REVIEW ARTICLE

Naturalistic Representation

DAVID PAPINEAU [1987]: *Reality and Representation*, Basil Blackwell 1987.

1 *Introduction*
2 *Anti-realism*
3 *Davidson*
4 *Intrinsic and Extrinsic Identification*
5 *Naturalizing Reference*
6 *Papineau's Teleological Theory*

I INTRODUCTION

David Papineau's book, *Reality and Representation* [1987], is rich, wide-ranging, and bold. It covers central issues in metaphysics, the philosophy of language, the philosophy of mind, and epistemology. It even urges views of mathematics and morality. To present this *Weltanschauung* in less than 250 pages demands a bit of scrimping on details. Papineau has coped with this problem remarkably well but, nevertheless, many will feel that there is too much hand waving and not enough tight and sustained argument. Papineau's style throughout is breezy, and nearly always clear.

In the first half of the book, Papineau discusses the nature of representation. He offers a naturalistic theory which he sees as a defense of realism about the external world. I shall examine this discussion.

I shall not be concerned with the second half of the book, interesting though it is. In this half Papineau develops a naturalized epistemology, arguing against the sceptic that we can avoid errors by reflecting on the reliability of our methods of belief-evaluation.

A theory of representation has the task of saying in virtue of what a token has the representational content that it has. Naturalistic responses to this task appeal ultimately to one or more of three causal relations between representations and reality: (i) the *historical* cause of that particular token; (ii) the *reliable* cause of tokens of that type; (iii) the *purpose* of tokens of that type, or of the mechanism that produces them, where the purpose is explained causally along Darwinian lines. Papineau's theory is of the latter, *teleological*, sort.

My main concern will be to examine that theory and compare it with other naturalistic theories. However, I shall start with a consideration of the context within which Papineau presents his theory. Though Papineau's own 'problematic' is admirably and uncompromisingly realistic and naturalistic, he presents his theory in a dialectic with theories from other, largely hostile, problematics. This procedure would be worthwhile only if it involved a critique of the central assumptions of the other problematics and hence of the theories in it. Papineau's discussion does not involve such a critique. As a result, the discussion mostly 'goes past' the other theories.

I shall consider first Papineau's dialectic with anti-realism and second his dialectic with Davidson.

2 ANTI-REALISM

The dust jacket claims that Papineau 'defends realism from a naturalistic perspective'. I don't think that he does because he never really comes to grips with the underlying assumptions of the various anti-realisms.[1]

Papineau sets out to defend realism by arguing that error is always possible: our representations can be wrong about reality. What is the connection between the possibility of error and realism?

Papineau is rightly struck by the importance of what I call 'the independence dimension' of realism: it is essential to the realist view that reality be independent of our minds, our language, our capacities to know about it, and so on. Papineau captures a good part of this in his claim that 'reality is independent of thought, in the sense that there is no conceptual link between how things are and how humans take them to be' (pp. ix–x). In contrast, anti-realism holds that 'at some level judgement and reality must fit each other' (p. xii); that at some level error is impossible. This is the connection.

It is important to see that two directions of dependency—two bases for a lack of error—might be in question here, and realism is only concerned to reject one. Realism must reject the dependency of reality on thought. It must deny that entities depend for their existence and nature on our cognitive activities. But realism can be neutral on the idea that thought might be so dependent on reality as to be error-free, for that idea, however implausible, has no metaphysical implications. To think that our representations cannot err because they determine reality is to have an implausible theory of reality. To think that reality determines that representations cannot err is to have an implausible theory of mind and language. The first sort of implausibility is exemplified in the popular tradition of constructivist anti-realism, stretching from Kant onwards. The second sort is exemplified in the *Verstehen* tradition in the social sciences, which need not be *metaphysically* anti-realist. The second

---

[1] How much is my disagreement with Papineau a 'merely verbal' one over what to *call* 'realism'? I think, very little, for reasons to be found in my [1984] and [1991].

sort can even be contemplated by someone who is steadfastly naturalist and realist, as one of the Jerry Fodors showed [1984, 1990]. Indeed, if we include Papineau's important qualification, 'at some level', and interpret it generously, it is *not obviously* implausible that reality does make our representations right. And even though it is central to Papineau's teleological theory that a belief need not usually be true, he himself thinks that a belief of a certain type must 'have had a history of usually being true' (p. 89).

A naturalistic defense of realism must tackle the assumptions that lead to constructivism. It must tackle the Cartesianism and pervasive apriorism that underlie attempts to arrive at a world view from epistemology or semantics. Such a critique is necessary to expose constructivism as the philosophical conceit that it is.

Papineau does not offer such a defense of realism. His concern with 'what the analysis of representation implies for the relation between judgement and reality' (p. xii) is not with the direction of dependency that is relevant to realism but with the other direction. His enterprise is really to argue for a naturalistic theory of representation that allows for error, and that *largely takes realism for granted.* That is a thoroughly respectable enterprise, one in which many are engaged. But then it is a mistake for Papineau to think that he is dealing with 'the fundamental issue' of the realism debate, and to think that he is answering anti-realists like Hilary Putnam and Michael Dummett (pp. xiii, 113–22). They are in different problematics which a naturalistic theory of representation largely passes by.

The possibility of error is not to be identified with realism. Even less is the possibility of divergence of opinion. Papineau moves from the former identification to the latter on the ground that divergence without error leads to 'the apparently absurd conclusion that different humans live in different realities' (p. 10). *But this 'apparently absurd conclusion' is precisely what most anti-realistics in this century have believed!* Such different realities are con- structed, for example, by Goodman's 'versions', by Kuhn's 'paradigms', and by the 'discourses' of the structuralists and post-structuralists that dominate French intellectual life and many literature departments around the world.[2] That Papineau's account of anti-realism misses its most popular version illustrates how distant he is from the mainsprings of anti-realism.

