountered any real cats, they would have caused a token of 'cat'. So the counterfactual supporting correlation is between the token and cats or robot-cats. Fodor endorses this argument (and makes a similar point against Dretske's 'learning-period' solution to the disjunction problem).

(c) The token refers to both robot-cats and cats. Baker rejects this position because this does not provide a solution to the disjunction problem. Fodor actually endorses this position, but maintains that the disjunction problem still admits his solution.[23] Fodor says that it is consistent to hold that 'cat' refers to both robot-cats and cats when it's just an accident that one learned it from robots (as in Baker's case), while maintaining that it would not be disjunctive where one could learn it only from robots, or from cats, but not both. Fodor reports Dretske as giving a similar verdict on an example of his where there are both H_2O and XYZ on Twin Earth, but a local speaker has learned 'water' from H_2O only. Dretske gives the disjunctive meaning in this case.

---

22. Lynne Rudder Baker, 'On A Causal Theory of Content,' Philosophical Perspectives 3 (1989): 165-186.

23. Fodor, A Theory of Content, 104.

This is less of a concession in Dretske's case than it is in Fodor's. On Dretske's account, the information that a state carries at the end of the learning period is fixed as its semantic content. In an environment that contains both XYZ and H_2O, a state that is caused by H_2O will also be such that it would be caused by XYZ, owing to the indistinguishability of H_2O and XYZ. This being the case, the presence of this state actually carries the information that either H_2O or XYZ is present. When the environment contains only XYZ, there is no possibility that there is H_2O present and so the state would carry only the information that XYZ is present. The 'intuition' that Fodor reports Dretske as sharing with him is fully supported by Dretske's theory.

Fodor, on the other hand, appears to be effectively adding a new condition to his theory in making this reply to Baker. Assuming that in her example the correlation between 'cat' and cats is asymmetrically dependent on the correlation between 'cat' and robot-cats, Fodor is conceding that an asymmetric dependence is not sufficient for content. This would be significant if true, for Baker argues that the features of her unusual case can be generalized to our ordinary circumstances:

> [S]uppose that Sally has ... seen 1,001 mules, each of which has produced in her an M-token; then one day, for the first time, she sees a horse, which also produces in Sally an M-token. By parity of reasoning, we should say that Sally's first M-token represents mule-or-horse. The same story could be told about (almost?) any symbol.[24]

In order to avoid this conclusion, Fodor must indicate some relevant difference in the cat/robot-cat case to make it go disjunctive. That it is much harder to tell the

---

24. Lynne Rudder Baker, 'Has Content Been Naturalized?' In Meaning in Mind: Fodor and His Critics, edited by Barry Loewer and Georges Rey (Oxford: Blackwell, 1991).

difference between cats and robot-cats than it is to tell the difference between horses and mules would not be sufficient.

What Fodor might appeal to in reply to this objection is made clearer in the light of his discussion of Twin Earth. Fodor has two ways of dealing with the Twin Earth example. One way depends on the 'actual causation' assumption that nothing can be in the extension of a symbol unless it has caused tokens of the symbol. The other way is a defense of the pure informational theory, which has no actual causation requirement. The defense depends on 'water' being a kind term. This being the case, says Fodor, speakers intend to apply the term only to things which are the same kind as the local samples. Fodor's claim is that this intention has the effect of making the application of 'water' to XYZ asymmetrically dependent on its application to H_2O. Given this intention, speakers only apply 'water' to XYZ when they are unable to distinguish it from H_2O. Were speakers able to distinguish between XYZ and H_2O they would apply 'water' to H_2O and not to XYZ. As a consequence, Fodor says:

> [T]here are nearby worlds where you get the H_2O/"water" connection without the XYZ/"water" connection, but no nearby worlds where you get the XYZ/"water"connection without the H_2O/"water" connection. I.e. it's nomologically possible for the XYZ/water [sic] connection to fail without the H_2O/water [sic] connection failing, but not vice versa.[25]

Having shown that the intention of speakers to use 'water' as a kind term is sufficient to establish an asymmetric dependence of the XYZ/'water' correlation on the H_2O /'water' correlation, Fodor is able to make the required distinctions to save his theory from Baker's objection. In the example where a person learns to apply 'cat' to robot-cats only, the term 'cat' refers to both cats and robot-cats so

---

25. Fodor, A Theory of Content, 116.

long as the person has no intention of using 'cat' as a kind term for only cats or robot-cats. Were the person to have such an intention, this would establish an asymmetric dependence of the one correlation upon the other. The explanation of the difference between Fodor's response to this example and his likely wish to say that in Baker's other example that Sally's M-token refers to mules and not horses is that in the latter example, Sally has an intention of using M-token as a kind term.

To note that speakers sometimes have intentions to use certain terms for kinds is not on its own sufficient for Fodor's purposes. It is important what kinds they intend to use them for. There is nothing to stop someone from having an intention to refer to a kind that includes both H_2O and XYZ, or both mules and horses. So Fodor is effectively supposing that the explanation of why 'water' refers to H_2O and not H_2O or XYZ, and, closer to home, why 'mule' refers to mules and not mules or horses, is that we intend 'water' to refer only to H_2O and 'mule' to refer only to mules.

Stated so openly, this supposition makes the theory appear at best trivially true and at worst circular. This is no more than mere appearance, however. The basis of the theory is that the asymmetric dependence of one nomic relation upon another is a consequence of the semantic properties of a symbol. When a state *S* refers only to objects of type *A*, *A*s (under appropriate circumstances) will elicit tokens of S, and any non-*A*s that cause tokens of S, the *B*s, only cause tokens of *S* because they are confused with *A*s. This being the case, if the language changed so that the semantic connection between *S* and *A*s was broken, the connection between *S* and *B*s would also break. Yet breaking the connection between *S* and *B*s (by being very, very careful about what *S* is applied to) would not break the connection between *S* and *A*s. If *S* refers to *A*s, therefore, the correlation between *B*s and *S* will asymmetrically depend upon the correlation between *A* and *S*. If there is such a thing as reference, there will be the asymmetric dependencies that

Fodor's theory says there are. In proposing that the existence of an asymmetric dependence is sufficient for reference, Fodor is assuming that there are no such dependencies in the absence of a referential connection. This assumption is not obviously true, but the asymmetric dependence engendered by reference is striking enough that the theory is worth taking seriously. Stated in semantic terms, Fodor's theory is acceptable. Of course, as the purpose of the project is to give a naturalistic theory of content, he is not entitled to any of the semantic assumptions. Unfortunately, these are needed to make the theory work.

With respect to Fodor's third condition, if speakers were to break the connection between *S* and *B*s by applying *S* only to things that it is true of, the connection between *S* and *A*s would not thereby be broken (although fewer *A*s would cause *S* owing to uncertainty as to whether they were really *B*s). This does not mean that if *B*s did not cause *S* then *A*s still would. In terms of possibilia, the set of worlds in which *B*s do not cause *S* is the superset of the set of worlds in which *B*s do not cause *S* owing to people being maximally careful to apply *S* only to things it is true of. It is possible to argue, as Putnam does, that the worlds in the subset are more distant from the actual world than some of the other worlds in the superset.

With respect to Fodor's second condition (stated in terms of nomological possibility), although if *A*s did not cause *S* (owing to *S* no longer referring to *A*s) then *B*s would not cause *S*, it is not the case that it is nomologically impossible for *B*s to cause *S* and *A*s not to cause *S*. Although *B*s actually cause *S* because they are confused with *A*s, this does not make it a law that if *B*s cause *S* then *A*s do. As the property of being an *A* and the property of being a *B* are distinct properties, it is possible for any one of them to be nomically related to the property of being a cause of *S* without the other so being.

There is one way to state the asymmetric dependence without any semantic assumptions. This is by formulating an adequate naturalistic theory of content, then replacing the semantic assumptions behind asymmetric dependence with the corresponding naturalistic facts. I can see no reason to believe that asymmetric dependence can be used as the basis of a naturalistic theory of content, rather than being explained by such a theory.

CHAPTER 5
FINAL CAUSES AND CONTENT

The outstanding problem for theories of intentionality based on the insights of the causal theory of reference is the problem of error, or the disjunction problem. The problem is that the content of a particular state cannot be thought of as simply what it is caused by, as some tokens of representative states are non-veridical. One of the currently more popular solutions to this problem is to hold that the representational properties of some mental state tokens are derived from functions, either that of the tokens themselves or the representing mechanisms of which they are states. It is supposed that this would provide a solution to the problem of error because, for example, if a state's function or purpose is to indicate that a dog is present but is sometimes caused by the presence of a cat, then the cat-caused tokens may be dismissed as malfunctions irrelevant to the semantic properties of the state.

There are general considerations that have been taken to lend credibility to this sort of view as a theory of the basis of content. Dennett, for example, invites us to consider the following thought experiment: Suppose that you wish to live to see the twenty-fifth century, and the only known way of doing this is to have your body frozen and thawed out several centuries hence. Naturally, you will be concerned to keep your body safe and supplied with sufficient energy to keep it frozen for four hundred years. One good way to do this is to build a robotic capsule sophisticated enough to find the resources your body needs. This device

must be able to identify places of safety, energy sources and potential dangers and seek out and avoid them as necessary.[1]

Dennett's suggestion is that if one is inclined to regard such a survival machine as having no more than derived, rather than original, intentionality, then one should see oneself in the same light. His point is that we are really no more than survival machines 'designed' by a process of evolution 'for the purpose of preserving your genes until they can replicate'. He concludes '[O]ur intentionality is derived from the intentionality of our "selfish genes"! *They* are the unmeant meaners, not us!'[2]

I take it that Dennett does not literally mean that genes possess intentionality. Rather, he means that it is our genes, and more generally the process of evolution which produces design-like phenomena in nature, that are the source of the phenomenon of intentionality. The phrase 'unmeant meaners' is a reference to the doctrine that some things, such as ourselves, have 'original' or genuine intentionality. On this view, other things, such as artifacts, only have intentional properties derivatively, owing to beings with original intentionality bestowing these properties upon them. Dretske, for example, puts forward a view of this type when he distinguishes type I and II representing systems, which represent only what they have been designated as representing. It is obvious that some representations only have that status owing to the way they are treated by beings that already have intentionality. The markings on this page, for example, only mean what they do owing to our using them to mean that.

It is not credible to suppose that all intentionality is derived in this way. Not only would this give rise to an infinite regress (of the sort that led to the posit of

---

1. Daniel Dennett, The Intentional Stance (Cambridge, MA: MIT Press, 1987), 295-98.

2. Ibid., 298.

unmoved movers), but there don't seem to be any intentional agents around to provide our meanings for us. The doctrine of original intentionality, as Dennett characterizes it, goes further and declares that the intentionality of some beings, i.e. us, is 'utterly underived'[3]. I would not say that all the philosophers to whom Dennett attributes this view really believe this. Many, such as Dretske and Fodor, believe that our intentionality is to be accounted for naturalistically, and so in that sense is 'derived' from non-intentional processes.

Even so, these philosophers make a distinction between the basis of our intentionality and that of simpler systems. They are inclined to say of an artifact that its apparently contentful states are so only derivatively owing to its being designed by beings who have full intentionality. Dennett replies to this that our apparently contentful states only have the properties that they do owing to the design-like work of a process of evolution: 'It follows from the truth of Darwinism that you and I are Mother Nature's artifacts, but our intentionality is none the less real for being an effect of millions of years of mindless, algorithmic R and D instead of a gift from on high.'[4] So, concludes Dennett, if we persist in denying artifacts genuine intentionality, we must also deny it of ourselves.

It is true that we know that beings capable of entertaining mental states with content exist, and that these beings have been produced via a process of evolution. It does not follow from this that content itself is a matter of function, or even if it is that these functions are to be accounted for in terms of natural selection. There is something to what Dennett says, in that given the assumptions of naturalism, we should expect ourselves and some suitably sophisticated artifact to be in the same boat so far as our having genuinely contentful mental states goes. However, even

---

3. Ibid., 288.

4. Daniel Dennett, Darwin's Dangerous Idea (New York: Simon and Schuster, 1995), 426-7.

some of the philosophers he condemns as believers in 'original' intentionality would accept this much. This point does not go any way towards showing how intentionality is realized in such a device, or in ourselves.

Some general criticisms of teleological explanations are presented by Gould and Lewontin in their 'The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme.'[5] One of the criticisms of adaptationist reasoning presented in that paper is the assumption that all biological phenomena occur because of some beneficial consequence to the relevant organism. Another is that the apparent explanatory power of teleological explanations rests simply on their being good stories, 'just so stories', as Gould and Lewontin put it.

A spandrel, as Gould and Lewontin describe it, is a triangular area formed by mounting a dome on a base of archways which meet at right angles. These areas, such as the ones in St. Mark's in Venice, are often highly decorated. Gould and Lewontin say that although it appears that the purpose of spandrels it to display decorations, this is not the reason why they are to be found in churches. Rather, they are an architectural consequence of building a dome over arches. While it is possible to dispute the example, noting that the spandrels of St. Mark's have been shaped specifically for the display of mosaics, the basic idea is sound enough. It is a mistake to believe that everything exists because of its effects. Some things exist because they are effects.

Some texts on labor, for example, begin with a discussion of why the process is so painful in human females, particularly in comparison with other mammals. One would expect rational creatures, *ceteris paribus*, to avoid painful experiences

---

5. Gould and Lewontin, 'The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptionist Programme.' Proceedings of the Royal Society, Vol. B205, 581-98.

and so this feature of human reproduction seems to be rather unadaptive. It is possible to form some adaptationist speculative hypothesis (and some authors do), such as that the experience helps to produce some bond between mother and child. However, The phenomenon can be well enough explained without supposing that it is adaptive. Given the adaptiveness of the basic human body shape, particularly that it is bipedal, and the state of development of human infants at birth, some pain is going to be involved in getting a fairly large child through a fairly small pelvic space. One need not look for some special explanation of the existence of a feature so long as it can be explained as a consequence of other features, and is not so unadaptive as to significantly endanger the survival of the organism. This, I take it, is the essential point of Gould and Lewontin's spandrel example. This is not an objection to adaptationist thinking in general, just the assumption that every feature of an organism is an adaptation and may be given a teleological explanation.

Still, given that a process of evolution takes place, some features are to be explained by their adaptiveness. For an example of such an explanation, consider the aquatic ape theory of human development. This theory holds that certain non-apelike features of human beings, such as a relative lack of body hair, distribution of body fat and bipedal stature developed from a water-dwelling species intermediate between humans and apes. This theory cannot be dismissed simply out of hand, for it is known that aquatic mammals such as whales and dolphins did evolve out of land-based mammals. The aquatic ape theory is that we evolved out of a species that returned to the ocean and then, for some reason, returned once again to the land.

This is certainly a fascinating story, but that has no bearing on the likelihood of its truth. Teleological theorists look at the effects traits have on the organism which possesses them and seek to explain their existence on the basis of these

effects. The assessment of adaptationist hypotheses presents a major problem, for the best confirmation of a theory is its prediction of new phenomena, not merely its fitting the known phenomena. Not only are such predictions hard to come by for adaptationist hypotheses, the known phenomena themselves may be rather thin. The aquatic ape theory begins simply with the observation of certain human characteristics and notes that they would be adaptive for a swimming creature. This is less satisfactory than fitting some direct evidence of the existence of such a hybrid creature, such as a fossil record.

The teleologist may respond to these objections by noting that, as evolution does take place, some adaptationist story will be true of most of the complex features of biological creatures. This is so, but it is of no help in deciding which particular adaptationist story is true. Fortunately for teleological theories of content, this sort of objection is not directly damaging. As a philosophical theory, the teleological approach is intended to provide an explanation of how intentionality is possible. This enterprise may be illustrated with examples, such as the frog's eye telling the frog's brain the location of flies, but it does not matter for the success of the theory whether any particular such example is actually correct. What matters is that the theory shows how intentionality can be accounted for in teleological terms. A psychological theory that made use of intentional notions would be vulnerable to these sorts of objections, for then a claim would be being made about the content of some specific state based on its evolutionary history. Even if any particular teleological explanation were too weak and untested to be accepted, this would not defeat the project of providing a teleological theory of intentionality. This sort of philosophical theory does not depend on making specific claims about the content of actual states, only that for some states, or mechanisms of which representations are states of, some adaptationist story is true of them. The aim of this theory would be to show how states with contents could

have arisen, not to describe their selectional history. If the teleological theory is successful at showing how intentionality is possible, in the absence of any alternative it should be the preferred theory.