3 DAVIDSON

Papineau's view of realism leads him to start his defense with a consideration of Davidson. For Davidson subscribes to a 'principle of charity' which makes error impossible and diversity unintelligible (give or take a bit). So, on Papineau's criteria, Davidson counts as a strong anti-realist. From my

[2] I have criticized constructivism in Devitt and Sterelny [1987], pp. 199–221.

perspective, Davidson is a good example of the failure of these criteria, because there is no clear evidence that Davidson is an anti-realist in any interestingly metaphysical sense.[3]

Set aside the realism issue. Is Davidson an appropriate starting place for Papineau? I think not. The start is with the argument from 'On the Very Idea of a Conceptual Scheme' (Davidson [1984], pp. 183–98), which Papineau aptly calls, following Rorty ([1979], p. 78), 'the transcendental argument to end all transcendental arguments' (p. 22). But what is a nice naturalist like Papineau doing with an argument like *that*? Papineau responds sensibly (pp. 31–9) that the principle of charity, which is at the core of the transcendental argument, is false and should be replaced by the more moderate 'principle of humanity' (*cf.* Grandy [1973]). But he never attempts to confront the assumptions, whatever they may be,[4] that underlie Davidson's position. Rather, he takes a naturalistic and realistic approach to language and mind that seems to be quite at odds with the Davidsonian approach.[5]

Though Davidson's approach has different origins it is, I think, akin to the earlier-mentioned *Verstehen* approach.[6] Perhaps Papineau thinks so too, though he is not explicit on the matter. Anyway, when he later considers the *Verstehen* interpretation of the principle of humanity, he gives the appropriate naturalistic response. 'There is no distinctive mode of *understanding* by which we account for human beliefs' (p. 120); we look for *explanation* here as everywhere else in science. But then if there is nothing special about our approach to representation, we don't need a mawkish label, 'principle of humanity', to capture that approach. And there is no point in reaching this principle from a consideration of Davidson's semantics where that consideration does not confront the foundations of that semantics. From a naturalistic perspective, the principle of charity is obviously false[7] and the principle of humanity is the following application of the truism that the scientist should 'seek the best explanation': ascribe to an organism only those beliefs which it is explicable that it should have (whether the beliefs are true or false, rational or irrational).

Perhaps Davidsonian semantics has some other role to play in Papineau's theory. Certainly, Papineau spends several pages on the details of this semantics (pp. 18–39) and seems to approve of it (apart from the principle of

---

[3] I made the same mistake as Papineau about Davidson in an earlier work ([1984], p. 180; but compare [1988], p. 165).

[4] The influence of Quine on Davidson is obvious, but the precise relationship of Davidson's semantics to Quine's is difficult to discern. I have argued that Davidsonian semantics, despite its physicalistic overtones, combines 'semanticalism' (Field [1972], p. 92) about truth with instrumentalism about reference ([1984], pp. 163–71).

[5] In a revealing passage, one influential Davidsonian puts semantics outside physicalistic science (McDowell [1978], p. 128).

[6] Devitt and Sterelney [1987], pp. 238–49; see McDonald and Pettit [1981] for a lengthy argument to this effect.

[7] As I have attempted to argue: [1981], pp. 115–8, [1984], pp. 172–9.

charity). However, it is unclear why he does. So far as I can see, that semantics plays no role in his theory of representation. Indeed, *how could* his theory, which is primarily a theory of reference, be part of a semantics that denies any place to a theory of reference (Davidson [1984], pp. 215–25)?

To sum up these two sections, I think that Papineau is mistaken in presenting his theory of representation as an answer to anti-realism and as a development from Davidsonian semantics. Those views are from alien problematics that Papineau does not really confront.

I move now to Papineau's theory.

4 INTRINSIC AND EXTRINSIC IDENTIFICATION

Philosophers typically urge two ways of identifying mental representations: a 'narrow' way in terms of functional-role properties that supervene on what is 'inside the skin'; and a 'wide' way in terms of the truth-conditional properties. Papineau calls these 'intrinsic' and 'extrinsic' identifications (p. 22.)[8]

Two questions arise: (1) Why do we need these two forms of identification? (2) What is the relation between them?

Concerning (1), there tends to be a *presumption* that we need the familiar extrinsic identification and then *arguments* along the following lines to show that we *also* need the intrinsic indentification. First, the extrinsic identification is too 'coarse-grained' to serve the purpose of psychological explanation: the difference between believing that Jocasta is on the bed and believing that Mum is on the bed is crucial to explaining Oedipus' behaviour even though the *beliefs have the same truth conditions. Second, there are cases where a* representation lacks truth conditions because a part of it lacks a referent, and yet we still want to ascribe a content to the representation. Third, arguments from functionalism, methodological solipsism, and the computer analogy, lead to the idea that only what supervenes on what is inside the skin, as truth conditions and reference do not, is relevant to psychological explanation. Papineau is influenced by these sorts of argument (pp. 43–54, 91–8).

A problem with this is the presumption that we need the extrinsic identification. The more the case builds for an intrinsic, functional-role, identification of representations, the less does there seem to be a place for the extrinsic, truth-conditional, identification. What do we need that identification *for?* If the functional role we ascribe to a representation explains its psychological role, what does ascribing a truth condition explain? Often it seems as if the mere familiarity of *truth* and hence of *truth condition*, prevents this question from becoming pressing. However, once it is noted that this familiarity may arise from a *deflationary* notion of truth, the question should

---

[8] Papineau endorses the common conflation of the functional identification of representations with their syntactic identification (p. 47). I have argued against this conflation (1989], pp. 371–4).

become pressing. For if truth is deflationary, the property of having certain truth conditions can have no role in the explanation of a representation. And no account of that property, teleological or otherwise, is required.[9]

Papineau is sensitive to this problem of need (pp. 46–54). His solution is along lines first proposed by Putnam ([1978], pp. 99–103): we need to ascribe truth to representations to explain the success of actions (pp. 54–7, 81–4). I shall return to this idea in Section 6. In a different context, and in passing, Papineau mentions another popular solution: we need to ascribe truth to representations to learn about the world from them (p. 97). This idea, first proposed by Hartry Field ([1972], pp. 370–1), is much more promising, though it has proved very hard to work out the details.[10]

What about (2), the question of the relation between the two ways of identifying meaning or content? The tendency has been to take these 'two factors' as requiring distinct semantic theories[11] and as being largely independent of each other.[12] This bifurcation of meaning strikes me as insufficiently motivated and implausible. I think that the intrinsic factor is part of the extrinsic factor.

To believe in an extrinsic factor is to believe that at least part of the meaning of a sentential representation is its property of having certain truth conditions. We need then to explain that property to show that we are *entitled* to it. The most plausible explanation is in terms of the syntactic structure of the representation and the referential properties of the concepts that occupy the structure. We must then explain syntax and reference. Syntax is a matter of inferential role. What about reference? I suggest that three sorts of theory of reference are possible and that very likely each sort has some application.