For the teleological proposal to constitute a solution to the problem of error, some satisfactory account of function needs to be presented. McGinn, for example, in his Mental Content describes a state's function as 'what it is designed to do, what it is supposed to do, what it ought to do'. A considerably less vague account of function is required. There are basically three types of theories of function that have been advanced. The first type makes function a matter of the capacities and organization of the system. This will not do for the purpose in hand, for we need an account that applies to a system which is failing to perform its function, or performing it less than adequately. The second type of theory accounts for functions in terms of effects which are, in some sense, good for a system. Woodfield, in his Teleology, defends such a view, where goods are understood in terms of contributions towards achieving the goals of beings. This sort of theory is unsuitable for the naturalist who wishes to give an account of what it is for an organism to have goals in terms of certain of its states having functions. The third type largely derives from Larry Wright's work on teleology where a function of a thing is something it does which explains its existence.

The most promising account of functions for those who wish to give a teleological theory of content is that based on Wright's account. In his Teleological Explanations, Wright suggests that the basis of opposition to teleology is often the uses to which it has been put, such as arguments for vitalism, panpsychism and a divine creator, rather than any problems with teleological notions in themselves. The main objections to teleological explanations as such, rather than their applications, he sees as being the following:

1. A cause cannot occur later than its effect, but teleological explanations explain events in terms of subsequent events, hence they cannot be causal explanations.

2. Teleological explanations illegitimately ascribe human mental characteristics to things other than human beings.

3. Teleological explanations are inimical to scientific research.

All of these objections are made in the following passage, where the authors have just given a chemical explanation for plant behavior:

> Contrast this *cause and effect* explanation with some of the *teleological* explanations of stomatal movement sometimes seen in print: that stomata open in the morning "so that the plant can secure carbon dioxide for photosynthesis" and that stomata close at night "in order to save water." Such teleological explanations, crediting the plant with intelligent and purposeful behavior, are easy to formulate but totally inadequate in explaining plant responses. Teleological explanations get the cart before the horse by converting a possible result into a cause. If botanists were satisfied with teleological explanations for plant behavior, research aimed at the discovery of the actual course of events would cease.[6]

Wright proposes to meet these objections in the following ways:

1. Wright observes of the charge of reversing the order of cause and effect that 'it is not clear how this view could have survived even modest scrutiny'. There is nothing, he says, in the ascription of goals, purposes or functions that requires us to reverse the order of cause and effect. Suppose one teleologically explains my going to the shop by citing my desire to purchase some bread. It is the desire that

---

6. Victor A. Greulach and J. Edison Adams, Plants: An Introduction to Modern Botany, 261. Quoted in Larry Wright, Teleological Explanations: An Etiological Analysis of Goals and Functions (Berkeley: University of California Press, 1976), 8-9.

causes me to go, not the subsequent act of purchase. The purchase is the goal of the action, not its cause. Wright wishes to make the same point with respect to functional explanations. In the botanical example, he claims that the functional explanation is not incompatible with the chemical explanation. When one says that the stomata close in order to save water, one is saying that saving water is their function, not that the saving of water resulting from the closing is the cause of the closing.

2. Wright takes the paradigm case of teleological explanation to be human purposive behavior. The objection is that teleological explanations of non-purposive phenomena are therefore illegitimate. Wright's claim is that teleological explanations are to be understood as an extension of intentional explanation, but only metaphorically. The feature of the human cases that transfers to the non-human cases, he says, is "when we say 'A in order that B,' the relationship between A and B plays a role in bringing about A." The value of the metaphor is bringing out this aspect, rather than any conscious purpose. To understand the teleological language as literally ascribing a conscious purpose, he says, is so bizarre in most cases, that no one is misled, just as no one is misled in other uses of metaphors, as when one says that a certain individual is a square peg in a round hole.

3. Wright says that if teleological explanations prove unsuitable for scientific purposes even though they have been shown to be free from logical or philosophical difficulties, then this would mean no more than that they have a limited application. More seriously, he believes this objection to be a result only of a misunderstanding of teleology, and so also believes that it may be dissolved by giving an adequate account of the notion. In order to deal with this objection Wright must not only show that a teleological explanation does not contradict an independent causal explanation but must also show what it is that the teleological explanation is explaining and in what such an explanation consists. An assessment

of his success in achieving these tasks can only be made on the basis of his analysis.

Wright's Analysis

Wright's work is based on the analysis of Charles Taylor[7]. On Taylor's account, behavior is teleological if it occurs because it is required for some end. The basic idea is that a teleological answer to the question 'why did he do that?', is of the form 'he had to do it'. This is clearly too strong. In most cases the answer 'he thought he had to do it' would be more accurate, although still not quite right. Wright also recognizes that Taylor's simple formulation is too strong, noting that nothing a predator does in stalking its prey is strictly required in order to obtain food. The predator could find its prey already dead, for example. So Wright weakens the analysis: teleological behavior is not that which occurs because it is required for some end, but that which occurs because it will in fact achieve the end. Recognizing that not all purposive behavior will achieve its end (for someone can try to do something and fail), Wright finally, modifies the analysis to the claim that behavior is teleological when it occurs because it tends to bring about certain results:

S does *B* for the sake of *G* iff:

(i) *B* tends to bring about *G*.

(ii) *B* occurs because (i.e. is brought about by the fact that) it tends to bring about *G*.

Wright calls these conditions (T). (T) is actually more vague than it may at first sight appear. Wright tells us that "teleological behavior is behavior with a consequence etiology; and behavior with a consequence etiology is behavior that

---

7. Charles Taylor, The Explanation of Behavior (London: Routledge and Kegan Paul, 1964).

occurs because it brings about, is the type of thing that brings about, tends to bring about, is required to bring about, or is in some other way appropriate for bringing about some specific goal."[8] In (T), 'tends to' is to be understood to include all the above mentioned consequence etiologies. Wright suggests that the way to understand the general claim that a given behavior is appropriate for bringing about a goal, is that it is the right sort of thing to try to do in the circumstances. This formulation is naturally understood as the claim that it is the sort of thing an intentional agent would try to do in the circumstances.

Behavior may be for the sake of a given end, even though it isn't even nomologically possible for it to achieve that end. The real key to goal-directed behavior in the conscious case is that it is done because some agent thought that it would, or at least might, bring about a certain end. It does not seem that this can even be reduced to it being in some sense rational to believe that it might bring about the end, for the desperate may attempt the irrational (as presumably the irrational do). In my view, the prospects for an account of goal-directed behavior which captures the conscious cases as well as those of any possibly non-intentional systems are dim, as to capture the latter it would have to ignore that which is central to the former. We may hope for a comprehensive naturalistic account of goal-directed behavior, but I expect it will be available only subsequent to a naturalistic account of intentionality.

(T) is an account of goal-directed behavior. This is, of course, not the same as an account of function, for not everything that has a function exhibits behavior. Chairs, for example, have the function of supporting sitting individuals, but they don't do anything. The most basic class of functions are those 'underwritten by the presence of human intent' or as Wright calls them, the 'conscious functions'.

---

8. Wright, Teleological Explanations, 38-39.

Things with such functions include artifacts designed for a specific purpose, and also things that are pressed into service for tasks other than that for which they were designed (if they were designed at all). For example, a rolled up newspaper stuffed under a door may be said to have the function of blocking a draft.[9] If the newspaper was not put under the door intentionally then it could not be said to have the function of blocking the draft, even though it would still have this effect. Wright maintains that his etiological condition distinguishes between effects of a thing which are functions and those which are not. This condition is that the function of a thing is an effect of it that explains why the thing is where it is and also why it has the nature that it does.

Analogously to the formulation of the Taylor analysis, Wright offers the following set of conditions, (F):

The function of *X* is *Z* if and only if:

(i) *Z* is a consequence (result) of *X*'s being there

(ii) *X* is there because it does (results in) *Z*.

According to Wright, the main difference between (F) and (T) is that in (T) what is central is 'trying' behavior: actual success of the behavior is not necessary for the ascription of a goal to it. In contrast, he says, 'the most central cases of functions are all ones in which it can be said that X actually does Z'. He immediately qualifies this, by admitting that something may have a function that it never has and never will perform. but insists that it is required that the thing at least be able to do *Z* (under 'appropriate conditions'). According to Wright, we can account for function attributions in cases without human involvement in the same way as provided by (F) with only the 'slightest change in nuance'. In the case of conscious functions the consequence etiology is provided by conscious

---

9. Not all agree that such a paper would have this function.

selection and Wright's suggestion is that natural selection will provide the consequence etiology in the case of natural functions.

How the Analysis copes with the problems

(1) Backwards causation.

Wright is able to easily dismiss this objection. Wright considers the backwards causation objection as presented by B. F. Skinner:

> A spider does not possess the elaborate behavioral repertoire with which it constructs a web because that web will enable it to capture the food it needs to survive. It possesses this behavior because similar behavior on the part of spiders in the past has enabled them to capture the food they needed to survive. A series of events have been relevant to the behavior of web-making in its earlier evolutionary history. We are wrong in saying that we observe the purpose of the web when we observe similar events in the life of the individual.[10]

Wright suggests that Skinner's objection to natural functions stems from a misleading grammatical feature of (F). Skinner supposes that to say, as (F) does, that the spider possesses its web making ability because that helps it to catch food is to make a reference to that specific spider's future food catching. This would go against the actual etiology of the phenomenon. The consumption of a particular fly by a particular spider is not part of the etiology of that very spider's web making ability. Wright suggests as a more accurate rendering of a teleological explanation: 'spiders possess the ability to spin webs because web spinning helps catch food.'

---

10. B. F. Skinner in Braybrooke, Philosophical Problems of the Social Sciences, 52, quoted in Wright, Teleological Explanations, 9.

This makes it appear that the best way to put matters is to say that an *X* is now present because in the past *X*s have done Z. Wright urges us not to accept this formulation because it blurs the distinction between the functional and the vestigial. For example, both kidneys and appendixes exist because of the function they had in the past, but only kidneys currently have a function. To put the teleological explanation in the past tense would in this way be misleading. He says that the same criticism would apply to an attempt to render the conscious functional explanations in the past tense. Wright suggests that 'It's there because it does that' should be taken as shorthand for 'it's there because things like it in the appropriate way have that sort of property.'[11] He says that when this is understood 'the etiology is clear and the functional insight preserved.'

This is not clear, however. Any explanation as to why a particular entity exists, if it is not a conscious functional explanation, must make reference to the properties of entities other than it. In the conscious case, an explanation of the existence of a given entity may make reference to its properties, because the being that created or repositioned it would have the presence of that very object with those properties as its goal. Even in this case, the actual explanation would be that the thing is as it is, not because of its actual properties, but rather because some being intended there to be an entity with certain properties there. It would be possible to add that the thing whose presence is being explained actually has the properties it was intended to have, but this adds nothing to its causal explanation.

Wright says that it would in fact be misleading to insist on using the past tense for the following reason. Suppose one said, for example, 'The Titanic sank because it was the case that when one makes a big hole in the hull of ship it sinks'. This, he says, may be taken to imply that one could do such a thing now and not

---

11. Wright, Teleological Explanations, p. 90.

have the ship sink. However, whether anyone is likely to draw this faulty inference or not is irrelevant to the explanation of the phenomenon. If some causal correlation between properties is part of the explanation of a given phenomenon, it can only be the past holding of the correlation that matters for this. It cannot be part of the causal explanation of a phenomenon that a causal correlation will continue to hold after the phenomenon occurs. The explanation would continue to be correct even if the laws of physics subsequently changed. Similarly, if past *X*s being *Z* explains why a current *X* exists, that *X*s in general continue to be *Z* can only be irrelevant to this explanation.

Some of Wright's more recent followers have not followed him in making this claim. However, Wright's worry that we would be unable to distinguish between vestigial and currently functional cases seems well founded. If teleologists following his basic account reject the requirement that in order for a thing to have a function it must currently perform or be capable of performing the function, then vestigial phenomena, having just as much a functional explanation as actually functional ones, will incorrectly be ascribed functions. Just as in the conscious case, whether something does or is able to perform a function is no part of the explanation of why it came to exist.[12] On Wright's account, a thing's having a functional explanation need not guarantee that it has a function.[13] The obvious way to attempt to make the distinction is to suppose that the vestigial are those

---

12. That an entity has a particular function does not even feature in the explanation of why it continues to exist, for that explanation would still rely on its past performances of its function, not present or future ones. Perhaps the only exception is the explanation of why a particular thing exists at a particular moment, when its existence is dependent upon its performing its function at that moment.

13. This has relevance for those who base an account of intentionality on inferring functions from functional explanations. The present problem is that a particular functional explanation does not imply that the thing explained has the function any longer.

things with functional explanations that are no longer capable of carrying out their function. This won't do if one wants to be able to say that there are things that have functions that they are unable to perform. For cases of natural selection, the solution to this difficulty would seem to be to appeal not to the current functionality of the thing, but the functionality of its immediate ancestors. A thing could be said to have a certain function if it satisfies Wright's analysis and its immediate ancestors performed or were capable of performing the function. That there is some vagueness in how immediate a thing's ancestors have to be in order for it to still have a function is not troubling, so long as it matches whatever vagueness there already is in the notion of the vestigial.

(2) The failure to explain.

As the essence of Wright's account is that having a certain sort of causal explanation is constitutive of possessing a function, the objection that the account's supposed functional explanations fail to be genuinely explanatory is most troubling. To take the plant example quoted by Wright, the explanation request is 'why do the stomata of this plant close at night?'. The teleological explanation is that the stomata close in order to save water. Wright's claim is that the explanation of the existence of some phenomena may be given by the effects that phenomena of that type have. In this case, the closing of the stomata has the effect of saving water, and the phenomenon occurs, according to Wright, because it has this effect. The objection is that the genuine explanation of the phenomenon is chemical.

In the introduction to Teleological Explanations Wright grants (as one must) that this alternative explanation is a legitimate one, but insists that this is compatible with the legitimacy of the teleological explanation. It is unlikely that all phenomena with teleological explanations are instances of causal

overdetermination. So either these are explanations of subtly different phenomena, or they are appropriate explanations of the same phenomenon within different contexts of explanation. The latter seems to be the most promising, despite its threat of a resurrection of the distinction between efficient causes and final causes.

To take a conscious case as a parallel example, one might ask what a rolled-up newspaper is doing stuffed under the door. To be told that somebody put it there would be inadequate. The answer to the question of how the paper got where it is is obvious, the explanation request assumes that whoever put it where it is was trying to serve some end by putting it there, and is asking what this end is. In the case of the plant, for teleological explanations to be legitimate, there would have to be an explanation request to which the chemical answer is not appropriate but the teleological answer is.

The chemical explanation of the closing of the stomata can explain why any particular instance of closing occurs. What it cannot explain is why there is this process of closing at all. The chemical explanation may exhibit the causal sequence that leads to closing, and so explain why closing occurs on any given occasion, but it does not explain why this type of plant has this capacity. Unfortunately, neither does any teleological explanation that makes an appeal to natural selection. Such a teleological explanation cannot explain why a certain type of object came to possess a given trait, but it appears to be capable of explaining why things of that type continue to possess that trait. This appearance is not quite accurate, however. It is not strictly true that a plant possesses stomata that close because of the effects such stomata have on plants of that type. It turns out that this form of explanation is shorthand for some other explanation, which does not quite follow Wright's pattern.

The causal explanation of an organism's possession of a functional trait and the explanation of its possession of a non-functional trait (possibly vestigial) will

be effectively the same. The immediate explanation will be based on the type of DNA possessed by the organism and the explanation of why the organism has the type of DNA that it does will usually be that its ancestors had DNA of that type.[14] A particularly useful trait is no more likely to be passed on by an organism to its immediate descendents than one which is not particularly useful. Rather, if the organism survives to reproduce all its traits are passed on.[15] Obviously, an organism that possesses many traits important to survival that are adapted to its environment is more likely to survive than one which possesses considerably fewer such traits. Yet this does not mean that the adaptiveness of a given trait is responsible for the propagation of that very trait. It would be quite possible for highly adaptive traits to fail to be passed on because of shortcomings elsewhere in the organisms that possess them, and also for the comparatively useless to be carried along with other more useful traits (as is the case with the vestigial).

Wright's account of functions demands that we be able to separate out the contribution made by a particular trait to the survival of an organism. It is plausible to suppose that this demand may be satisfied to some degree, as some traits are obviously more adaptive than others. The account further demands that each trait's contribution is what explains its own existence. It is not plausible to suppose that this demand may be satisfied, for this explanation is in fact the same for all (non-mutant) traits possessed by a given organism, no matter how adaptive or unadaptive: the trait exists in this organism because it was possessed by an ancestor organism that was sufficiently adapted to its environment for it to survive to reproduce.

---

14. Of course, mutations would be exceptions to this generalization.

15. This is true only for simple organisms that reproduce by fission. For more complex organisms that reproduce sexually, only one half of its traits will be passed on to its descendents. However, each trait of an organism has an equal chance of being passed on to its immediate descendents.

This suggests that Wright's account be modified by weakening the explanatory requirement. Instead of accounting for a thing's function as an effect of it that explains its existence, instead it should be said that a thing's function is an effect of it that makes a contribution to its explanation. In the case of natural selection, a simpler way to put this is to say that a thing's functions are those of its effects which increase the survivability of the organism of which it is a part. Wright's analysis, then, turns out to be similar to that offered by others, such as Canfield, who suggests that *X* has a function *Z* when *X*s doing *Z* increases the probability of the survival of the system of which *X* is a part.[16] The difference is that Wright's analysis is historical, maintaining that it is the past performance of *Z* by things of type *X* that confers the function *Z* on present *X*s.

(3) Illegitimate anthropomorphism.

Timothy L. S. Sprigge has proposed a *reductio* of Wright's analysis.[17] The argument is that according to Wright's analysis, every law-governed event will be teleological, as it occurs because it tends to achieve the state of affairs consisting in the satisfaction of the law. For example, if *A* causes *B*, the argument is that *B* occurs because it produces the state of affairs *C*, which is the satisfaction of the law '*A* causes *B*'. In order to fit the Wright analysis, it would have to be the case that *C* is an effect of *B*, i.e. that *B*s cause *C*s, and that a particular *B* now exists because in the past *B*s have caused *C*. It seems to me that this objection is more than a little forced. *C* is not really an effect of *B*, for *C* is not temporarily subsequent to *B*. *C* occurs cotemporaneously with *B*. Even if this were not the

16. J. Canfield, 'Teleological Explanation in Biology', British Journal of the Philosophy of Science 14 (1963).

17. Timothy L. S. Sprigge, 'Final Causes', Proceedings of the Aristotelian Society Suppl., 1971.

case, *C* would not be caused simply by *B*, but by *B* following *A* in accordance with a law. Furthermore, present *B*s would not exist because past *B*s have produced *C* (or at least this would depend on the account of laws on offer).

Amongst the other counter-examples to Wright's analysis in the literature, the most challenging is Boorse's example of a pipe which leaks a dangerous gas that prevents anyone from effecting its repair. This is problematic for Wright, as the leak persists owing to its releasing the gas, but the release of gas is clearly not a function of the leak. A solution to these counterexamples has been suggested by Michael Levin:

> The difference between hearts and Boorse's leak, to which intuition seems to be responding, lies in a difference between the role of the efficient cause of the leak – mounting gas pressure, say – and the role of the efficient cause of hearts, the process that builds hearts from protein. Not only does the heart exist because of what it does, namely circulate blood, heart-building occurs because of what *it* does, namely build hearts. The formation of heart-muscle tissue is itself explained via the fitness advantage it confers on organisms by making (fitness-enhancing) hearts. In contrast, the pressure in the pipe did not mount because it was leading to a leak.[18]

Levin calls the requirement that the cause of a thing that possesses a function have a functional explanation the 'chain condition'. As Levin points out, the chain condition also applies to conscious functions. This is because the mechanism of beliefs and desires that permit the formation and execution of plans will itself be explained by the adaptiveness of the behavior it produces.

---

18. Michael Levin, 'Plantinga on Functions and The Theory of Evolution,' Australasian Journal of Philosophy 75 (1997), 89.

Nevertheless, Wright's analysis does not capture what we ordinarily mean by 'function'. Wright notes that one may consciously select something without having what we would ordinarily describe as a reason for selecting it. He calls these cases 'mere discrimination' as opposed to other cases where 'some advantage is at least implicit'.[19] These cases he calls 'consequence-selection'. Wright claims that it is consequence selection that is the basis for assignment of functions rather than mere discrimination. This seems to be correct. Faced with a situation in which a choice must be made but there is no particular reason to choose one rather than another of the alternatives, some arbitrary selection criterion will be employed which does not confer a function.

Consider an ice explorer who finds that he has to shoot and eat half of his dogs. Suppose that half of his dogs have white coats and half do not. The explorer decides to keep the dogs with white coats and eat the rest. By carrying out this process of selection, no new functions have been conferred on the dogs, or on their coats. As the idea behind consequence selection is that the thing is chosen because of what it does (or is expected to do) for the one who chooses it, Wright should count this selection process as one of mere discrimination rather than consequence selection. In making this particular cut, neither the dogs, nor their coats, have been selected for their effects.

Wright claims that natural selection can be seen as an extension of consequence selection. The obvious obstacle to seeing natural selection in this way is that no being is actually doing the selecting in the case of natural selection. In response to this objection Wright gives the following reply:

---

19. Wright, Teleological Explanations, 85. The advantage is presumably that of the being who is doing the selecting.

> Consequence selection, by contrast with mere discrimination, de-emphasizes volition in just such a way as to blur its distinction from natural selection on precisely this point. In the conscious cases, the consequences are selection *criteria*. So we can say in these cases too that given X, Z, and the environment, which includes the selection criteria, X will be selected automatically. . . .[20]

Wright concludes that 'From an etiological point of view, it should be clear how difficult, not to say obscurantist, it is to drive much of a conceptual wedge between conscious and natural consequence-selection.'[21]

These remarks are not convincing. Selection criteria are operative in cases of mere discrimination just as much as they are in cases of consequence selection. The stated difference between mere discrimination and consequence selection is not that there is no principle of selection in cases of mere discrimination but that the selector gains some advantage in employment of the principle in cases of consequence selection, and there is no such selector, and so no such advantage, in the case of natural selection.[22] One might think that the intended parallel in the case of natural selection is benefit to the organism, but not only is this not what Wright says, it will not do the required work on its own. The principal reason for this is that the creatures themselves are not the ones doing the selecting. In the case of the ice explorer discussed above, it turned out to be good for the dogs that they had white coats, but no function was conferred on their having white coats by being selected for survival.

This issue is important, for whatever makes the difference between consequence selection and mere discrimination is the difference between a process

---

20. Ibid., 86.
21. Ibid., 87.
22. See footnote, ibid., 85.

of selection that is function-bestowing and one that is not. If we are to regard natural selection as a function-bestowing process, then it will be because it shares the distinguishing feature of consequence selection.

In attempting to draw a parallel between natural selection and consequence selection, what Wright probably has in mind is the fact that it is no accident that certain creatures are favored in natural selection. Given the nature of the environment, those beings most adapted to it are the ones most likely to flourish, and so there is a reason why certain types of creatures are naturally selected rather than others, just as in consequence selection there is a reason why certain things are chosen by a conscious being. They are chosen because the choosing being sees them as means to one of its ends. In consequence selection, objects are where they are, having the properties that they do, because someone expected them to have a certain effect, not because they do have this particular effect. This is the principal reason why there are artifacts which are very poor at performing their functions. Conscious functions, therefore, are tied to goal-directed behavior in a direct way in that they are means to ends.

This feature of consequence selection does not directly distinguish between it and mere discrimination. Even in cases of mere discrimination, the thing chosen will often be desired for some end, too. The ice explorer who chooses to keep his white dogs still wants the dogs to pull his sled. Furthermore, in some sense of 'reason' there is a reason in mere discrimination cases why some things are chosen rather than others, for some selection criterion will be employed, even if it is based on some random event like the toss of a coin.

There are relevant differences between mere discrimination and consequence selection, however. The difference which makes a process of selection a case of consequence selection rather than mere discrimination is in the relation between the selection criterion employed and the ultimate end of the chooser. In

consequence selection, the criterion employed is relevant to the end, in that things which meet the criterion are thought more likely to realize the end than things which do not. Even though things chosen by mere discrimination will generally also be wanted for some end, it is not thought that their meeting the criterion makes them more likely to realize the end than things which do not.[23] Employment of the selection criterion in consequence selection results in the attribution of a function, for that choice is made in order to realize some end. Employment of the selection criterion in mere discrimination does not, for meeting the criterion is irrelevant to whether the thing chosen will achieve the end for which it is desired.

It remains to be seen whether natural selection should be regarded as analogous to the non-function-bestowing mere discrimination or the function-bestowing consequence selection. For conditions (F) to be satisfied there has to be an entity *X* which has an effect *Z*, and *X* exists (in part) because it produces *Z*. Features that have been naturally selected appear to fit (F) because they exist now, in part, owing to the causal powers of their ancestors. In arguing that natural selection is akin to consequence selection, Wright makes the point that in cases of consequence selection by a conscious agent, the chosen consequences meet selection criteria: the agent wishes there to be entities with those effects. Wright attempts to get around the absence of a selecting agent in the case of natural selection by supposing that the nature of the environment determines that there are entities with certain effects. In other words, the environment plays the role of chooser in natural selection.

The fault with this suggestion, as the basis of a theory of functions, is that the nature of the environment does not supply the selection criteria in the same way

---

23. Things which meet the criterion of a consequence selection may turn out not to be suitable for realizing the end, and the criterion employed in mere discrimination may serendipitously turn out to be relevant to doing so.

that the intentions of a conscious agent do in the case of consequence selection. The distinguishing feature of consequence selection is not that there is some criterion of selection but that the criterion of selection is relevant (or at least is intended to be relevant) to the end of the chooser.

Natural selection differs from consequence selection in this crucial respect, for it is not a goal directed process. For any given environment and organism, there are a number of ways in which its descendents may become better adapted to that environment and factors other than purely adaptationist ones will be involved in which of these paths will in fact be taken. This stands in contrast to a conscious chooser, who is aiming at a specific goal. Although there may be many alternative means to a given end, and the chooser may in fact choose one that is less than optimal, the goal provides a yardstick against which the success of the choice may be measured. In the case of natural selection, there is no such measure. There is nothing that makes it possible to say of the creatures that have actually survived that they have done so 'in error', owing to their failure to measure up to the goal of the process. In natural selection, the environment is not, as it were, exhibiting 'trying behavior'.

As conscious functions are means to ends, there is a disanalogy between natural selection and conscious functions, as the natural selection of certain creatures does not serve any end. So literal attributions of function on the basis of natural selection really are illegitimate anthropomorphisms, as being goal-directed is what makes consequence selection function bestowing.

I have so far argued that were Wright's analysis of function correct, then natural selection should not be regarded as a function-bestowing process. Wright's analysis, however, is not correct. Consider again the example of the ice explorer discussed above. In a function bestowing selection process, a thing is selected because of what it does. The ice explorer, in deciding not to shoot the white dogs,

cannot bestow a function on their white coats, for there is nothing that the whiteness of the coats does that explains their existence. However, it would be possible for the explorer to adopt a criterion of selection which did have this feature. Suppose instead that he decides not to shoot those dogs which blend in best with the surroundings, thus preserving the dogs with the white coats. It will now be possible to explain the (continued) existence of the dogs with white coats by their blending in with the surroundings. Yet the new selection criterion is just as arbitrarily selected, relative to the explorer's ends, as the old one.

It would be decidedly odd to suppose that in such a situation the dogs with white coats had acquired the function of blending in with the surroundings. The dogs themselves have the same function that they always did, i.e. hauling the sled, and their coats have only their former biological functions. Contrast this with a situation in which the explorer wishes to keep a low profile and so takes only dogs with white coats on the expedition. In this situation it does make sense to say that the whiteness of the dogs coats has the function of blending in with the surroundings. This is because the dogs' blending in with their surroundings is now part of achieving the goals of the explorer. In the case of the arbitrary selection criterion, whether the dogs actually do what they have been selected for plays no role in achieving the goals of the one doing the selecting.

The problem with Wright's analysis is that he equates 'x exists in order to do y' with 'x exists because it does y'. These are closely related notions, in that very often when something is selected because it has a particular effect, it will be in order for it to have that effect. The above example illustrates that a thing may exist because of what it does without it existing in order to do that thing.

This point is most clearly seen where the selection criterion involves a capacity that will not be utilized. For example, NASA had a difficult problem in deciding which of the Mercury astronauts to send first into space, for each was as

well qualified as any other. One solution to this problem would have been to organize a snooker tournament, with the winning player being sent on the first mission. If Alan Shepard had won the tournament, his skill at snooker would explain why he was the first American in space, but it would not have been the case that he was sent into space in order to play snooker.

A defender of Wright's analysis may insist that the explorer's selection would fail to be function-bestowing according to his analysis because no benefit accrues to the explorer from selecting those particular dogs. This means that Wright's point about benefit in distinguishing mere discrimination from consequence-selection is something additional to conditions (F) rather than an explication of them. If so, and Wright wishes to offer a comprehensive account of functions, (F) plus the additional condition should apply to all other cases of functions. As noted above, there is no analogous benefit in the case of natural selection.

The difference between a function-bestowing selection process and a non-function-bestowing selection process does not lie in the explanation of why the thing is where it is with the nature that it has, but what the purpose of its being there is, and this is something that is only present in a goal-directed process. As being directed towards a goal is an essential part of a function-bestowing selection process, the natural modification to (F), is to add that the explanation of X, in terms of X causing Z, be such that X's causing Z is a means to some goal. This prevent natural selection from being function-bestowing, because it is not a goal-directed process. That aside, if being goal-directed is to be understood in intentional terms, an account of functions that depended upon an account of goal-directedness could not be used to give an account of intentionality.

Teleology without function

The failure of Wright's account as an analysis of 'function' does not mean that it may not form the basis for a theory of intentionality. This failure means that those teleological theories based upon Wright's account lose a degree of their intuitive support, given that the phrases commonly used in stating the theories, such as what a state is 'supposed to' represent, are not to be taken literally. However, it is still open to the teleologist to hold that the Wright causal-explanational model may still be used to determine content. Although it can no longer be literally maintained on the basis of the analysis, for example, that it is the function of a state to carry certain information, it may nevertheless be the case that the contents assigned by a theory using the Wright analysis meet the constraints on a successful account of intentionality. The teleologist may simply let each 'function' assigned by Wright's analysis be known as a 'function*' and go on to insist, for example, that the content of a state is whatever it is its function* to represent.

Dennett provides a simple example of this sort of account of content. Consider a soda vending machine which is designed and built in the United States. It is designed to accept only U.S. quarters. When a quarter is inserted and accepted by the machine, it goes into state *Q*. Being a fallible device, sometimes it rejects a quarter and sometimes it accepts an object that isn't a quarter. In fact, the machine is sensitive to the insertion of objects of a certain size and weight (call this class of objects '*K'*) into its slot. The machine will serve adequately as a quarter-detector, as in its environment (the United States), most K things are also quarters. Furthermore, it is clear about this machine that it serves as a quarter-detector rather than a *K*-detector, or a very bad dime-detector, because of the intentions of its designers, builders and owners. The machine has been manufactured and employed by these people for the purpose of gathering quarters. It is only relative to these intentions that we may regard state *Q* as being caused in

error when the machine accepts a non-quarter. In so far as the vending machine exhibits intentionality, it is only derived.

Dennett's purpose is to show that whatever intentionality we possess is as derived as that of artifacts such as the vending machine. That nothing intrinsic to the machine makes it a detector of U.S. quarters is shown by the fact that such a machine may be exported to Panama and serve just as well as a detector of Panamanian quarter-balboas (as these coins are struck from the same stock as U.S. quarters). This sort of example indicates that with artifacts, it is not merely the intentions of the designers that may result in the assignment of functions, but also the purposes to which the objects are put, which suggests a parallel with the case of evolved beings. A U.S. vending machine that accidentally found its way to Panama would put in successful service, whereas one that was misdirected from, say, Japan, would not. The Wright analysis may then be used to assign a function* to the mechanism which decides whether to accept or reject an object inserted into the machine's slot, which in this case would be to detect quarter-balboas, as that activity contributes to the explanation of why the mechanism persists where it is.