Description theories are one sort. They were once popular for proper names and natural-kind terms but are more plausible for terms and concepts like 'bachelor' and 'pediatrician'. According to a description theory, the reference of a person's concept is determined by the reference of certain other concepts associated with it by the person. These associations are also matters of inferential role. Description theories explain the reference of some concepts in terms of the reference of others and so are essentially incomplete. There must be some concepts whose reference is not parasitic on others, but is explained in terms of direct relations to world. Only one kind of relation is appealing to the naturalist: a causal relation. So, for the ultimate explanation of reference, we look for what we might call 'pure-causal' theories.[13] These might be historical,

---

[9] For more on deflationary truth, see Quine [1970], pp. 10–13, Grover, Camp, and Belnap [1975], Leeds [1978], Field [1986b], Brandom [1988], Devitt and Sterelny [1987], pp. 162–5.

[10] Two attempts are: Schiffer [1981], Field [1986b]. My own most recent attempts are [1984], pp. 91–9. Devitt and Sterelny [1987], pp. 169–70.

[11] Field [1977] has been very influential in this respect.

[12] See particularly McGinn [1982], pp. 211, 230, Loar [1982], pp. 280–2, [1983], p. 629. For an earlier version of the idea, see Putnam [1975], p. 269.

[13] Searle [1983] disagrees. I have responded [1990].

reliablist, or teleological (Section 1). The third sort of theory is a mixture of the other two, a 'descriptive-causal' theory: reference is determined partly by direct causal links to reality and partly by inferential associations with certain other concepts.

Now the key point so far as (2) is concerned is that the explanation of syntax, and each of the possible explanations of reference, involve functional roles that are internal to the mind.[14] This is obvious with a description theory: the dependence of one concept on another is a functional matter. However, even with a pure-causal theory, inner processing must take place if the organism is to take referential advantage of the causal action of the environment.

In sum, the extrinsic factor of meaning or content—the property of having certain truth conditions—already includes a great deal of inner functional roles. It is not clear then why the reasons for believing in an intrinsic factor will not be satisfied by those roles. If they are satisfied, those reasons do not motivate a distinct functional-role semantics and do not lead to two distinct meaning factors. The intrinsic factor can be obtained from the extrinsic by abstracting from truth-referential meaning those parts that are determined by what is outside the skin. The intrinsic is the internal part of the extrinsic.[15]

If this is right, it eases the earlier problem of saying why we need the extrinsic factor: for the extrinsic is just a little bit more than the intrinsic and so once we have shown that we need the intrinsic we are not that far from showing that we need the extrinsic. In contrast, the usual theories that treat the two factors as distinct, and as the subjects of different semantic theories, make the problem particularly acute.

Where does Papineau stand on (2)? His teleological theory, to be discussed in the next section, is an example of a pure-causal theory. Papineau thinks that this theory applies only to 'observational' concepts. For the others he proposes what is, in effect, a description theory (pp. 78–9, 92–3).[16] However, he allows that all concepts may be 'to some degree non-observational' (pp. 93–4). In so doing he is, in effect, contemplating that some concepts may be covered by causal-descriptive theories. So Papineau allows for theories of each of the above three sorts. Furthermore, he thinks that the descriptive element in determining reference is made up of the functional roles that go into the intrinsic factor. So there is no reason why he should not go along with my view of how the intrinsic is to be unified with the extrinsic.

So far, so good. However, there are two aspects of Papineau's view that are not good. First, his version of a description theory for a non-observational concept is simply dreadful: the 'concept refers to that entity whose role in the causal structure of the world mirrors the role of the concept itself in the cognitive structure' (p. 93). This might seem to work for concepts that refer to

---

[14] For the opposite view on reference, see Block [1986], p. 665, Wettstein [1988], p. 415.

[15] For more on this see my [1989], pp. 377–81.

[16] He still wants to talk of these concepts as having a purpose, however (p. 80).

unobservable natural kinds identified by their causal roles (though I don't
think it really does). It does not begin to work for many other concepts that
cannot plausibly be seen as 'observational'[17] and will therefore need a
description theory (or a descriptive-causal theory). 'Bachelor', for example,
probably depends for its reference on its causal links to 'unmarried' and yet
that causal link is not 'mirrored' by one between bachelors and unmarried
objects. Similarly, 'pediatrician' and 'doctor', 'sloop' and 'boat', and so on
through most of the dictionary. The functional-role of a concept determines its
reference not by mirroring reality but by relating it to other concepts that refer.

Second, Papineau, like almost everyone else, is a meaning holist (pp. 97–8).
Indeed, he is committed to an extremely individualistic version of holism: if the
functional role of a concept in your head, or in my head last week, differs *in the
slightest degree* from one in my head now, then the concepts must differ. I have
argued against holism elsewhere (forthcoming a, b) and will say no more here
beyond drawing out the disastrous consequence of holism for Papineau.

According to Papineau's holism, all of a concept's functional relations go
into its intrinsic meaning. So all go into determining reference. So, *any* mistake
in a theory will lead to reference failure. Yet all interesting theories are
probably a bit mistaken, at least. So, probably all such theories fail to refer.[18]
Faced with this disaster, Papineau fudges: we have 'good pragmatic reasons for
ascribing referents' (p. 97) and so, he implies, it is still all right to do so. If we
can answer question (1), which Papineau thinks we can, we do indeed have
good reasons for ascribing referents, though these reasons would be no more,
nor any less, 'pragmatic' than those for ascribing, say, mass to objects. The
problem that Papineau is dodging is that *on his theory of reference we are not
entitled to ascribe referents*. What we want is one thing, what we can get,
another. Either Papineau's theory for 'non-observational' concepts must go or
he must give up ascribing reference to them (as part of a genuinely explanatory
theory of mind).

5 NATURALIZING REFERENCE

The naturalization of reference (or intentionality) comes down ultimately to
some sort of pure-causal theory. Each sort must be historical, reliablist, or
teleological, or some combination of these. In this section we shall consider
some examples.

1. Papineau's teleological theory is a development of that in 'Representation
and Explanation' [1984]. In the same year Ruth Millikan offered a detailed

---

[17] Papineau does not explain what 'observational' concepts are beyond saying, unhelpfully, that
they 'stand for observational states of affairs' (p. 79).