Fodor has criticized this attempt to provide the basis for a theory of intentionality owing to its indeterminacy. His claim is that given a naturalistic account of teleology, an alternative teleological story can be told about the content-bearing states. For example, the visual system of frogs is sensitive to small dark moving spots on the retina.[24] As almost all the things that tend to cause these spots in the frogs natural environment are flies, this mechanism acts as a fly-detector. However, instead of supposing that the frog detects flies, one may instead claim that the frog actually detects small black moving dots. Because enough of the

---

24. Lettvin, Maturana, McCulloch and Pitts, 'What the Frog's Eye tells the Frog's Brain,' Proceedings of the Institute of Radio Engineers, 1959.

small black dots in the environment are also flies, it makes no difference to the success of the evolutionary explanation. It does, however, make a difference to the resulting assignment of content. On the teleological account, if the function (or the function*) of the mental state is to indicate the presence of black dots, then that is its intentional content. Fodor concludes, 'Darwin cares how many flies you eat, but not what description you eat them under'.

Some teleologists may be tempted to reply to this objection by appealing to Dretske's influential paper, Misrepresentation. Dretske declares that only when we understand 'nature's way of making mistake' will we understand our own (underived) capacities for misrepresentation, and hence representation. Dretske wishes to base an understanding of intentionality upon the notion of a natural sign. The foremost problem with natural signs and their natural meanings is that they are incapable of misrepresentation: an occurrence of a natural sign means that *p* only if *p*. Suppose that *d*'s being *G* is a natural sign of *w*'s being *F*. Dretske makes the natural suggestion that if it is the function of *d* to indicate the condition of *w* then it does not matter whether or not *w* actually is *F* for *d*'s being *G* to mean this. Dretske's definition of functionally derived meaning is as follows:

(M_f) *d*'s being *G* means$_f$ that *w* is *f* = *d*'s function is to indicate the condition of *w*, and the way it performs this function is, in part, by indicating that *w* is *F* by its (*d*'s) being *G*[25].

To illustrate this, he gives an example of a fuel gauge which has the function of indicating the amount of fuel remaining in the tank. The way it does this (in part) is by its needle pointing to the inscription 'full' when the tank is full, and pointing to 'empty' when it is empty.

---

25. Dretske, 'Misrepresentation.' In Belief, edited by Radu Bogdan (Oxford: Oxford University Press), 21. The reference to function in the definiens may be replaced by a reference to function*.

Dretske's project is to show how a natural sign may misrepresent by looking for what it is supposed to $mean_n$, rather than what it does $mean_n$. The 'supposed to' is to be cashed out in functional terms. Of course, 'function' itself needs to be cashed out. Dretske tells us where to look for this:

> The obvious place to look for natural functions is in biological systems having a variety of organs, mechanisms, and processes that were developed (flourished, preserved) because they played a vital information-gathering role in the species' adaptation to its surroundings.[26]

To illustrate how M_f is supposed to work, Dretske asks us to consider simple organisms with pressing biological needs. His main example concerns marine bacteria possessed of magnetosomes that are sensitive to the earth's magnetic field. In the northern hemisphere these bacteria move towards geomagnetic north and in the southern hemisphere towards geomagnetic south. The survival value of this arrangement is to prevent the bacteria from entering toxic water, as they can only survive in the absence of oxygen. Were bacteria from the northern and southern hemispheres to be transplanted they would continue to head towards the respective geomagnetic poles and be destroyed.

Assuming that a teleological/informational account of representation will assign the magnetosomes of these bacteria the function of representing something, the obvious thing to say is that the relevant states of the bacteria represent the direction of oxygen-free water and are misrepresenting when they lead transplanted organisms to their destruction. The problem is that it is not immediately clear whether it is the direction of oxygen-free water or (for northern bacteria) the direction of magnetic north that is being represented. On the latter

---

26. Ibid., p. 25. This appeals to a notion of information, which itself needs to be cashed out. Dretske's proposal is a species of Wright's account of functions.

way of describing things this would mean that the mechanism was working perfectly well when it led the organism, transplanted into a foreign hemisphere, into toxic waters. Adapting Fodor's objection, Darwin cares what direction the bacteria go in, but not how that direction is to be described.

Dretske presents his problem in the following way. Given that a system needs *F* and has a mechanism, *M*, which enables the organism to obtain *F*, how does the mechanism carry out its function? It may do this either by directly indicating the presence of *F*s or by directly indicating the presence of something else, *G*s. The latter technique will be reliable when there is a correlation between *F*s and *G*s in the system's environment, so that *G* is a natural sign of *F*.

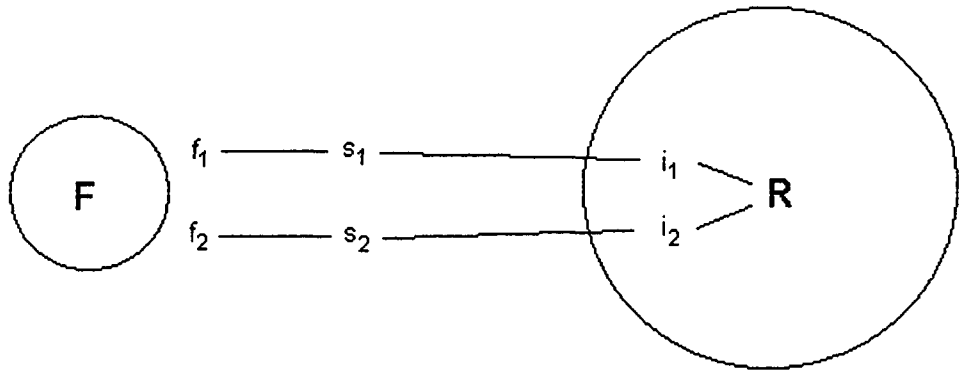
Given that the sensory mechanism of the bacteria is magenetic and not chemical, Dretske suggests it is unreasonable to regard the mechanism as indicating the presence of oxygen, but rather (for northern bacteria) one should take it to be the direction of magnetic north.[27] This means that if a northern bacterium is transplanted into the southern hemisphere, then when it moves towards the surface and dies, it will not be misrepresenting anything. However, notes Dretske, the original problem reappears, for now we can describe the function of the magnetosome as either representing the direction of geomagnetic north or the direction of the surrounding magnetic field. If the latter description is chosen, says Dretske, the internal states of the organism will always be accurate in the same way that a compass is, i.e it will be a system that is incapable of error.

Dretske's solution to this problem is that a necessary condition of a state *R* representing *F* is that there be more than one information carrying route from *F* to *R*. The presence of a property of *F*, f_1, will bring about (or be correlated with)

---

27. It seems to me that one would want conditions such as this to be a consequence of one's theory of functions. In fact, this is not a consequence of a Wright-style account of functions, this slack having to be taken up by *ad hoc* measures.

environmental events s_1, which will bring about internal events, i_1, of the system, which will result in *R*. In order for *R* to represent *F*, other properties of *F*, f_n must bring about another sequence of events (s_n, in, *R*). Dretske claims that it is impossible to say that *R* $means_f$ anything more proximal than *F*. Even though s_1 is a triggering event (ultimately) of *R*, *R* cannot $mean_f$ s_1, for it cannot $mean_n$ s_1.[28]



According to this account, indeterminacy of function for magnetosomes arose because the mechanism has only one way of detecting the direction of the relevant magnetic pole. If the bacteria had been equipped with another way of doing this (and resulting in the tokening of *R*) we would have been able to rule out the interpretation of the relevant internal state, *R*, as $meaning_f$ the direction of the surrounding magnetic field, for then *R* could not $mean_n$ that, even under 'optimal conditions'.

There is a little more to Dretske's account than this. It has been objected that *R* does $mean_n$ the disjunction of f_1 and f_2 (or of s_1 and s_2) and so may be taken to $mean_f$ these disjunctions. In response to this, Dretske has added an associative learning requirement. This requirement is that the system is such that after

---

28. This inference is curious, for it is not impossible for something to have the function of representing *F*, even though it does not, and cannot, carry the information that *F* is present. The very point of appealing to $meaning_f$ is because no actual representation $means_n$ what it represents. Perhaps Dretske has in mind that in 'optimal conditions' something that $means_f$ *F* would $mean_n$ *F*.

exposure to a stimulus in the presence of *F*, *R* will be triggered by the occurrence of the stimulus alone. There will be no time-invariant $meaning_n$ amongst the intermediate elements in the causal chains leading to *R*. The disjunction of intermediate causes of *R*, of which *R* is a natural sign, will change each time the system learns to token *R* in the presence of a new stimulus. However, because each of these intermediate entities is a natural sign of *F*, *R* will continue to $mean_n$ *F*.[29] In the light of this modification to the view, we can see the actual number of channels of information does not matter, it can be more than one or just one. What does matter is that it is possible for the organism to acquire at least another channel than the one(s) it already has.

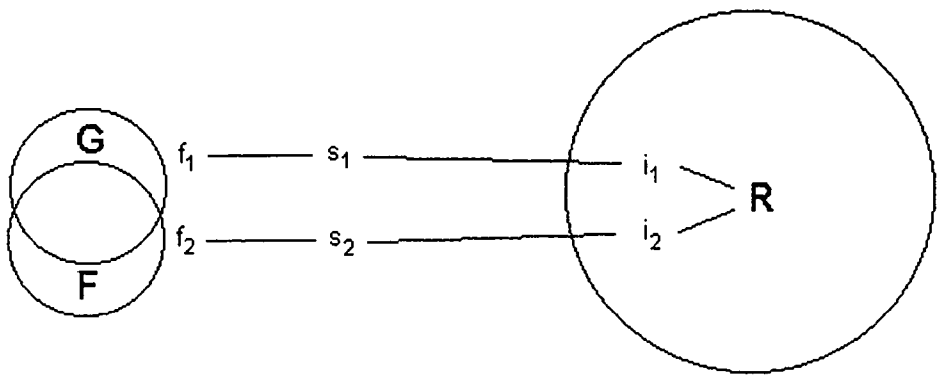
One curious aspect of this argument is that it relies on the premise that a necessary condition of a representation *R* $meaning_f$ *F* is that *R* is a natural sign of *F*. In short, that if something does not $mean_n$ *F*, then it does not $mean_f$ *F*. Given that errors do in fact occur, this is false. If anything actually does have the function of meaning something, it will in general not be a natural sign of that thing. This was the main reason for appealing to functional meaning in the first place. It is Dretske's assumption that a theory of functions will leave it indeterminate what entity in a causal chain to a representation is to be taken as the content of the representation. I take his intention to be the provision of a supplement to a theory of functions which will take up the slack. Although the premise he appeals to is false, perhaps the following line of reasoning would serve in its stead: It would be unmotivated to declare f_1 rather than f_2 to be the content of *R*. If one were to do this, f_2's causing *R* would have to be regarded as an error case, and, by hypothesis, the theory of functions provides no basis for doing that.

---

29. A reason to doubt the usefulness of this move is that the stimulus that now triggers R will also be a 'natural sign' (in so far as such a thing is possible) of the properties of *F*, f_1, f_2, etc.

So, the theory of functions would leave it indeterminate whether *R* represented the thing at the end of the casual chain, *F*, or a disjunction of the intermediaries. At this point, Dretske is able to employ his associative learning requirement to determine *F* as the content of *R*.

Supposing that Dretske's solution to the given problem of indeterminacy is acceptable, he has shown that there is a principled way of distinguishing between the elements in a causal chain leading to an information carrying token. This provides an effective reply to Fodor's objection to teleological accounts of content, where frogs are reinterpreted as snapping at patterns of light, or retinal images, or the like. Another way of presenting Fodor's argument, which appears to me to be the way he intended it, is to suppose that the frogs' internal states represent the usual object in the causal chain, i.e. flies, but do not represent them as flies.

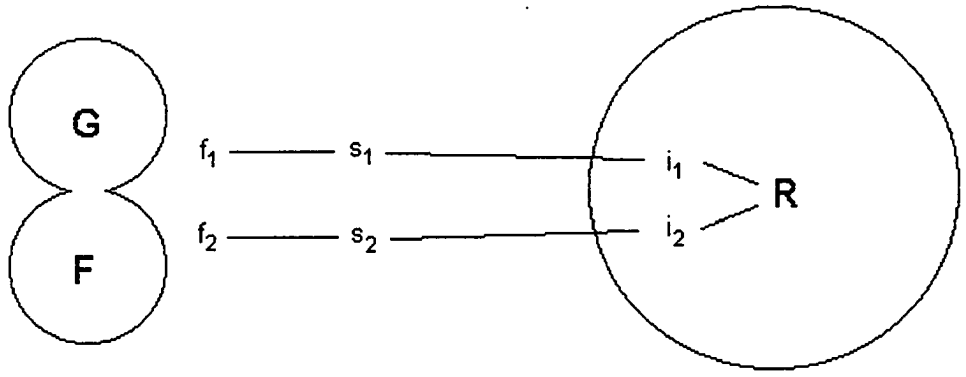


Given that the local *F*s are also *G*s (rather than merely correlated with *G*s), Dretske cannot show that *R* does not mean_f *G* with the same reasoning he used to show that *R* does not mean_f s_1 or f_1. If it turns out that there is a route to *R* from *F*s which are not *G*, but no route to *R* from *G*s which are not *F*, then Dretske would be able to claim that *R* does not mean_n *G*, and so does not mean_f G.[30] As noted

30. Officially, it is not required that there is an actual route, only that the system could come to acquire such a route. This solution is ineffective if there is a route to *R* from *G*s which are not *F* (i.e. errors) or if such a route could be learned.

above, Dretske is not really entitled to this premise, so he would have the harder task of showing that, assuming some functional theory of content, it is not open to ascribe *R* the function of representing *G* and treating the non-*G* *F*s which cause *R* as error cases.

Even if Dretske has managed to provide a solution to the distal-cause problem, the main problem facing informational theories, the disjunction problem, remains to be addressed.



One might use Dretske's argument in an effort to solve the disjunction problem by attempting to show that it is only *F* that *R* will be a sign of as a result of a learning process. The idea behind the associative learning condition is that the new stimulus, *cs*, will be sufficient to produce *R* on its own. *R* is held to continue representing *F* in these cases because *cs* is supposed to be a natural sign of *F*. There is of course no reason why a particular *cs* should also be a natural sign of *G*. However, to rule out *F*-or-*G* as the interpretation of *R*, it would have to be the case that the mechanism could not be conditioned to produce R in the presence of *cs* and *G*. I can think of no reason why this should be the case.

This is not really a criticism of Dretske's work, for it was never presented as provide a solution to the disjunction problem. He is assuming that that sort of

indeterminacy will be taken care of by the theory of functions. Godfrey-Smith accuses Dretske of begging the question in his definition of functional meaning.[31]

Godfrey-Smith's point is that Wright's account will not support Dretske's definition of functional meaning. On Dretske's definition, *d*'s being *G* means that *w* is *F*. Godfrey-Smith asks, "[W]hy *F*? Why is it not *w*'s being *F* or *H*, where *H* is a property of *w*'s easily mistaken for *F*-ness?" One would expect Dretske to respond to this complaint by appealing to the theory of functions. Godfrey-Smith's objection, however, is that even given that *d* has the function of indicating the condition of *w*, we cannot suppose that states of *d* (such as *d*'s being *G*) have the function of indicating the states of *w* (such as *w*'s being *F*). This is because, he says, Wright's theory of functions will not support the attribution of functions to states of mechanisms rather than mechanisms.

The reason why Godfrey-Smith says this is that although being in particular states may be beneficial to an organism, the nature of that state is not the result of the nature of previous states of the same type, for there is no way for the properties of a state to be such as to bring about future states of the same type. For example, the visual system of human beings has evolved, we may say, in order to gather information about our surroundings. There are no features of a particular visual, experience, however, which make it likely that through natural selection there will be more visual experiences with the same character.