[18] Theories that keep the two factors distinct avoid this disaster: the holistic intrinsic factor has no
role in determining reference. Of course, this leaves the problem of what does determine
reference.

teleological theory in *Language, Thought, and Other Biological Categories* [1984], and Jerry Fodor suggested a partly teleological theory in 'Semantics, Wisconsin Style' [1984]. As Papineau says, 'this was clearly an idea whose time had come' (p. xvi). And a very interesting idea it is too. It is also the most difficult of the three ideas for naturalizing reference. I shall discuss it in the next section.

2. Fodor's teleological suggestion was a response to another very interesting idea: the reliablist, 'informational', or 'indicator', theories of Dennis Stampe [1979] and Fred Dretske [1981]. The idea is that a token represents what tokens of that type are *caused by.* The token 'carries the information' that a certain situation holds in much the same way that three rings carry information about the age of a tree. Tokens of that type are 'reliably correlated' with a situation and hence 'indicate' that situation. This idea faces a major problem. How can it allow *error?* Occasionally we see a muddy zebra but *mis*represent it by thinking 'horse'. So, some zebras are among the things that would cause tokens of 'horse'. What 'horse' is reliably correlated with is really the presence of horses or muddy zebras or the odd cow or. . . . So it should refer to horses, muddy zebras, the odd cow. . . . The problem is that many things that a token of a certain type does not refer to, including some denizens of Twin Earth, *would cause* a token of that type.

To solve this problem, the informationalist must claim that circumstances in which muddy zebras cause 'horse' are not appropriate for fixing its reference: a token of 'horse' represents what such tokens are caused by in 'normal' circumstances.[19] The problem then is to give a naturalistic account of 'normality'.

3. At this point it is very tempting to appeal to teleology. If this appeal does not go too far, it can be combined with an informationalist approach. The informationalist must appeal only to the purpose or 'teleofunction' (Millikan [1990b]) of the *mechanism* that brings about the representational state. He appeals to Darwin to show that certain circumstances are the 'normal' ones for the functioning of that mechanism. States represent what they indicate *in those circumstances.*[20] Sometimes, however, the appeal seems to go further without noticing it, effectively abandoning the informationalist approach altogether in favour of a teleological approach like Papineau's. This happens when the appeal is not simply to the function (meaning *teleo*function from now on) of the mechanism that produces the state but to that of *the state itself.*[21] If the state

[19] Normal circumstances are sometimes confused with 'optimal' circumstances (*e.g.* in Stalnaker[1984], see also [1986], which is a response fo Field [1986a] and Schiffer [1986]). On this confusion, see Godfrey-Smith [1989].

[20] See, *e.g.*, Dretske [1988].

[21] Early informationalist discussion paid little, if any, attention to the distinction between the functions of states and of mechanisms; *e.g.*, Enc [1982], Fodor [1984] and [1990], Dretske [1986]. The importance of the distinction is emphasized in Godfrey-Smith [1989] and Fodor forthcoming. I draw on these two papers in discussing the distinction in the next section.

represents what it is supposed to respond to then its reliable correlations are beside the point.

The hybrid approach—informationalism together with the function of the mechanism—seems promising. However, there are apparently overwhelming problems for a theory of representation that has *any* dependence on reliable causation, as Peter Godfrey-Smith (forthcoming) has shown; indicator semantics seems doomed. The basic problem is simple: harmless 'false' positives'.[22] It is common in nature for an organism to represent a situation as one in which there is a predator, or food, or whatever, and to be *more often wrong than right*; so what it indicates is *mostly* not what it represents. This situation is common because it has an evolutionary pay-off; so the circumstances of error are as 'normal' as could be. Consider the typical bird that is the prey of hawks. A high proportion of the time that it registers the presence of a predator it is wrong; it has responded to a harmless bird, a shadow, or whatever. These false positives do not matter to its survival. What matters is that it *avoid false negatives*; what matters is that it registers a hawk when there is one. The price the bird has to pay for that is frequently registering a hawk when there isn't one. What nature has selected is a *safe* mechanism not a *certain* one.

4. Fodor has abandoned the teleological approach to representation— indeed, he has become a severe critic of that approach—in favour of a novel sort of informationalism ([1987], forthcoming). Like the informationalist, he thinks that it's a law that horses cause 'horse'. The problem then is that it is also a law that the odd cow causes 'horse'. Fodor's solution is that 'horse' nevertheless refers only to horses because the latter causal relation holds *because* the former does, but not vice versa: the relation to the odd cow is *asymmetrically dependent* on the relation to horses.

Fodor has defended his theory with great ingenuity against a range of counter examples. My concern comes earlier: I don't see how the theory covers any *examples*. What we need are some paradigm cases of '*A*' referring to *A*s where the following is the case: *A*s cause '*A*', *B*s cause '*A*' and *it is obvious that the latter causal relation is asymmetrically dependent on the former.* The asymmetric dependency that Fodor claims in the case of 'horse', for example, is far from obvious. At first sight, the basic law in that case seems to be that horses, muddy zebras, the odd cow, ... cause 'horse'. This seems to be the law on which the others asymmetrically depend: it is because of this law that horses cause 'horse' and that the odd cow causes 'horse'. We can even explain why that law seems basic, or at least *more* basic: it is because horses, muddy zebras, the odd cow, ... *all have 'horsy looks'* and it is horsy looks that makes us think 'horse'. If this were so, asymmetric dependency would yield just what we were trying to avoid: 'horse' refers to horses, muddy zebras, the odd cow. . . .[23]

---

[22] See also Stich [1985], Millikan [1989] and [1990].
[23] This paragraph draws on Godfrey-Smith [1989].

Fodor's response to this sort of objection is strange. Because 'this stuff is supposed to be philosophy' (forthcoming: 43) he feels little obligation to establish the existence of the required asymmetries. Rather he is concerned to refute the arguments of Brentano and others that a naturalistic account of intentionality is *impossible*. Certainly, if there *were* the asymmetries Fodor describes, they *would* explain reference. So his theory is a *possible* explanation, and thus answers the transcendentalists. However, there is also the question of whether this is the *actual* explanation, a question that concerns many (particularly those who have never had a transcendental moment in their lives). To show that it is the actual one, we need to show that there is a correlation between reference and asymmetries.[24]

It is tempting to argue along these lines. There are clear *semantic* asymmetries: for example, it is because 'horse' means HORSE, and not because it means MUDDY ZEBRA, that horses and muddy zebras cause 'horse'. How is this to be explained? Only a causal explanation would be naturalistically respectable. *So what else could the explanation* be but the one in terms of asymmetric causal dependencies? Unfortunately, this seems more like naturalistic desperation than an argument.