This isn't really an accurate criticism of Dretske's account as Dretske does not attribute functions to states of mechanisms. On his definition, *d*'s function is to indicate the condition of *w*, and this function is performed by *d*'s being *G* indicating that *w* is *F*. For any object *d* which has the function of indicating the

---

31. Peter Godfrey-Smith, 'Misinformation,' Canadian Journal of Philosophy (1989).

condition of *w*, and is capable of performing its function, some of its states must be coordinated with the states of *w*. If they were not, it would not be able to perform its function. So Dretske's claim is not that the signs themselves have functions, but rather that some states $mean_f$ certain things because they are part of a system with a representational function.[32]

Probably the most sophisticated teleological theory of intentionality is that developed by Millikan. Her approach is somewhat different from Dretske's but the theory is also based on Wright's theory of functions. In order to present Millikan's theory, some of its technical terms must first be explained. An object *x* has the proper function *F* if *x* exists with property *C* because ancestors of *x* performed *F* by having *C*. 'Ancestor' is also one of Millikan's technical terms. Something is an ancestor of *x* only if it is a member of the same reproductively established family as *x*. A first-order reproductively established family is a set of things which share a character owing to their being copied from a model, such as a set of xeroxed pages of text. When the members of a first-order reproductively established family have the proper function of producing another type of thing, tokens of that type form a higher-order reproductively established family. Human hearts, for example, form a reproductively established family, even though they are not directly copied from one another. This is because it is the proper function of certain genes to produce hearts, and these genes are directly copied from one another and so form a first-order reproductively established family. A Normal explanation is an explanation of how, historically, members of a reproductively established family have performed a proper function.

Millikan refers to something that possesses intentionality, such as a sentence or a belief, as an 'intentional icon'. Intentional icons are members of

---

32. Dretske makes this explicit in 'Misrepresentation', 25.

reproductively established families. Speakers acquire words, for example, by observing instances of the word used by others. As the existence of current tokens of a type of word can be explained by the history of its ancestors (the English language, for example, has proliferated because its past users found it useful), the words in a sentence will have a proper function, and the whole sentence will have a proper function owing to the functions of its parts. Like other function-performing entities, intentional icons are unable to perform their functions in isolation (hearts, for example, are unable to pump blood around the body without, amongst other things, a functioning circulatory system). Language users have what Millikan calls 'producer' and 'interpreter' devices, whose proper functions are, as one would expect, to produce and interpret sentences.

Elements of some intentional icons, such as words, can have what Millikan calls a 'real value'. Some (but not all) real values are also referents. Millikan holds that there is a sort of correspondence, or 'mapping' as she puts it, between intentional icons and the world. Intentional icons have real values only when they are supposed to map onto something. This is only a necessary and not a sufficient condition. An intentional icon may be supposed to map onto a given thing, but fail to do so. Should this situation occur, the icon would not have a real value. A term only has a real value (and hence a referent) when it actually maps onto the thing that it is supposed to map onto.

A referring expression, such as a proper name, will be a member of a reproductively established family with a proper function such that tokens of the name, when performing this function in accordance with a Normal explanation are supposed to correspond to the referent of the name. All those tokens that do perform this function will actually correspond to the referent, rather then merely being supposed to do so. This is not simply a relation between the name and its referent, as in a simple causal theory of reference. In Millikan's theory, the

referent of a term depends upon the role it plays within the sentence in which it occurs.

The mapping of the elements of a sentence onto their referents is determined by the role the sentence plays in enabling the functions of the producer and interpreter devices for the sentence. The idea is that there is some correspondence between the parts of a sentence and things in the world that is part of a Normal explanation of how the interpreter and producer devices perform their functions. The simplest example Millikan gives of how this works concerns bee dances.

The devices that produce and interpret bee dances have evolved because of their effectiveness in leading bees to nectar. When a bee that has spotted a supply of nectar returns to the hive, it does a dance. The device within the bee that produces the dance has a function of producing movements that will result in a direction and duration of flight towards the nectar. The interpreter devices within the bees at the hive have the function of producing a direction and duration of flight on the basis of the movements produced by the returned bee. The crucial element that permits these devices to perform their functions is the bee dance. In order for the producer and interpreter devices to direct bees towards nectar, there must be some regular relation between the dances and the direction and distance to sources of nectar. As Millikan puts it:

> Intuitively it is clear that in some sense of "mapping," the bee dance that causes watching bees to find nectar in accordance with a historically Normal explanation is one that maps in accordance with certain rules onto a real configuration involving nectar, sun, and hive. As such it is an indicative intentional icon.[33]

---

33. Ruth Millikan, Language, Thought, and other Biological Categories (Cambridge, MA: MIT Press, 1984), 99.

Although Millikan's theory of intentionality puts Wright's theory of functions to use differently than Dretske's, there are evident similarities. The basic claim is that there is a correlation between intentional states and types of object in the world, which correlation is determined by the contribution the state makes to an explanation of the success of the organism. The failure of Wright's theory of functions to determine the relevant correlations will afflict Millikan's theory as well as Dretske's.

Let us suppose that certain mechanisms are adaptive owing to their states carrying information about their surroundings. Adopting Wright's analysis, we are entitled to attribute a function* to the mechanism, and in line with Dretske's suggested definition of functional meaning, we may say that the states of the mechanism represent various features of the environment. This does not yet provide us with a theory of content, however. The role of the theory of functions is to constrain acceptable interpretations of the content-bearing states, as the information actually carried by the states includes the error cases.

One problem for a Wright-based account of function in performing this role is that it must be based on historical performance, as it is a species of causal explanation. One may take the view that the actual performance of the mechanism is the relevant condition in providing an account of the content of its states. If a mechanism is to be explained by its serving to indicate the nature of its environment, one should regard each state as being correlated with the features of the environment that cause it, rather than with the idealized performance preferred by teleologists. To take Dennett's example of the vending machine, the machine actually serves to detect and accept objects of a certain size and weight, the members of class *K*. The additional fact that contributes to the explanation of the success of the machine at its job is that *K* is largely correlated with quarters in its environment. In order to explain the success of the machine, it is unnecessary to

narrow the class its state *Q* is correlated with to quarters. It is not as if by doing this one is spared the need to include the nature of the environment in the explanation, for one must still add that there is a sufficient quantity of quarters in its environment for it to persist where it is (a vending machine placed in a poor location will be moved on, for example, even if it is working as designed).

The teleologist needs to show that a better explanation of the success of an entity is provided by narrowing the class that its states are functionally correlated with. Perhaps an explanation request for the success of a vending machine in a context where it is well known that there are lots of quarters around would be better answered by citing a correlation between *Q* and quarters, rather than *K*, particularly if the only people who are really familiar with *K* are the engineers that designed the machine. The problem with relying on such features of the context of explanation is that it makes which explanation is to be preferred, and so what the content of the states of the entity are, a relative matter. What content a state possessed would be a subjective matter, dependent on facts about the one who attributes the content.

A different, and probably more satisfactory, way for the teleologist to respond to this problem is to note that what matters for the success of a mechanism, whose function is indicate the nature of its environment, is making veridical tokenings rather than erroneous ones. For the success of the vending machine, for example, the absolute quantity of objects of kind *K* it accepts is immaterial. All that matters is how many of the *K* things it accepts are quarters. Similarly for the frogs, if what matters for their survival is how many flies they eat, then the best explanation of their success, and hence the function* attributed to some aspect of their visual systems, is that they are detecting flies, and not merely moving black dots. So, it appears, there is something to be gained by narrowing the class correlated with the content-bearing state. This is to show what it is,

amongst the class of things that the mechanism is capable of detecting, the detection of which is responsible for the mechanism's success.

Unfortunately for teleological theories of content, this line of reasoning is not satisfactory. Furthermore, I do not think that there is any principled way of restricting the classes attributed as contents by a teleological theory of intentionality in the ways that the proponents of those theories wish. In part, this pessimism stems from reflection on Putnam's Twin Earth thought-experiment. As is well known, the point of this thought-experiment is that the referents of our terms are partly determined by our environment. The reason for this is because we have a fairly clear idea of what relation a substance has to bear to local samples of water in order to fall into the extension of 'water'. In particular, we believe that it has to have the same chemical structure as our local samples. The example also shows that if there is nothing in human practice that rules out XYZ as being the referent of 'water', then English speakers from Earth may call XYZ 'water' without being in error (it doesn't matter whether they are brought to Twin Earth or the twin-water is brought to Earth). If, on the other hand, it is already fixed by our practice that 'water' refers only to H_2O, then such speakers would be in error. If our practice does not fix this, it does not follow that it is an error to exclude XYZ from the extension of 'water'. Rather, if past practice does not legislate on the issue, there is no fact of the matter as to whether XYZ is water or not, and the language-using community may proceed either way without committing any linguistic error.

In this respect, human beings' mental states and their contents appear to be similar to the states of less sophisticated entities, such as the vending machine or the frogs. As Dennett puts it, Panama is Twin-Earth for U.S vending machines. The question when a U.S. machine shipped to Panama stops being a quarter-detector and starts being a quarter-balboa-detector is uninteresting, as there just

doesn't seem to be any real fact at stake. There is no problem in saying of a machine working in the U.S. that it has the function of detecting quarters, or that one in Panama has the function of detecting quarter-balboas, but it seems silly to expect that we must always be able to say of a machine whether it is one or the other. Similarly for frogs. Suppose that some frogs are raised in captivity and are fed on artificial food pellets launched at them by their keepers. If one is to assign a function to the frog's visual system, it is most natural to assign it the function of detecting pellets. Proficiency in the detection of the pellets is, after all, what ensures survival in this environment. For a frog that has been introduced to this environment from the wild, it does not seem that there has to be an answer to the question as to when it stops detecting flies and starts detecting pellets.

Dennett infers from such observations that the contents of human mental states are just as derived and subject to indeterminacy as the contents (or 'contents') that we ascribe to frogs and vending machines.[34] He welcomes this conclusion as it supports his idea that there is no real distinction between the sort of content that we have and the apparently second-rate, non-genuine, merely derived content that we ascribe to such relatively unsophisticated entities. That our contents are potentially indeterminate does not show that they are derived, let alone that they are derived in the same way as those attributed to other biological organisms and artifacts, on the basis of function. More importantly, there are disanalogies between these cases which point to a deep flaw in teleological theories of content.

When frogs, vending machines, or human beings are transported out of their respective environments to others where their established states will be tokened in

---

34. Dennett, Darwin's Dangerous Idea (New York: Simon and Schuster, 1995), p. 411.

the presence of objects which superficially resemble the objects which formerly caused them (their respective 'Twin-Earths'), Dennett expects the contents of these different entities to be subject to similar indeterminacies

Consider, however, what happens when the foreign objects are transported into the home environments. When a Panamanian quarter-balboa is inserted into a U.S. vending machine and accepted by it, any acceptable theory of functions will classify this as a failure to perform its function.[35] Now, when edible pellets are launched into the frogs' natural habitat and the frogs detect them and ingest them, is this an error or not? It would be if it is a function of the frogs' visual systems to indicate simply the presence of flies. Dennett appears to think that this would be the correct way to describe this case, for he says, in response to Fodor's objection, 'We can use the frog's environment of selection... to distinguish between the various candidates [for interpretation of the frog's states]'.[36] In the environment in which the frogs have been selected, they have encountered only flies.

Dennett willingly grants that there is the potential for indeterminacy, as when more pellets are introduced into the environment, selection for detection of pellets may begin to occur. Dennett points out that far from being a flaw in the account, this is a desirable feature, for it provides for the possibility of a change in function subsequent to a change in the environment. I take it that what Dennett has in mind is that when a system is capable of making what we are inclined to describe as errors, it is more flexible and hence adaptable to changes in the environment. What would be failures to perform the function in one environment, would contribute to survival and hence be functional in a different environment.

---

35. Acceptable, that is, for the purpose of serving as the basis for a theory of content.

36. Ibid., p. 408.

However, if we are following Wright's analysis, these cases are to be treated differently. The U.S. vending machine that accepts a Panamanian coin is failing to perform its function, while a frog that snaps at an edible pellet has a visual system that is performing its function. The reason for this difference is simply because in accepting the foreign coin, the vending machine is not doing something which contributes to its survivability, while in snapping at the pellet, the frog is. One may agree that the vending machine in the described circumstances is failing to perform its function, but hold that if the frogs have been selected for fly detection, then either the frog is in error, or that it is at least indeterminate whether it is an error or not. I believe that this is a mistake.

According to the Wright analysis, a thing's function is an effect of it that has increased, or does increase, the probability of its survival. This being the case, there are no grounds on the basis of this analysis for declaring the states of the frogs' visual systems that direct them to snap at artificial pellets to be in error. The relevant states of the frogs' visual systems, let us say, are sensitive to objects of type *K* (which includes both flies and non-flies, such as the pellets). Detection of some of the things in class *K* is detrimental to the frogs. Were the zoo keepers to launch cyanide pellets past the frogs, then the frogs' detection and ingestion of the cyanide pellets would lead to their deaths. In conjunction with the reasoning above, Wright's analysis leads us to regard the members of the subset of *K* which have a detrimental effect on the frogs as not being in the class of things that the mechanism has the function of detecting, *K*'.[37]

---

37. Teleological theories claim that *K*' constitutes a type, some instances of which do cause the representing state. As the frogs will fail to detect some of the members of *K*', so *K*' is not a subset of *K*, instances of the type of object that *K*' constitutes are to be found amongst the members of *K*.

The class *K'* will be more extensive than one may anticipate. Consider something the detection of which would be harmful to the frogs, but will not kill the frogs off quickly. For example, suppose the zoo keepers launch carcinogenic, but nutritious, pellets past the frogs. It is, perhaps, not clear whether such things are to be included in *K'* or not. Although the frogs who detect and eat the carcinogenic pellets will not die immediately, they will fare less well than ones who don't eat them. There will, in other words, be selection against seeking the pellets over time. An application of Wright's analysis would, therefore, lead one to suppose that the carcinogenic pellets are to be excluded from the favored class, *K'*. However, suppose that the zoo keepers switch the frogs to a carcinogen-only diet. Those frogs whose visual systems are not able to detect the carcinogenic pellets will quickly die, while those frogs who can will probably reproduce before they die of cancer. This suggests that the notion of what is harmful or beneficial to an organism, and consequently what is to be included in or excluded from the favored class *K'* cannot be specified absolutely, but only relative to other features of the environment.

I claim that if any member of *K* is to be put into the favored class *K'* owing to its detection having had some survival value for the frogs, then everything in *K* which either has had or would have some survival value for the frogs, owing to their detection of it, also belongs in *K'*. Teleologists will probably wish to object to this claim on the grounds that frogs' visual systems have been selected for their past success at detecting flies, rather than food pellets (carcinogenic or otherwise). The idea is that in attributing content to a particular state one has to look at its causal history. Whatever type of thing its sensitivity to has led to the success of the organism will be the content assigned to that state. This is a central element of the view, and I do not think that it can be maintained.

The appeal to actual selectional history to rule out the 'hypothetical' cases (such as food pellets rather than flies) as being in the extension of a representing states is of no help. This is because it amounts to no more than the claim that it is the type of object that has actually caused the state that constitutes its content, rather than the type that includes the 'hypothetical' cases. This begs the question, for the theory of functions itself is supposed to determine what type of object is in the extension of the state. The problem is that the appeal to teleology is supposed to abstract a certain type of object, *K'*, from the class *K* of all the things that would cause the state in question to be tokened. The only criterion for this, if we are using Wright's analysis, is contribution to the explanation of the existence or persistence of the entity in question. On this criterion, the 'hypothetical' objects are just as much members of *K'* as are flies that the frogs have not yet encountered.

An attempt to restrict *K'* to just flies is likely to result in the extension of the state turning out to contain no more than the objects that have actually caused it, with a consequent contribution to the explanation of the continued existence of the frog. This is, of course, not the class of flies, but a proper subset of that class. This result would be disastrous, for then each new fly would count as an error (even if it has been seen before, but not caught and eaten) although as soon as its detection has contributed to the explanation of the success of the frog it would be part of the extension.[38] What is desired is for the content of the state to be the class of all flies, including the 'hypothetical' ones that never happened to pass by the frogs. There seems to be no non-arbitrary way of including these 'hypothetical' cases without also including the hypothetical ones.[39]

---

38. I don't know how long it takes for an ingested fly to become part of the explanation of the (continued) existence of a frog.

39. The teleologist could appeal to the notion of biological species at this point. This has at least two unwelcome consequences. Firstly, it makes reference relative to our particular way of choosing to individuate species, and secondly It restricts determinate

Perhaps teleologists will not find the forgoing too troubling. Perhaps, after all, the content (or 'content') of the relevant states of the frogs' visual systems is properly understood as being the class of "airborne food item[s]".[40] Teleologists are not entitled to be quite so sanguine. The Wright analysis considers an effect to be functional when it contributes to an explanation of the thing that it is an effect of. The sensitivity of states of a mechanism to things of type *K*' may be explanatory of the mechanism's success, but nothing makes it necessary that the explanation be the same for each member of *K*'.[41] This being the case, not all the things that the frogs catch need be food items. Perhaps pellets containing steroids or other performance-enhancing drugs would be detected by the frogs. The ingestion of such things would make some contribution to the explanation of the success of the frogs, in addition to that provided by its ingestion of food, and so the relevant state of the frogs would include these things in its extension too.