5. Finally, consider the historical causal approach. This approach attempts to explain the reference of a token not in terms of what *would* cause tokens of that type but in terms of what *did* cause that token. This approach is derived from the 'causal theories of reference' of Keith Donnellan [1972], Saul Kripke [1980], and Hilary Putnam [1975]. The central idea of these theories, which seems especially plausible for proper names and natural-kind terms, is one about *reference borrowing*: a person's term or concept refers to an object in virtue of a network of causal chains stretching back through communication situations involving other people whose term or concept referred to that object. This idea was not offered in a naturalistic setting but it is appropriate to place it in such a setting.[25] So placed it is largely complementary to the teleological and informationalist approaches. I have already pointed out that these approaches concern those of a person's concepts that do not depend for their reference on the person's associated concepts. In that respect these approaches are concerned with the ultimate explanation of reference. If the theory of reference borrowing is right, then their concern is ultimate in a further respect: they are

---

[24] Fodor applies his pure-casual theory not merely to relatively 'observational' concepts like 'horse' but to *all* concepts. He does this because he doubts that the definitions needed by description theories are available ([1981], pp. 257–316). Presumably, he doubts also that the partial definitions needed by descriptive-causal theories are available, though he does not argue this. His theory of concepts like 'bachelor', 'pediatrician', and 'paperweight', does seem really incredible. It seems most unlikely that there is any law that the appropriate objects do cause the concept. So it is even less likely that reference comes from asymmetric dependence on such a law.

[25] Field [1972], Devitt [1974], [1981]. Devitt and Sterelny [1987]. The idea was also for language not mind. I think [1985], in contrast to Evans [1982], that it is appropriate to apply the idea to the mind.

concerned with concepts that do not depend, via reference borrowing, on *other people's* concepts. Their concern is with the concepts of a person that depend for reference neither on other concepts of that person—hence not covered by a description theory—nor on the concepts of other people—hence not covered by a theory of reference borrowing.

Those who follow informationalist approaches do not mention the theory of reference borrowing. This is not so important if I am right that the theory is complementary to those approaches. However, it is important that anther idea, less obviously present in writings of Donnellan, Kripke, and Putman, also goes unmentioned. For this idea may be a contribution to the (doubly) ultimate explanation. Indeed, informationalists in trouble often help themselves to this idea, apparently without noticing the departure from pure informationalism.[26] The idea is that a person 'at the bottom of the network'—someone who does not borrow reference—has a concept the reference of which is determined by the actual causal interactions between that person and the referent. Thus, we depend on Aristotle's contemporaries for the reference of 'Aristotle' but *their* concepts referred to him because they were, as a matter of fact, caused by him; the concepts were 'grounded in' him.

This idea also seems promising, but it also faces an apparently overwhelming problem, the '*qua*' problem.[27] Consider my natural-kind concept 'horse'. This is grounded in a few horses. But those objects are not only horses, they are mammals, vertebrates, and so on; they are members of very many natural kinds. In virtue of what is my concept grounded in the objects *qua* horses rather than *qua* any of the other natural kinds of which they are members? So in virtue of what does it refer to all and only horses? Why does the concept formed by those groundings not 'project onto' the members of these other natural kinds? The problem is worse. What restricts the kinds in question to *natural* kinds? The objects in which 'horse' are grounded may be pets, investments, brown, and so on, and they *are* horses or cows, horses or cows or kangaroos, and so on. In virtue of what are the groundings not in them *qua* members of those kinds? Attempts have been made to solve this problem,[28] but none of them seems to come close to doing so.

The earlier problem of error for informational theories is, in effect, a version of the *qua* problem. What the problem shows is that the informational and historical theories are relying at bottom on some unexplained intentionality. This is unsatisfactory, of course, because the object of this naturalistic exercise is precisely to explain intentionality.

---

[26] *E.g.* Dretske's first attempted solution to the problem of error ([1981], pp. 193–5). See also Dretske [1988], p. 73. It is hard to see how a pure informationalist can handle the Twin Earth problem in a way that is not *ad hoc*: 'water' is as reliably correlated to xyz as to H_2O.

[27] Sterelny [1983], p. 120. The problem for natural-kind terms was noted by Papineau [1979] and Dupre [1981] and for proper names, by Devitt ([1981], pp. 60–4).

[28] Sterelny [1983], Devitt and Sterelny [1987], pp. 72–8.

It is against this record of failure that we must assess Papineau's teleological theory. Because the record is so dismal it behoves those who think, as I do, that there must be something to reference and intentionality to look tolerantly at teleological theories. Alas, we soon discover that a lot of tolerance is needed.

6 PAPINEAU'S TELEOLOGICAL THEORY

According to Papineau, a belief token represents its truth condition because the 'biological function' or 'purpose' of that type of token is to be present when that condition obtains; and a desire token represents its satisfaction condition because the biological function or purpose of that type of token is to give rise to that condition. 'To talk about something's "function", in the biological sense, is to talk of some effect of that thing which accounts for its being there' (p. 64); some past process of natural selection favoured that thing *because* it produced that effect (p. 65).

These are very difficult matters. Unfortunately, this is a place where Papineau does a lot of hand waving. I shall list some worries. I think that the cumulative effect of these is fairly decisive against Papineau's theory and rather worrying for the teleological program in general.

1. Why suppose that beliefs and desires have functions at all? One might be tempted to think that it is obvious that they have functions for, after all, 'we have them, and we have successfully evolved'. But this is to be tempted by naive Panglossianism of the sort so beautifully criticized by Stephen Jay Gould and Richard Lewontin [1979]. Lots of the properties that species have are not adaptive: they were simply brought along by others that were adaptive or are there 'by chance'; poor eyesight in polar bears and chins in humans are examples.

2. Nevertheless, it is plausible to suppose that, in some sense, beliefs and desires do have functions. In considering this, the earlier distinction between the function of a *mechanism that produces states*—or the function of those states *in general*—and the function of *a particular state produced by that mechanism* is important. It is plausible that hair has a function, but very implausible that each particular hair has a particular function. Closer to home, it is very plausible that the eye has an informational function, but much less plausible that each informational state of the eye has a particular informational function. Similarly, it is plausible that the mechanisms that produce beliefs and desires—or beliefs and desires in general—have functions. It is much less plausible that each particular belief or desire has a particular function. Indeed, it may be close to incredible. Yet that is the view that must be embraced by a pure teleological account like Papineau's. (The former, more plausible, view has the disadvantage that its explanation of content is incomplete. If we complete it with an informationalist element we get the hybrid account we viewed dimly in the last section.)

3. Papineau thinks that a desire was selected because it gives rise to its satisfaction condition.[29] If the selection is genetic, that requires that the satisfaction of the desire improves the organism's chances of reproducing. This rules out the possibility of any of these innate desires being *irrational* in that they do not serve reproductive interests. This seems Panglossian. Consider, for example the human desires, beyond a safe level, for salt, sugar, saturated fat, or drugs.

What about all the desires that are not innate and hence cannot have been genetically selected?[30] Consider, for example, the desires to be rich and famous, that the Redskins will win occasionally, and that anti-realism will perish. Papineau does seem to intend his teleological theory to apply to such desires, acquired during the lifetime of an organism (p. 69). It is hard to see how such a theory could work. We would need an account of non-genetic selection that matched these desires to their satisfaction conditions. Perhaps we could start with the assumption that satisfying a desire is *psychologically* rewarding,[31] whether it is genetically rewarding for the species or not. So a desire that leads to such satisfaction is selected for and its function, hence its content, is the condition that it brings about. One problem with this is that many desires, like those above for example, are likely to remain unsatisfied. Yet they still have content. Furthermore, the theory seems to have a severe indeterminacy problem. Any worldy condition brought about by a desire can be described in many different ways. Under one of these descriptions the condition satisfies the desire and that description expresses the content. *But, which description is that?*

In sum, we seem to have hardly begun the account of *how* your average desire was selected because it brings about a particular worldy condition under a particular description, the condition which, under that description, satisfies the desire. It's a nice story, but it's just so hard to believe.

4. A belief's function is to be present when its truth condition obtains. How does being so present lead to its selection? Papineau claims that 'beliefs will in general have biologically advantageous effects only in so far as they have effects which satisfy desires' (p. 70). To complete the account, we need to answer two questions.

First, why does satisfying desires lead to a belief's selection? The talk of what is 'biologically' advantageous suggests that the concern here is with innate beliefs that have been selected genetically. But then satisfying desires would not lead to a belief's selection unless that satisfaction were reproductively advantageous. This would require that no desire be 'irrational', which is

---

[29] Compare Millikan [1986], p. 63.
[30] See Millikan [1990a] for her handling of this problem and that of irrational desires.
[31] It could be psychologically rewarding only if the person was *aware* that her desire had been satisified. The word 'satisfy' may suggest such awareness, but we should note that the Redskins might win occasionally, hence satisfying a person's desire in *some* sense, and yet the person not be aware of this.

implausible for innate desires, as we have just noted, and even worse for acquired ones. However, despite the suggestion, it is clear from an earlier passage (pp. 66–7) that Papineau's teleological theory for beliefs, like that for desires, is intended to cover acquired states. To explain belief selection, then, he needs the above assumption that desire satisfaction is *psychologically* rewarding.

Second, how is desire satisfaction tied to the presence of the belief's truth condition? 'The truth condition of a given belief type is that circumstance whose presence guarantees the success of actions based on that belief' (p. 56). This is surely way too strong.[32] Truth does not guarantee success. The satisfaction of many desires depends on things that are quite outside the control of the desirer: on the natural order; on the socio-economic situation; on the personal foibles of others; and so on. Sadly, but often, no amount of true believing will make any difference (Philosophy won't buy me love). Aside from that, true beliefs will not do the job unless you have *enough* of the *right* ones. What you need is not truth but *truth where it matters*.[33]

5. For Papineau, a belief token has a content only if it has a function. It has a function only if its type has been selected. Its type has been selected only if it has regularly led to desire satisfaction. It has regularly led to desire satisfaction only if it has 'had a history of usually being true' (p. 89). So a belief token has no content until its type has a history of truth. The worry about this is not the apparent paradox of truth without content, for that can be explained away. The worry is that a state can surely be a belief, with a determinate content, before its type has had a chance to be selected.[34]

6. The insistence that a belief must have had a history of usually being true runs into the problem that was so disastrous for the earlier hybrid theory consisting of informationalism with a dash of teleology (3 in Section 5). This is the problem of false positives. Many species have survived only by mostly misrepresenting the world. What got selected, *from the beginning*, were beliefs whose tokens were *true when it mattered for survival*, even though they may have been mostly false. In this crucial respect, Papineau's full-blown teleological theory is no advance over the partly teleological hybrid theory.

7. It is a consequence of Papineau's theory that a physical replica of me, created 'by some cosmic accident a few seconds ago', would not have beliefs

---

[32] Millikan's story is slightly different ([1986], pp. 68–71; see also Godfrey-Smith [1988]). According to her the truth condition of a belief is the state of the world essential in the explanation of how the belief has historically assisted the performance of the functions of those parts of the cognitive system whose activities are affected by the belief. In particular, the belief has assisted in the satisfaction of desires.

[33] This paragraph draws on my argument that truth does not explain success ([1984]: 88–91).

[34] I partly owe this point to Peter Godfrey-Smith, who in turn credits Bill Ramsey for a similar point against Dretske [1988].

that were about anything. Papineau agrees that this is counter-intuitive but bravely bites the bullet (pp. 72–5).[35]

8. All naturalistic theories of representation face the problem of compositionality. Representations are said to get their content, in one way or another, from causal interactions with reality. But the interaction (in humans at least) is always between a *state of affairs* and a *complex* syntactic structure made up of simple representations or *concepts*. Somehow or other the referential properties of the concepts have to be abstracted from these interactions. Other complex representations can then derive their truth conditions from the referential properties of the concepts they contain. Papineau handles this problem as follows. First, 'the biological purpose of any given concept is to allow us to form certain beliefs' (p. 76).[36] On the basis of his story of the purposes of beliefs, he goes on to redescribe the purpose of concepts as being to refer to objects (pp. 76–7). Perhaps this will work, but it is *very* swift. And one wonders about the effect of compositionality on our earlier worries. Will it make them better or worse?