Consider again the vending machine discussed above. It was noted that, unlike the case of the frogs snapping at food pellets, a U.S. vending machine that accepts a non-quarter will be, on the basis of the Wright analysis, failing to perform its function. Whether this is true or not depends upon the nature of the non-quarter. What makes a U.S. quarter genuine is the fact that it has been legitimately produced either by the U.S. Mint, or that institution's authorized agent. Naturally, it is extraordinarily difficult to discern directly whether a piece of metal has this sort of causal history, which is why vending machines just test for size and weight, a certain combination of which factors is highly correlated with genuine U.S. quarters.

---

reference to biological species. Biological species aren't really any more of a natural kind than anything else.

40. Ibid., p. 410.

41. Not even Millikan's requirement that it be a Normal explanation necessitates this.

If I have access to the right sort of equipment, I will be able to produce pieces of metal that are all but indistinguishable from genuine U.S. quarters. If I insert these very good fakes into a vending machine, it will accept them. Given that the people who empty the machine are unable to distinguish my very good fakes from genuine quarters, the machine will flourish even if I am the only one who uses it, and my very good fakes are the only coins it collects. In this situation, the Wright analysis implies that the function* of the machine is to collect not just quarters, but also very good fake-quarters.

I have so far argued that the type *K'* that teleologists favor as the content of the representing states of organisms or artifacts should include those objects which the relevant mechanism is capable of detecting and which would contribute to the explanation of its success were it to do so. As a consequence, acceptance of Wright's analysis commits teleologists to including some objects in *K'* which they wish to exclude. Even objects which the mechanism is not capable of being detected should also be included in *K'*. Recall that some members of the class attributed as the extension of a representing state will fail to cause a tokening of that state. Not only are dogs on a dark night mistaken for cats, for example, but sometimes cats on a dark night are mistaken for dogs, and so fail to cause 'cat' tokens. Nevertheless, a teleological theory would aim to place those cats that fail to cause tokens of 'cat' in the extension of 'cat' because, it is hoped, the theory will yield the class of cats as the extension of 'cat'. This is because the theory is supposed to show that cats are the type of object represented by 'cat' tokens. However, there must be some principle that permits extending the extension from the things that do cause the tokens to the ones that do not.

The only principle that is available is that given by the theory of functions. When the relevant state of an organism or artifact is tokened, the entity produces some sort of response. Having detected the presence of things of type *K*, frogs

attempt to catch the *K* things with their tongues, while vending machines dispense cans of soda. Given this, we should expect that *K'* is the type of object in the presence of which it is advantageous to the organism or artifact to put forward its usual response. Without any other non-arbitrary way to include things within *K'* that pass undetected, the content of a given state may be inflated to near-vacuous proportions. For example, anything that the owners of the vending machine would be willing to accept in return for their soda would count as being of the same type as the quarters that the machine actually accepts. Depending on the interests of the owners, this will be a rather large class, including such things as precious metals, rare postage stamps and even foreign currency. Similarly for frogs, anything that would increase the survivability of the frogs when ingested, whether it comes in fly-like form or not, would count as being of the same type as the flies that they do ingest to their benefit. In fact, it is not even necessary that the benefit to the frogs results from their ingestion of the fake-flies. A fake-fly that was non-harmful to the frogs but had the benefit of improving their skills at catching real flies, a sort of target-practice device, would also count as being of the same type as real flies, on the Wright analysis.

It is important to note that these things count as being members of class *K'* even before frogs or vending machines come into contact with them. A defender of the teleological analysis may be willing to concede that the detection of these things is the function of the entities in the above examples, but only because they enter into the causal history of the tokening of the relevant states of those entities. My point, however, is that the even the 'usual' cause of the states, combined with the Wright analysis, is insufficient to determine the functions that we actually ascribe.

Conclusion

I think it is clear that the Wright account fails as an analysis of 'function'. This is not merely because of the logical difference, noted above, between saying that a thing exists because of what it does, and saying it exists in order to do what it does. The most simple reason for the failure is because the functions* that the Wright analysis attributes to things diverge from the functions that we ordinarily ascribe to them, as in the case of the vending machine that has the function*, but not the function, of collecting fake quarters.[42]

The problem with attempting to base a theory of intentionality on the success of the causal theory of reference is that the fact that tokens of a state are caused by certain things is not in itself enough to specify the content of that type of state. The standard problem of error is a species of this problem, in that solving that problem would be to show why some of the causes of a state are semantically irrelevant. More generally, the problem is finding some principle that will determine the extension from the class of things that actually causes the state to be tokened. In the specification of linguistic reference, this is done by theory. For example, in demonstrating the extension of 'water' by pointing to a glass of water and saying 'That stuff is called "Water"', I will be relying on some theory as to what counts as being of the same stuff as the sample in the glass. Some analogue to the role played by theory here has to be found in a theory of intentionality. The teleological proposal is the most promising attempt to fill this lacuna.

The appeal to selectional history is supposed to determine the extensions of mental states that will serve for a theory of intentionality. However, teleological

---

42. One might suppose that the owners of the machine are just wrong about its function, as after all if the fakes are so good, they will still prosper. However, if the fakes are good enough to fool them but not the bank (and they have sufficiently many other vending machines not to go broke), I think they would be strongly inclined to disagree.

theorists tend to have a prejudice in favor of the categories that we standardly use in assigning contents. Of the traditional objections associated with teleological theories, this is closest to the charge that teleology involves illegitimate anthropomorphism. The problem is that the theory of functions has to project from the actual causal history of a state to a type of object which serves as its extension. It isn't enough for the teleological analysis to just pick out from amongst our standardly used categories which ones are to be preferred as the content of a state. This wouldn't serve as the basis for a theory of content, for it offers no account of how the contents it assigns are determined.

Given the divergence between function and function*, in so far as it is the functions of entities that are relevant to intentionality, a theory of functions* provides no basis for a theory of content. It does seem to be functions that are required, for when we adopt the intentional stance, as Dennett puts it, it is functions, not functions* that we ascribe to other creatures and to artifacts. Furthermore, an attempt to ground a theory of intentionality on functions would be unsuccessful, because 'function', unlike the inadequate function*, is itself an intentional notion.

CHAPTER 6

INSCRUTABILITY, RELATIVISM AND REALISM

One reaction to the troubles outlined above for a naturalistic theory of content would be to hope that some version of these theories will be made to work. Perhaps one may. Although none of these theories is successful in determining reference, some of them are successful in determining something, which may provide the basis for some future theory of content. For the time being, one has to face the serious possibility that reference, as Quine has argued, is inscrutable.

This conclusion is highly counter-intuitive. The main reason for this, I believe, is that our conscious thoughts appear to us to be fully determinate. From the first person perspective, it seems just obvious that reference is scrutable. Searle has argued that this evidence of consciousness shows that Quine's inscrutability argument is in fact a refutation of his linguistic behaviorism. This won't do. Searle is correct to point out that Quine's doctrine of ontological relativity is misleading in its anodyne suggestiveness, but one cannot appeal to the determinateness of mental content as an answer to concerns about inscrutability in the absence of a naturalistic account of reference for mental states. I have argued that the difficulty in providing such an account is considerable.

Another reason for resisting this conclusion is that it may be taken to imply some form of anti-realism. Putnam has argued for some time that the prospects for a naturalistic account of reference are dim and he uses this as the justification for his 'internal realism'.[1] Whether or not Putnam's internal realism follows from his arguments against the possibility of providing a causal theory of reference, it does

---

1. Putnam's arguments for internal realism first occur in his Meaning and the Moral Sciences. Later developments of these arguments are to be found in Reason, Truth and History, The Many Faces of Realism and Representation and Reality.

not follow from the arguments I have presented above. That there are multiple adequate interpretations for our language does not in itself imply anti-realism concerning that of which we speak. In order for that consequence to follow, the interpretations must be incompatible. In my defense of Quine's inscrutability thesis, this is not the case. Different ontologies are not ascribed to speakers of the language on each interpretation, only different satisfaction conditions for the sub-sentential elements of the language. Putnam challenges 'metaphysical realists' to provide their one true description of things as they are in themselves, hoping to point out that this is a mere prejudice and that there is no such one true description, and so no way that the things in themselves truly are. It may be that there is no one true description, but so long as all the adequate descriptions are compatible, anti-realism does not follow.

Putnam's Arguments Against Causal Theories of Reference

In Meaning and the Moral Sciences, Putnam considers the following way of defining reference.[2] Suppose there is a speaker who is reliable in that his utterances have a high probability of being true. If one were to give a truth definition for this speaker's language, the assignment of objects as referents to the terms would determine which of his sentences were true. One way to attempt a definition of the reference relation is to define it as the relation which gives a maximally charitable truth definition for this reliable speaker. One of the reasons Putnam gives for rejecting this suggestion is that what one refers to depends on "the global structure of his 'linguistic' (and inductive) behavior"[3].

---

2. Hilary Putnam, Meaning and the Moral Sciences (London: Routledge and Kegan Paul, 1978), 39.

3. Ibid., 40.

What Putnam means by this is suggested by the example he gives of whether Bohr was referring to the same entity by 'electron' in 1904 as he was some 30 years later following his advancement of the quantum theory. Settling this question will involve appeal to 'the principle of the benefit of the doubt'. Bohr in 1904 has a number of beliefs he expresses using the term 'electron' and in 1934 he has a number of significantly different beliefs he would express using the same term. If we define 'electron' as that thing to which all Bohr's beliefs apply in 1904, then that term fails to refer. In deciding whether Bohr was referring to the same thing as in 1934, or to anything at all, we have to employ what Putnam calls 'the principle of the benefit of the doubt' in determining which beliefs are crucial to determining the reference of the term and which are not.

When a speaker specifies the referent of a term by description, and that description fails to refer owing to mistaken factual beliefs of the speaker, Putnam tells us that the principle of the benefit of the doubt directs us to assume that the speaker would accept reasonable reformulations of the description. So, although nothing exactly fits Bohr's 1904 description of an electron, electrons partially fit this description. Actual electrons have negative charge as do Bohr's 1904 'electrons', for example. Putnam's point is that in 1904 Bohr would not have insisted that electrons must have all the properties that he then thought they did. He would have been willing to accept that he was wrong about some of the properties of electrons. Only if it turned out that nothing came at all close to his 1904 description, would we say that there are no electrons. This happened with phlogiston, for example. As Putnam points out, Bohr himself must have afforded his earlier self the benefit of the doubt, as he did not introduce a new term for the particle described by the quantum theory.

The problem raised by this for giving a definition of reference is that of finding a precise account of which redescriptions someone would accept for those

terms for which reference is fixed by description. For some redescriptions, it is reasonable to suppose that they would be accepted. For others, this supposition would be unreasonable. To give an account of what it is reasonable to count as important in determining the referent of the term, says Putnam, would be to give an account of full human capacity. For the same reason, Putnam holds that it is unrealistic to expect to find an algorithm for translation.

Putnam suggests that translation is largely a matter of 'rationalizing behavior'. This means that the translation manual for an individual will specify beliefs that, together with the desires of the individual, will yield good explanations for his behavior. Putnam goes on to suggest that the interest-relativity of explanation accounts for the truth of Quine's thesis of the indeterminacy of translation.

In order to demonstrate this, Putnam first establishes that a request for explanation presupposes some set of interests. His example concerns a professor X who is found naked in the women's dormitory at midnight. A bad explanation of this phenomenon, says Putnam, is that X was in the dormitory at midnight - ε and that X could neither leave the dormitory or get dressed by midnight without exceeding the speed of light. The sort of explanation we are interested in, says Putnam, concerns the psychological causes of the state of affairs. Putnam also mentions Alan Garfinkel's example of the bank robber who responds to the question 'Why do you rob banks?' with the answer 'That's where the money is'. Garfinkel points out that this answer would be satisfactory if the question were asked by another robber, but not if asked by a (non-robber) priest. This is because a why-question presupposes a range of relevant alternatives. The robber would really be asking 'Why do you rob banks rather than other places?' whereas the priest would really be asking 'Why do you rob banks rather than lead a law-abiding life?'

Putnam attempts to use these facts about explanation requests requiring a context to lend support to Quine's thesis of the indeterminacy of translation. Consider two interpretations of a speaker's language, Int1 and Int2. Int1 interprets the speaker as referring to rabbits with 'gavagai' and Int2 interprets him as referring to undetached rabbit parts. Quine's claim is that there is no fact of the matter as to which of these interpretations is correct. Applying his notion that translation is a matter of explaining behavior, Putnam argues that the indeterminacy thesis has the air of implausibility that it does owing to the particular interests we bring to the explanation of behavior. Given these interests, we find the simplest explanation of someone who catches a rabbit, say, that he wants a rabbit to eat, believes that he sees a rabbit, and shoots at the rabbit he sees. The alternative explanation, that he wants some undetached rabbit parts in order to detach and eat them, believes he sees some undetached rabbit parts and shoots at said parts, seems 'absurd to us, *given the way we structure the explanation space*'[4].

Although our interests incline us to prefer the interpretation of 'gavagai' as referring to rabbits rather than rabbit parts, and so there may be little indeterminacy 'for us', beings with different interests in explaining behavior may be expected to favor a different interpretation. As an example, Putnam asks us to consider a race of tiny Martians whose language does not contain simple expressions for objects such as rabbits, but instead has simple expressions which refer to undetached rabbit parts. Such Martians, he suggests, would find Int2 a more satisfactory interpretation of Inf's language.

The attempt to argue from the alleged interest-relativity of explanation to the truth of the indeterminacy thesis is not very compelling. Putnam has not given us good reason to think that explanation is interest-relative. As Michael Devitt has

---

4. Ibid., 44.

pointed out, Putnam's examples show not that the value of an explanation is relative to interests, but that explanation requests are often elliptical. This is explicit in the example of the two ways of clarifying the explanation request put to the bank-robber. When it is made clear which why-question is being asked, it is clear that the robber's answer is relevant to the one question but not to the other. Similarly, when we ask why Professor X was found in the girls dormitory, we are asking for the events which are causally responsible for his being there. As Devitt puts it, '[W]e *have* interests which make us ask *that* question, but a good answer to it is good absolutely...'[5]

Furthermore, whatever interest-relativity of explanation that obtains does not show the indeterminacy thesis to be true. It may be, as Putnam suggests, that diminutive Martians would prefer an interpretation of Inf's language which assigns undetached rabbit parts as the referent of 'gavagai', but this does not show that any such interpretation is adequate. The most these claims concerning the interest-relativity of explanation could show is the basis of the resistance to the indeterminacy thesis. It may be that the naturalness of Int1 given our interest in rabbits makes it very difficult for us to even seriously entertain Int2 as a possible interpretation. If true, however, this does not show that Int2 is a possible interpretation.

Putnam appeals again to the argument that meaning and reference depend upon 'discounting differences in belief' in Representation and Reality. It would be extraordinary for two beings to share all the same beliefs and so it is practically impossible in interpretation to ascribe all of one's beliefs to the one being interpreted. This being the case, one will construe someone as referring to, say,

---

5. Michael Devitt, Realism and Truth (Princeton, NJ: Princeton University Press, 1984), 184.

gold with a given term even though on the full interpretation the person has different beliefs to oneself about gold. In practice, Putnam suggests, we count two beings as meaning the same thing or referring to the same thing by a given word according to some intuitive judgment of reasonableness. According to Putnam, this presents a problem for a physicalist/functional reductionist theory of reference, for it will have to formalize these judgments of reasonableness. His point is that it is unrealistic to expect to formulate a definition of coreferentiality on this basis. This, he says, would be no easier than to 'survey human nature *in toto*'.[6]

One might suppose that this sort of objection applies only to establishing the coreferentiality of theoretical (or nonobservational) terms. Putnam invokes Quine's thesis of the indeterminacy of translation to show that even for what we consider to be observational terms, such as 'rabbit', where each speaker is able to distinguish whether the term applies to an object, there can be a problem in judging synonymy. He notes that stimulus synonymy is not a necessary condition for synonymy, even in the case of observational terms. A Thai speaker, for example, may not associate the same stimulus meaning with 'meew' that an English speaker does with 'cat', but this does not mean that 'meew' is not coreferential with 'cat'.

It is clear enough that it will not always be the case that two beings who are counted as having coreferring expressions in their language will be in absolutely identical functional states. Still, it may be hoped that it is possible to define the relation of coreferentiality such that the different functional states that such beings would be in would be in the same equivalence classes under that relation.

---

6. Hilary Putnam, Representation and Reality (Cambridge, MA: MIT Press, 1988), 75.

Putnam's reasons for thinking that this hope is unlikely to be realized are that to interpret even the simplest words in a language requires interpreting the discourses as wholes and that the range of possible discourses encompasses all possible beliefs that sentient beings may form, including every possible scientific and mathematical system that such beings may formulate. This is to do more than simulate an ability that human beings actually have, for no human being could hope to grasp all possible beliefs of all possible sentient beings.

A Lacuna in the Causal Theory

It may be thought that the situation is not quite so grim. Putnam's pessimism is based in part on an assumption of holism, and some, such as Fodor, have used a causal theory of reference to support an atomist position. If Putnam were relying on the truth of holism in his arguments against the possibility of formulating a causal theory of reference, he would be in effect begging the question. The essential point may be made without assuming the truth of full holism, however.

For externalists, direct causal connection to things is not necessary in order to be able to refer to them. Assuming that the word 'electron' refers, I am able to refer to electrons by using it owing to *some* suitable causal connection via textbooks and their authors to experiments involving electrons. If there are any extraterrestrial beings, I am able to refer to them by the term 'extraterrestrial' not owing to any causal connection between my use of the word and such beings, but because I can describe them in terms that do refer owing to causal connections to their referents, such as 'Beings that do not originate from this planet'. As Putnam points out, this lack of actual causal connection obtains even with simple words

such as 'rabbit' or 'cow'.[7] The extension of 'rabbit' includes not only those rabbits that stand in the right causal relation to the present use of the term, but all other rabbits. This shows that it won't do to suppose that the content of a term may be simply determined, such as being given by just the objects that have caused it in the past. It must be an object of a specific type, that includes objects that the speaker has had no causal interaction with.

To suppose that the extension of a term consists of the class of things of the same kind as the things which have caused it (in some special way) is not sufficient, for as Putnam notes, "'of the same kind' makes no sense apart from a categorial system which says what properties do and what properties do not count as similarities. In some ways, after all, anything is 'of the same kind' as anything else."[8] Even if one grants that causation plays a central role in the theory of reference, one must recognize that theory, and hence the beliefs of the speaker, enters into the specification of what type of object is causally related in the right way to the term. These judgments of agreement and disagreement are how we project from the objects which cause speakers utterances to the classes which are the extensions of their terms.

The criterion of adequacy for a theory of reference is agreement with our ordinary judgments of linguistic reference. I assume that Putnam's account of how linguistic reference is specified is largely correct. This means that causal connections play some role in determining reference. However, a causal connection between a term and an object of a certain type is not in itself sufficient to determine that the term refers to objects of that type.[9] In specifying what the

---

7. Hilary Putnam, Reason, Truth and History (Cambridge: Cambridge University Press, 1981), 53.

8. Ibid.

9. Going to particulars doesn't help either, owing to the indeterminacy of ostension.

terms of their language refer to, speakers rely on theories about the referents. Consider Putnam's example of 'water'. Speakers of English are generally able to reliably identify certain liquids as water. Putnam suggests that the reference of 'water' is to be specified as whatever bears the same-liquid relation to the liquids commonly identified as such. We are presently in the position of being able to say something in detail about what this relation is, for our chemistry is sufficiently advanced for us to distinguish between the structure of water (H_2O) and other liquids. Those who don't know enough about chemistry to employ this theory may nevertheless succeed in referring to water owing to what Putnam calls 'the linguistic division of labor'.

It would be highly undesirable to insist that speakers' being able to employ a sophisticated theory about the nature of the referent is necessary in order to determine reference. This is because such a requirement would mean that the word 'water' did not refer to water before our civilization developed an advanced enough chemistry to discern its composition. It is very plausible to suppose that English speakers before 1750 did refer to the substance H_2O by their term 'water'. The way to account for this would be to suppose that speakers employ some, at least tacit, theory concerning the nature of water. They no doubt assumed that it had some sort of structure, and that substances which superficially resembled it but were structurally very different (which differences would manifest themselves under some, perhaps very uncommon, circumstances) would not be the same type of substance. This condition is of course considerably vaguer than the more precise condition we currently employ in specifying the reference of 'water', but is probably sufficient to enable speakers of English before the development of

modern chemistry to agree with the moral of Putnam's twin-earth story.[10] There is some degree of vagueness in our condition for two substances being the 'same liquid', too. For example, it is unclear whether such subtle differences as liquids composed of different types of hydrogen isotopes and oxygen should count as different types of water or as liquids which are not water.

The situation is similar with respect to biological species. When specifying the reference of a term such as 'cat' one may take an exemplar and specify that the extension of 'cat' is comprised of all entities of the same type as it. In "The Meaning of 'Meaning'", Putnam claims that should our exemplars turn out to be non-animals but cunningly contrived robots, then it would have turned out that no cats are animals. If, in fact, the relation that we use in specifying the reference of 'cat' is the same-animal relation, this will not be the case: our term 'cat' would fail to refer. That issue aside, some account has to be given of what sort of relation speakers take the same-animal relation to be. A contemporary answer to this question is most likely be given in terms of evolutionary history. In so far as terms like 'cat', 'tiger' or 'whale' are socially deferential, this condition will apply to the terms for biological species used by speakers generally.

Clearly, this sort of condition would not be part of even the experts' specification of reference prior to Darwin. This does not in itself imply that the meanings of terms for biological species have changed, for that would depend upon whether the theory is incorporated into the contemporary meanings, which it probably isn't. Nor need it imply that the reference of these terms has changed, so long as the new theory picks out the same classes as the old theory. I think it's

---

10. The issue is not whether speakers before 1750 would have called XYZ 'water'. It is enormously probable that they would have, had they been exposed to it. The issue, rather, is whether they would be mistaken to do so. I am suggesting that they would be, owing to their, at least tacit, theory (or proto-theory) of water.

unlikely that the old theory did pick out exactly the same classes, however. It's not that reference for such terms used to be in terms of superficial characteristics such as the description theories that Putnam criticizes. Presumably, speakers prior to Darwin thought that the internal constitution of a biological entity played some significant role in determining what species it belonged to. Probably, however, they would see a case of parallel evolution where two creatures had the same physical constitution but different evolutionary history as the production of the same species at two different times or places rather than as two different species.

The Failure of the Causal Theory

It is clear that what theory speakers of a language use to class some things together and some not is not directly connected to the intrinsic nature of the things themselves. It would be possible for speakers to use different grouping schemes than the ones that we have settled on. In fact, past speakers of our language actually did, as some contemporary grouping schemes depend upon scientific theories of relatively recent vintage. Some possible schemes seem very odd to us, such as the predicates Goodman uses to present his 'new riddle' of induction. The most well known of these is 'grue'. There is a disjunctive condition on membership in the class of grue things. Something is grue if and only if it is either green and first observed before Jan 1st 2001, or blue and first observed after Jan 1st 2001. If we wish to account for the content of a representation in terms of its usual cause then, before the time specified in the definition of 'grue', all the representational states caused by green things are also caused by grue things. So, it might be suggested, there is just as much reason for a causal account of content, formulated before this time, to interpret what appears to be a representation of greenness as a representation of grueness.

It might be objected to this suggestion that although the representation is caused by some of the things that are grue, it will not be caused by all of the things that are grue. In particular, it will fail to be caused by grue things which are also blue. The problem with this objection is that no representation will be caused by all the things that it is supposed to represent. For all things, there are circumstances in which they appear to be something other than what they are. The difference is that the grue things that do not cause the representation we are inclined to interpret as referring to green things are not failing to cause it because they are mistaken for something else. This is no comfort to those who appeal to a causal theory in order to scrutinize reference, for we are presently attempting to reduce intentional notions such as recognition and replace them with purely naturalistic relations.

It is not possible (at least not obviously) to use the 'grue' example to argue for the indeterminacy of reference in the fashion of Quine's argument in 'Ontological Relativity'. This is because an interpretation of 'green' as 'grue' will not leave speakers' overall dispositions to verbal behavior unchanged. This is not what I am suggesting, however. What is being proposed is that a causal theory of reference is unable to explain why our terms do not have this interpretation. That is, the causal theory of reference does not explain why we have the speech dispositions that we do.

The fact that each thing is at once indefinitely many different types of object shows the hopelessness of Dretske's original solution to the disjunction problem. Recognizing that the content of a type of state cannot be identified with the information it actually carries, as it is sometimes tokened in error, Dretske wished to identify content with what causes the tokening of the state in some specific period, the 'learning stage'. The idea is that by keeping the student away from error-inducing causes during this special period, one can fix the content of the state

as the information it actually carries at the end of the period. There are numerous problems with this suggestion. Not least amongst these problems is that the notion of information, as Dretske wishes to use it, is not objective.

Even if, in trying to teach some student speaker a term for dogs, I manage to prevent my student from observing cats on dark nights and mistaking them for dogs, this does not mean that the student's term 'dog' carries only the information that a dog is present. Although the student's term 'dog' has only been brought forth in the presence of dogs, it has also only been brought forth in the presence of dog-or-cats. One can appeal to the teacher's intentions to rule this out as what has been taught, but that would of course be unacceptable in an attempt to ground a naturalistic theory of reference. Neither may one appeal to the causal sensitivity of the term to dogs. The very problem is that the term, even after a period of learning, is causally sensitive not only to dogs, but other things, including cats seen on dark nights.

Not only does Dretske's appeal to a learning period to fix the information carried by a term (or mental state) fail, any attempt to solve the disjunction problem by relying on the information actually carried under any circumstances is bound to fail for the same reason. If the class *K* contains every individual that causes a state, that the state appears to carry information other than that there is a member of *K* present is an illusion brought on by our tendency not to consider other categorial schemes than the familiar one. If one trained a system to respond to its environment in the way Dretske describes, one may be inclined to interpret some of its states as representing dogs, but this is not an objective feature of the relation between the system and the environment. The fact that the causal proclivities of the system would be identical whether it had 'mistakenly' encountered cats or not, shows that Dretske's learning period is irrelevant to what

the content of the state is, considered independently of the intentions of the instructors.

Instead of a theory based on what information the system does or has carried, instead one needs to look for a theory based on what information the system should carry. So a successful naturalistic theory of intentionality is most likely to involve some appeal to teleology. Unfortunately for the prospects of a naturalistic theory of content, naturalistic accounts of teleology are also based on what information a state has historically carried. The basic account is that the content of a particular type of state, *S*, of a mechanism, *A*, does not consist in the information that tokens of it currently carry, but in the information that past tokens of *S* have carried, where those tokens carrying that information contributes to the explanation of the present existence of the mechanism *A*.

The teleological objection to the suggestion that, so far as a causal theory of reference goes, a representation of greenness may be interpreted just as much as a representation of grueness is that it is the greenness of the things represented and not their grueness that is responsible for the success of the organism. Yet the green things observed before Jan 1st 2001 are just as much grue as green and whatever advantage it is that sensitivity to green things confers upon the organism, sensitivity to grueness confers that same advantage upon it. Whatever is so good about greenness, grueness, at least before Jan 1st 2001, is just as good. After Jan 1st 2001, grueness will not be as good, because then many grue things will be blue. So although the grueness interpretation may explain past success, it won't explain future success. On the views in question, however, it is only the causal history that counts for determining content.

A different way of making the point that it is sensitivity to greenness and not grueness that explains the success of the organism is to point out that had its environment been different (i.e. had the organism lived at a later time) the grue

things would be blue, not green, and the organism would fail to detect them. In the environment of the organism (i.e. before Jan 1st 2001), grueness is effectively correlated with greenness.[11] This means that *in the organism's environment* the representation does indicate grueness. Any adequate teleological theory will need to make use of this notion of indication-in-an-environment, so the above appeal is not conclusive. Had the organism's environment been different the grue interpretation would not fit the teleological theory, i.e. it would not be a good way of explaining the success of the organism, but it is the organism's actual environment that matters for this.

Unfortunately for the prospects of a naturalistic theory of content, naturalistic accounts of teleology are also based on what information a state has historically carried. The basic account is that the content of a particular type of state, *S*, of a mechanism, *A*, does not consist in the information that tokens of it currently carry, but in the information that past tokens of *S* have carried, where those tokens carrying that information contributes to the explanation of the present existence of the mechanism *A*.

The teleological proposal is an improvement on the suggestion that the content of a type of state should be identified with the information actually carried by its tokens at a particular time, for it adds the additional requirement that the information carried should contribute to an explanation. This has the effect of narrowing the class assigned as the content of the state. Suppose that a state, *S*, has been caused by both *F*s and *G*s. Further suppose that we are inclined to ascribe a content of representing *F*s to this state. The standard disjunction problem is that some non-arbitrary reason must be provided for excluding the *G*s from the

---

11. Grueness isn't actually correlated with greenness in the organism's environment because only observed green things are grue. Green things that will not be observed until after Jan 1st 2001 are not grue.

content of *S*. Dretske's proposal that if there is some special period during which *S* is caused only by *F*s then the content of *S* is *F* fails, because during this time *S* is still caused by *F*-or-*G*s.[12] The teleological solution is that if *S*'s being caused by *F*s contributes to the explanation of the mechanism of which *S* is a state, whereas *S*'s being caused by *G*s does not, then the content of *S* is *F* and not *F*-or-*G*.

However, everything that is an *F* is also a member of indefinitely many other classes. The teleological theory must show that of all the many candidate classes for the content of *S*, that it is only *F* that may be ascribed as the content. Let the members of the class *H* be those things, besides the *F*s, whose causing S would contribute to the explanation of the mechanism of which S is a state. If *H* is non-empty, then the teleological proposal cannot privilege *F* as the content of *S* over the larger class *F*-or-*H*.

To take a simple example, recall Dennett's example of the soda vending machine. Call the class of objects that the machine accepts in return for soda *K*. Let us suppose that the machine makes money for its owners. If one assumes that before it leaves the factory, the machine is 'trained' to accept only quarters by having genuine US quarters fed into its slot, this is obviously irrelevant to the information carried by the state the machine goes into, *S*, when it accepts an inserted object. Even before it leaves the factory, this is still the information that a *K* thing is present, even though of all the *K* things only quarters have been inserted into its slot. To suppose otherwise, one would have to find some non-arbitrary way of both excluding the non-quarters from the 'interpretation' of *S* and including within it the quarters that are non-*K*. Simply appealing to the things that are of the

---

12. This is not entirely fair to Dretske, as his official proposal is that the content of a state is the information it carries at the end of the learning period. This proposal has very limited application as there must be very few cases where a state has ever carried the information that we attribute to it as its content.

'same kind' as the quarters it has historically accepted will not do, for all the *K* things are of the same kind as those quarters. One could exclude the undesirable members of *K* from the interpretation of *S*, by taking *S* to represent just the class of objects that *S* has historically accepted. Not only would this exclude the genuine quarters that S will not accept, but it also excludes the genuine quarters that have not yet been inserted, whether the machine would accept them or not.

The teleological theory appears to improve matters by requiring that whatever class is put forward as the interpretation of *S* must figure in an explanation of the machine's success. It may appear that the teleological proposal will immediately fail for similar reasons as Dretske's solution to the disjunction problem. The explanation of the machine's success is that, historically, it has accepted objects of kind *K*. This is not the whole story, however. If the machine has accepted *K* things, but *K* things were not correlated with quarters, then the machine would not be successful. A vending machine that was very easy to fool would be removed by its owners before long. The explanation is not just that the machine accepts *K* things, but that in addition there is a high correlation between *K* things and quarters. So the machine is successful because most (or at least enough) of the things it accepts are quarters.

Unfortunately, this won't quite do. Although the things that the machine accepts which explains its success are quarters, this is not the only class those things belong to. Moreover, this is not the only class those things belong to that explains the success of the machine. Some classes one might suggest as an alternative interpretation of *S*, such as *K*, are not as good an explanation of the success of the machine as the class of quarters, for they contain members redundant to the explanation. This is not true of all such classes, however. Anything that the owners of the machine would be willing to accept as payment for their soda, including quarters but also including other things of value such as

precious metals and rare postage stamps, serves at least as well as the class of quarters for the purpose of explaining the continued presence of the machine.