9. There is a general epistemological worry about teleological theories of representations. In the paradigm cases of functions, it is quite clear before we make any teleolgical moves what the object in question *regularly does* (even if, like sperm, only a tiny proportion do it). Thus, it is clear initially that the heart regularly pumps blood, makes noises, and so on. We are then in a good position to determine its function. Its function is whichever of its activities explains its survival. Contrast this with our situation in assigning functions to beliefs and desires. There is nothing that any particular belief or desire regularly does. And this is no accident; it is of the very essence of a belief or desire to interact with a large number of other beliefs and desires in response to a large number of possible situations resulting in a large number of other possible situations. How are we to determine *which* of these situations to assign to the particular belief or desire? Until we have answered this, we cannot even *begin* to argue about the function of the belief or desire? For that argument is about which of these assignments explains the survival of the belief or desire.[37]

It is hard to avoid the suspicion that Papineau's teleological theory has not so much answered as begged this difficult question: the situation has been assigned to the belief by feeding in our knowledge of the belief's content.

Papineau has done a service in proposing his teleological solution to a long-standing and fascinating problem: the nature of representation and intentionality. Unfortunately, the worries I have raised seem fairly decisive against the solution. They do not show, of course, that another teleological theory could

---

[35] So does Millikan ([1984], p. 97). The non-historical account of functions urged by Bigelow and Pargetter [1987] would avoid this problem. According to them something has a function in virtue of its *present propensity* to lead to survival and proliferation.

[36] He is setting aside desires for convenience (p. 72).

[37] This paragraph owes a lot to Fodor forthcoming.

not be made good. However, they do rather shake one's confidence that one could. Perhaps Millikan, who has proposed the most detailed teleological theory (1984, 1986, 1989), should increase our confidence. The complexities of her theory have so far discouraged a detailed and informed critique.[38]

MICHAEL DEVITT

University of Maryland

REFERENCES

BIGELOW, J. and PARGETTER R. [1987]: 'Functions', *Journal of Philosophy* 84, pp. 181–96.

BLOCK, N. [1986]: 'Advertisement for a Semantics for Psychology', in P. A. French, T. E. Uehling Jr. and H. K. Wettstein (*eds.*) *Midwest Studies in Philosophy, Volume X: Studies in the Philosophy of Mind*, Minneapolis: University of Minnesota Press, pp. 615–78.

BRANDOM, R. [1988]: 'Pragmatism, Phenomenalism, and Truth Talk ', in P. A. French, T. E. Uehling Jr. and H. K. Wettsein (*eds.*) *Midwest Studies in Philosophy, Volume XII: Realism and Antirealism*, Mineapolis: University of Minnesota Press, pp. 75–93.

DAVIDSON, D. [1984]: *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press.

DEVITT, M. [1974]: 'Singular Terms', *Journal of Philosophy* 71, pp. 183–205.

DEVITT, M. [1981]: *Designation*. New York: Columbia University Press.

DEVITT, M. [1984]: *Realism and Truth*. Oxford: Blackwell.

DEVITT, M. [1985]: Critical Notice of Evans 1982, *Australasian Journal of Philosophy* 63, pp. 216–32.

DEVITT, M. [1988]: 'Rorty's Mirrowless World', in P. A. French, T. E. Uehling Jr., and H. K. Wettstein (*eds.*) *Midwest Studies in Philosophy, Volume XII: Realism and Anti-Realism*, Minneapolis: University of Minnesota Press, pp. 157–77.

DEVITT, M. [1989]: 'A Narrow Representation Theory of the Mind', in S. Silvers (*ed.*) *Representation: Readings in the Philosophy of Mental Representation*, pp. 369–402. Dordrecht: Kluwer Academic Publishers. Reprinted in W. G. Lycan (*ed.*) *Mind and Cognition: A Reader*. Oxford: Basil Blackwell, 1990.

DEVITT, M. [1990a]: 'Meanings Just Ain't in the Head', in G. Boolos (*ed.*) *Method, Reason and Language: Essays in Honor of Hilary Putnam*. Cambridge: Cambridge University Press.

DEVITT, M. [1991]: 'Aberrations of the Realism Debate', *Philosophical Studies* 61, pp. 43–63.

DEVITT, M. Forthcoming a. 'Meaning Localism'.

DEVITT, M. Forthcoming b. 'Holism, Analyticity, and Revisability.'

DEVITT, M. and STERELNY, K. [1987]: *Language and Reality: An Introduction to the Philosophy of Language*. Oxford: Basil Blackwell.

DONNELLAN, K. S. [1972]: Names and Identifying Descriptions', in D. Davidson and G. Harman (*eds.*) *Semantics of Natural Language*, Dordrecht, Holland: D. Reidel Publishing Company, pp. 356–79.

DRETSKE, F. [1981]: *Knowledge and the Flow of Information.* Cambridge, Massachusetts: MIT Press.

DRETSKE, F. [1986]: 'Misrepresentation', in R. Bogdan (*ed.*) *Belief: Form, Content, and Function.* Oxford: Clarendon Press, pp. 17–36.

DRETSKE, F. [1988]: *Explaining Behaviour.* Cambridge, Massachusetts: MIT Press.

DUPRE, J. [1981]: 'Natural Kinds and Biological Taxa', *Philosophical Review* 90, pp. 66–90.

ENC, B. [1982]: 'Intentional States of Mechanical Devices', *Mind* 91, pp. 161–82.

EVANS, G. [1982]: in J. McDowell (*ed.*) *The Varieties of Reference.* Oxford: Clarendon Press.

FIELD, H. [1972]: 'Tarski's Theory of Truth', *Journal of Philosophy* 69, pp. 347–75.

FIELD, H. [1977]: 'Logic, Meaning, and Conceptual Role', *Journal of Philosophy* 74, pp. 379–409.

FIELD, H. [1986a]: 'Stalnaker on Intentionality', *Pacific Philosophical Quarterly* 67, pp. 98–112.

FIELD, H. [1986b]: 'The Deflationary Conception of Truth', in G. MacDonald and C. Wright (*eds.*), *Fact, Science, and Morality: Essays on A. J. Ayer's Language, Truth and Logic.* Oxford: Basil Blackwell, pp. 55–117.