It may appear that this is not the case, for it seems that it is in virtue of accepting a larger proportion of quarters in particular, and not just things that are members of the larger class, that the machine remains where it is. I don't think that this objection is sound. The explanation of the persistence of the machine consists in it accepting a number of objects. These objects have in common, let us suppose, the property of being genuine US quarters. They also have in common the property of being things that the owners of the machine are willing to accept in return for their soda. Some non-arbitrary reason must be given for preferring an explanation in terms of the former property rather than the latter. That, amongst the things that the owners are willing to accept in payment, the only things they have historically received are quarters is not sufficient. If the machine just happens to have historically accepted only quarters that were minted between 1982 and 1999, some reason would have to be given for ignoring this class, in favor of the class of all quarters, as the interpretation of *S*.

The problem is finding some principle that will enable one to expand from the class of objects whose detection has contributed to the explanation of the (continued) existence of the mechanism to some larger class which includes objects which the mechanism has not encountered, some of which it would not detect even if it did encounter them. It is easy to overlook this problem owing to our commitments our own categorial schemes. When one recognizes that the theory of intentionality has to impose its own categorial scheme, the problem becomes acute. For a teleologically based theory of intentionality, the thing that should determine which class is relevant is the teleological condition. Unfortunately for these theories, it fails to impose the categorial scheme which we actually use in attributing both functions and contents. The teleological approach

may give us a reason for attributing one of our habitually used categories as the content of a particular state, but it is not sufficient to determine those categories.

Inscrutability and Realism

Putnam has referred to the phenomenon of alternative categorial schemes as 'conceptual relativity' and has argued for his 'internal realism' in part upon its basis. It would be unfortunate if a consequence of accepting this position is being forced to give up realism. Putnam's discussion of the issue is clouded somewhat by his Berkeleian insistence that he is the defender of commonsense realism, and that he only opposes what he calls 'metaphysical realism'. Putnam describes metaphysical realism as committed to the following claims:

'[T]he world consists of some fixed totality of mind-independent objects. There is exactly one true and complete description of 'the way the world is'. Truth involves some sort of correspondence relation between words or thought-signs and external things and sets of things.'[13]

In asserting the doctrine of conceptual relativity, Putnam takes himself to be giving up the metaphysical realist claim that there is a unique complete description of the world.

Putnam gives the following example as an illustration of conceptual relativity.[14] Consider a world which contains three individuals: x1, x2, x3. This world may be said to contain more than three objects, if one considers the sum of any two particulars to be an object. Ignoring the possibility of 'null' objects, this means that one could describe the simple world as containing as many as seven objects: x1, x2, x3, x1+x2, x1+x3, x2+x3, x1+x2+x3. The answer to the question

---

13. Putnam, Reason, Truth and History, 49.

14. Hilary Putnam, The Many Faces of Realism (La Salle, IL: Open Court, 1987), 18.

'how many objects are there?' depends upon what one considers to be an object. Putnam thinks that this presents a problem to metaphysical realists, as according to him, they are committed to the view that there can be only one correct description of reality, and so must think that one or the other of the descriptions is really correct.

Putnam describes the 'classic metaphysical realist' response in terms of the 'cookie cutter' metaphor. This response is that there is a single world (the dough of the metaphor) which may be divided up into classes of objects in different ways (the cookie cutters of the metaphor). Putnam says that metaphysical realism founders on the question 'What are the "parts" of this dough?'. Whatever answer is given to this question, whether the description given of the world before it is divided into objects is in terms of the three objects or the seven objects, the description of the world will not be neutral, it will have to be given in one or the other of the possible categorial schemes.

I am willing to agree that reality cannot be described independently of some description. The implications for realism of this fact are less clear to me. It is clear that there is no single language in which a description of the world must be given if it is to be complete. This is not true just in the trivial sense that it is possible to formulate the same theory in two syntactically different languages, but also in the sense that one and the same thing may be adequately described by two semantically different languages. The rival interpretations of 'gavagai' constitute an example of two semantically different languages. In the language attributed to Inf by Ling1 there is a term which is true of rabbits. There is no term which directly translates this in the language attributed by Ling2, who interprets 'gavagai' as referring to undetached rabbit parts. Nevertheless, both these languages are suitable for describing situations involving rabbits.

I agree with Putnam that it makes no sense to ask whether a description in Ling1's language or a description in Ling2's language is the right one. Both are capable of providing adequate descriptions of the world. If metaphysical realism is committed to denying this, then it is false. However, the fact that there can be distinct alternative descriptions in this sense does not have any implications for realism. If a speaker of Ling1's language denies that there are such things as undetached rabbit parts, then he is wrong. Similarly, if a speaker of Ling2's language constructs an expression which refers to rabbits, and uses it to deny the existence of rabbits, then he is wrong.

Although the world does not force us to use any particular set of concepts to describe it, and so in one sense there is no unique true description of it, the anti-realist claim, that there is no way the world is independently of the way we conceive of it, does not follow. If a true description in language *X* says that object *o* exists, and if someone asserts in language *Y* that *o* does not exist, then this latter assertion is false, it seems to me. All descriptions, in whatever language they are constructed, must be consistent with all true descriptions in all other languages.

The situation is different if we are faced with two descriptions, each of which we consider to be true, but which are not compatible with each other. To illustrate this possibility, Putnam uses the example of different accounts of the nature of points in Euclidean geometry, and the example of descriptions in terms of quantum field theory or quantum particle theory. There may be arguments from conceptual relativity that would lead one to give up realism. However, the degree of conceptual relativity which it is necessary to admit in order to make the arguments I have employed against naturalistic theories of reference does not compel one to abandon realism.

SELECTED BIBLIOGRAPHY

Antony, Louise. 'Naturalized Epistemology and Language.' In Naturalistic Epistemology, edited by A. Shimony and D. Nails, 235-257. Dordrecht: Reidel Publishing Company, 1987.

Baker, Lynne Rudder 'On A Causal Theory of Content.' Philosophical Perspectives 3 (1989): 165-186.

————. 'Has Content Been Naturalized?' In Meaning in Mind: Fodor and His Critics, edited by Barry Loewer and Georges Rey, 17-32. Oxford: Blackwell, 1991.

Block, Ned. 'Advertisement for a Semantics for Psychology.' In Midwest Studies in Philosophy, vol. X, edited by Peter A. French, Theodore E. Uehling, Jr., and Howard K. Wettstein, 615-78. Minneapolis: University of Minnesota Press, 1986.

Boghossian, Paul. 'Naturalizing Content.' In Meaning In Mind: Fodor and his Critics, edited by Barry Loewer and Georges Rey, 65-86. Oxford: Blackwell, 1991.

Boorse, Christopher. 'Wright on Functions.' The Philosophical Review 85 (1976): 70-86.

Burge, Tyler. "Individualism and the Mental." In Midwest Studies in Philosophy, vol. IV, edited by Peter A. French, Theodore E. Uehling, Jr., and Howard K. Wettstein, 73-121. Minneapolis: University of Minnesota Press, 1979.

Canfield, J. 'Teleological Explanation in Biology.' British Journal for the Philosophy of Science 14 (1963): 285-295.

Chomsky, Noam. 'Quine's Empirical Assumptions.' In Words and Objections: Essays on the Work of W. V. Quine, edited by D. Davidson and J. Hintikka, 53-68. Dordrecht: Reidel Publishing Company (1975).

Churchland, Patricia and Churchland, Paul. 'Response to Dretske's Precis of Knowledge and the Flow of Information.' Behavioral and Brain Sciences 6 (1983): 55-60.

Churchland, Paul. Scientific Realism and the Plasticity of Mind. Cambridge: Cambridge University Press, 1979.

Cram, Hans-Robert. 'Fodor's Causal Theory of Representation.' Philosophical Quarterly 42 (1992): 56-70.

Davidson, Donald. 'The Inscrutability of Reference.' In Inquiries into Truth and Interpretation. Oxford: Oxford University Press, 1984.

————. 'Reality Without Reference.' In Inquiries into Truth and Interpretation. Oxford: Oxford University Press, 1984.

————. 'Knowing One's Own Mind.' The Proceedings of the American Philosophical Association 60 (1987): 441-58.

Dennett, Daniel. The Intentional Stance. Cambridge, MA: MIT Press, 1987.

————. Darwin's Dangerous Idea. New York: Simon and Schuster, 1995.

Descartes, René. Meditations on First Philosophy. Translated by John Cottingham. Cambridge: Cambridge University Press, 1986.

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

————. 'Precis of Knowledge and the Flow of Information.' In Behavioral and Brain Sciences 6 (1983): 55-90.

————. 'Misrepresentation.' In Belief, edited by Radu Bogdan, 17-36. Oxford: Oxford University Press, 1986.

————. Explaining Behavior. Cambridge, MA: MIT Press, 1988.

————. Naturalizing the Mind. Cambridge, MA: MIT Press, 1995.

Enç, Berent. 'Intentional States of Mechanical Devices.' Mind 91 (1982): 161-182.

Evans, Gareth. The Varieties of Reference. Edited by John McDowell. Oxford: Oxford University Press, 1982.

———. 'Identity and Predication.' In Collected Papers. Oxford: Oxford University Press, 1985.

Field, Hartry. 'Logic, Meaning, and Conceptual Role.' Journal of Philosophy 69 (1977): 379-408.

———. 'Mental Representation.' Reprinted in Readings in Philosophy of Psychology, Vol. 2, edited by Ned Block, 78-114. Cambridge, MA: Harvard University Press, 1981.

Fodor, Jerry. 'Methodological Solipsism Considered as a Research Strategy in Cognitive Science.' Behavioral and Brain Sciences 3 (1980): 63-109.

———. 'Semantics, Wisconsin Style.' Reprinted in RePresentations. Cambridge, MA: MIT Press, 1984.

———. Psychosemantics. Cambridge, MA: MIT Press, 1987.

———. A Theory of Content and Other Essays. Cambridge, MA: MIT Press, 1991.

———. The Elm and the Expert. Cambridge, MA: MIT Press, 1994.

Frege, Gottlob. 'On Sense and Meaning.' In Translations from the Philosophical Writings of Gottlob Frege, edited by Peter Geach and Max Black, 956-78. Oxford: Basil Blackwell, 1952.

Godfrey-Smith, Peter. 'Misinformation.' Canadian Journal of Philosophy 19 (1989): 533-50.

Gould, Steven and Richard Lewontin, 'The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptionist Programme.' Proceedings of the Royal Society, Vol. B205 (1979): 581-98.

Hookway, Christopher. Quine: Language, Experience and Reality. Cambridge: Polity Press, 1988.

Hume, David. A Treatise of Human Nature. Edited by L. A. Selby Bigge, revised by Peter H. Nidditch. Oxford: Clarendon Press, 1978.

Katz, Jerry. The Metaphysics of Meaning. Cambridge, MA: MIT Press, 1990.

Kripke, Saul. 'A Puzzle about Belief' In Meaning and Use, edited by A. Margalit, 239-283. Dordrecht: Reidel Publishing Company, 1979.

————. Naming and Necessity. Oxford: Blackwell, 1980.

Levin, Michael. 'Plantinga on Functions and The Theory of Evolution.' Australasian Journal of Philosophy 75 (1997): 83-98.

Lettvin, J. Y., U. Maturana, W. McCulloch and W. Pitts. 'What the Frog's Eye tells the Frog's Brain.' Proceedings of the Institute of Radio Engineers (1959): 1940-51.

Loar, Brian. 'Subjective Intentionality.' Philosophical Topics 15 (1987): 89-123.

————. 'Social Content and Psychological Content.' In Contents of Thought, edited by Robert H. Grimm, and Daniel D. Merrill 99-110. Tucson: University of Arizona Press, 1988.

————. 'Personal References.' In Information, Semantics and Epistemology, edited by E. Villanueva, 117-133. Oxford: Blackwell, 1990.

————. 'Can We Explain Intentionality?' In Meaning In Mind: Fodor and his Critics, edited by Barry Loewer and Georges Rey, 119-136. Oxford: Blackwell, 1991.

Locke, John. An Essay Concerning Human Understanding. Edited by Peter H. Nidditch. Oxford: Clarendon Press, 1975.

Loewer. Barry. 'From Information to Inentionality.' Synthese 70 287-317.

McDowell, John. 'Singular Thought and the Extent of Inner Space.' In Subject, Thought and Context, edited by John McDowell and Philip Pettit, 137-68. Oxford: Oxford University Press.

McGinn, Colin. 'The Structure of Content.' In Thought and Object, edited by A. Woodfield, 207-258. Oxford: Oxford University Press, 1982.

————. Mental Content. Oxford: Blackwell, 1989.

Matthews, R. 'Troubles with Representationalism.' Social Research 51 (1984): 1065-97.

Millikan, Ruth. Language, Thought and Other Biological Categories. Cambridge, MA: MIT Press, 1984.

————. 'Thoughts Without Laws: Cognitive Science Without Content.' The Philosophical Review 95 (1986): 47-80.

————. 'In Defense of Proper Functions.' Philosophy of Science 56 (1989) 288-302.

Nagel, Ernest. 'Teleology Revisited.' The Journal of Philosophy 74 (1977): 261-301.

Neander, Karen. 'Functions as Selected Effects: The Conceptual Analyst's Defense.' Philosophy of Science 58 (1991) 168-184.

Papineau, David. Reality and Representation. Oxford: Basil Blackwell, 1987.

Putnam, Hilary. 'The Meaning of "Meaning."' Reprinted in Mind, Language and Reality. Cambridge: Cambridge University Press, 1975.

————. Meaning and the Moral Sciences. London: Routledge and Kegan Paul, 1978.

————. Reason, Truth and History. Cambridge: Cambridge University Press, 1981.

————. The Many Faces of Realism. La Salle, IL: Open Court, 1987.

————. Representation and Reality. Cambridge, MA: MIT Press, 1988.

————. Renewing Philosophy. Cambridge, MA: Harvard University Press, 1992.

W. V. Quine. 'Two Dogmas of Empiricism.' In From a Logical Point of View, second edition. Cambridge, MA: Harvard University Press, 1961.

————. Word and Object. Cambridge, MA: MIT Press, 1960.

————. 'Reply to Professor Marcus.' In The Ways of Paradox. Cambridge, MA: Harvard University Press, 1966.

————. 'Ontological Relativity.' In Ontological Relativity and Other Essays. New York: Columbia University Press, 1969.

————. 'On the Reasons for the Indeterminacy of Translation.' Journal of Philosophy 67 (1970): 178-183.

————. 'Reply to Chomsky.' In Words and Objections: Essays on the Work of W. V. Quine, edited by D. Davidson and J. Hintikka, 292-352. Dordrecht: Reidel Publishing Company (1975).

Récanati, François. Direct Reference: From Language to Thought. Oxford: Blackwell, 1993.

Salmon, Nathan. Frege's Puzzle. Cambridge MA: MIT Press, 1986.

Searle, John. Minds, Brains and Science. Cambridge, MA: Harvard University Press, 1984.

————. 'Indeterminacy, Empiricism, and the First Person.' Journal of Philosophy 84 (1987): 123-146.

————. The Rediscovery of the Mind. Cambridge, MA: MIT Press, 1992.

Sprigge, Timothy L. S. 'Final Causes.' Proceedings of the Aristotelian Society Supplementary Volume 45 (1971): 149-70.

Stampe, Dennis. 'Toward a Causal Theory of Linguistic Representation.' In Midwest Studies in Philosophy, vol. II, edited by edited by Peter A. French, Theodore E. Uehling, Jr., and Howard K. Wettstein, 42-63. Minneapolis: University of Minnesota Press, 1977.

Stich, Stephen P.. Deconstructing the Mind. Oxford: Oxford University Press, 1996.

Strawson, Galen. Mental Reality. Cambridge, MA: MIT Press, 1994.

Taylor, Charles. The Explanation of Behavior. London: Routledge and Kegan Paul, 1964

Taylor, Kenneth. 'Belief, Information and Semantic Content: A Naturalist's Lament.' Synthese 71 (1987): 97-124.

Wright, Larry. 'Functions.' The Philosophical Review 82 (1973): 139-168.

————. Teleological Explanations: An Etiological Analysis of Goals and Functions. Berkeley: University of California Press, 1976.