FODOR, J. A. [1981]: *Representations: Philosophical Essays on the Foundations of Cognitive Science.* Cambridge, Massachusetts: MIT Press.

FODOR, J. A. [1984]: 'Semantics, Wisconsin Style', *Synthese* 59, pp. 231–50.

FODOR, J. A. [1987]: *Psychosemantics: The Problem of Meaning in the Philosophy of Mind.* Cambridge, Massachusetts: MIT Press.

FODOR, J. A. [1990]: 'Psychosemantics or: Where do Truth Conditions Come from?' in W. G. Lycan (*ed.*) *Mind and Cognition: A Reader.* Oxford: Basil Blackwell.

FODOR, J. A. Forthcoming, 'A Theory of Content'.

GODFREY-SMITH, P. [1988]: Review of Millikan 1984. *Australasian Journal of Philosophy* 66, pp. 556–60.

GODFREY-SMITH, P. [1989]: 'Misinformation', *Canadian Journal of Philosophy* 19, pp. 533–50.

GODFREY-SMITH, P. Forthcoming. 'Indication and Adaptation'.

GOULD, S. J. and LEWONTIN, R. [1979]: 'The Spandrels of San Marco and the Panglossian Paradigm: a Critique of the Adaptationist Program', reprinted in E. Sober (*ed.*) *Conceptual Issues in Evolutionary Biology: An Anthology.* Cambridge, Massachusetts: MIT Press, 1984, pp. 252–70.

GRANDY, R. [1973]: 'Reference, Meaning, and Belief', *Journal of Philosophy* 70, pp. 439–52.

GROVER, D. L., CAMP, J. L., JR. AND BELNAP, N. D., JR [1975]: 'A Prosentential Theory of Truth', *Philosophical Studies* 27, pp. 73–125.

KRIPKE, A. [1980]: *Naming and Necessity.* Cambridge, Massachusetts [A corrected version of an article of the same name (plus an appendix) in D. Davidson and G. Harman (*eds.*) *Semantics of Natural Language.* Dordrecht, Holland: D. Reidel Publishing Company , 1972.]

LEEDS, S. [1978]: 'Theories of Reference and Truth', *Erkenntnis* 13, pp. 111–29.

LOAR, B. [1982]: 'Conceptual Role and Truth-Conditions', *Notre Dame Journal of Formal Logic* 23, pp. 272–83.

LOAR, B. [1983]: 'Must Beliefs be Sentences?', in P. D. Asquith and T. Nickles (*eds.*) *PSA 1982, Volume 2.* East Lancing, Michigan: Philosophy of Science Association, pp. 627–43.

McDONALD, G. AND PETTIT, P. [1981]: *Semantics and Social Science*. London: Routledge and Kegan Paul.

McDOWELL, J. [1978]: 'Physicialism and Primitive Denotation: Field on Tarski', *Erkenntnis* 13, pp. 131-52.

McGINN, C. [1982]: 'The Structure of Content', in A. Woodfield (*ed.*), *Thought and Object*. Oxford: Clarendon Press, pp. 207-58.

MILLIKAN, R. G. [1984]: *Language, Thought and Other Biological Categories: New Foundations for Realism*. Cambridge, Massachusetts: MIT Press.

MILLIKAN, R. G. [1986]: 'Thoughts Without Laws: Cognitive Science with Content', *Philosophical Review* 95, pp. 47-80.

MILLIKAN, R. G. [1989]: 'Biosemantics', *Journal of Philosophy* 86, pp. 281-97.

MILLIKAN, R. G. [1990a]: 'Truth Rules, Hoverfiles, and the Kripke-Wittgenstein Paradox', *Philosophical Review* 99, pp. 323-53.

MILLIKAN, R. G. [1990b]: 'Speaking Up For Darwin', in B. Loewer and G. Rey (*eds.*) *Meaning in Mind: Fodor and his Critics*. Oxford: Basil Blackwell.

PAPINEAU, D. [1979]: *Theory and Meaning*. Oxford: Clarendon Press.

PAPINEAU, D. [1984]: 'Representation and Explanation', *Philosophy of Science* 51, pp. 55-73.

PAPINEAU, D. [1987]: *Reality and Representation*. Oxford: Basil Blackwell.

PUTNAM. H. [1975]: *Mind, Language and Reality: Philosophical Papers, vol. 2*. Cambridge: Cambridge University Press.

PUTNAM, H. [1978]: *Meaning and the Moral Sciences*. London: Routledge & Kegan Paul.

QUINE, W. V. [1970]: *Philosophy of Logic*. Cambridge, Massachusetts: Harvard University Press.

RORTY, R. [1979]: *Philosophy and the Mirror of Nature*. Princeton: Princeton University Press.

SCHIFFER, S. [1981]: 'Truth and the Theory of Content, in H. Parret and J. Bouveresse (*eds.*) *Meaning and Understanding*. Berlin: Walter de Gruyter, pp. 204-22.

SCHIFFER, S. [1986]: 'Stalnaker's Problem of Intentionality', *Pacific Philosophical Quarterly* 67, pp. 87-97.

SEARLE, J. [1983]: *Intentionality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press.

STALNAKER, R. [1984]: *Inquiry*. Cambridge, Massachusetts: MIT Press.

STALNAKER, R. [1986]: 'Replies to Schiffer and Field', *Pacific Philosophical Quarterly* 67, pp. 113-23.

STAMPE, D. W. [1979]: 'Toward a Causal Theory of Linguistic Representation', In P. A. French, T. E. Uehling Jr. and H. K. Wettstein (*eds.*), *Contemporary Perspectives in the Philosophy of Language*. Minneapolis: University of Minnesota Press, pp. 81-102.

STERELNY, K. [1983]: 'Natural Kind Terms', *Pacific Philosophical Quarterly* 64, pp. 110-25.

STICH, P. [1985]: 'Could Man Be an Irrational Animal? Some Notes on the Epistemology of Rationality', in H. Kornblith (*ed.*) *Naturalizing Epistemology*. Cambridge, Massachusetts: MIT Press, pp. 249-67.

WETTSTEIN, H. [1989]: 'Cognitive Significance Without Cognitive Content', in J. Almog, J. Perry and H. Wettstein (*eds.*) *Themes from Kaplan*. Oxford: Oxford University Press, pp. 421-54.