4

Naturalizing Content

PAUL A. BOGHOSSIAN

Introduction

The conviction that intentional realism requires intentional reductionism has the philosophy of mind in its grip. Thus, Jerry Fodor:

> I suppose that sooner or later the physicists will complete the catalogue they've been compiling of the ultimate and irreducible properties of things. When they do, the likes of *spin*, *charm*, and *charge* will perhaps appear on their list. But *aboutness* surely won't; intentionality simply doesn't go that deep. It's hard to see, in face of this consideration, how one can be a Realist about intentionality without also being, to some extent or other, a Reductionist. If the semantic and the intentional are real properties of things, it must be in virtue of their identity with (or maybe their supervenience on?) properties that are themselves *neither* intentional *nor* semantic. If aboutness is real, it must be really something else.[1]

It is worth noting – if only because it so seldom is nowadays – that this rationale for the naturalistic conviction begs a question that doesn't obviously deserve to be begged. Why, indeed, must we think that no property can be real unless it is identical with, or supervenient upon, the properties that appear in the catalogues provided by physics? There is, I think, no *obvious* answer.

For one thing, *identity* of intentional properties with physical properties would appear to be out of the question. Not only does nothing seem to be identical to anything else, but there are several, to all appearances decisive, reasons why the intentional in particular is not identical to the physical in particular. If the naturalistic conviction is to have any face plausibility whatever, then, it must be formulated as a supervenience claim. And what the naturalist needs is an argument why, in general, it is a condition on a property's being real that it supervene on the properties recognized by physics. There are, to be sure, specific local areas in which some sort of supervenience thesis seems correct. For example, mere reflection on the *concept* of a moral property reveals that moral properties weakly supervene on non-moral properties: someone isn't so much as competent with moral concepts, if he treats non-evaluatively equivalent cases differentially from a moral point of view. But this sort of deliverance – which, it is worth emphasizing, is in any case only of a *weak* supervenience thesis – does not appear to be forthcoming in general.[2] It is simply not true that mere reflection on the concept of an arbitrary property

discloses that property's supervenience – however weak – on the physical. In particular mere reflection on the concept of an intentional property doesn't.

If any of this is right, we are owed an explanation why we ought to believe in the supervenience thesis.[3] Let us pretend, however, and for the sake of argument, that we have been given one. Now, Fodor seems to believe that his commitment to the supervenience thesis requires him to provide

> a *naturalized* theory of meaning; a theory that articulates, in non-semantic and non-intentional terms, sufficient conditions for one bit of the world to *be about* (to express, represent, or be true of) another bit[4].

It will prove useful to pause a while here to reflect on the connection: how, precisely, does commitment to the supervenience thesis imply a commitment to a naturalized theory of meaning in Fodor's sense?

The answer is that a naturalized theory of meaning is needed – as I should like to put it – to render the supervenience thesis intelligible. In its absence, a supervenience thesis linking the intentional and the physical must be regarded as hopelessly mysterious and cannot be accepted. I shall explain.

A set of properties A supervenes on another set B just in case no two things can differ in their A-properties without differing in their B-properties. It follows that if A supervenes on B, then for every property P in the supervenience set A, there exists a property Q in the subvenient set B, which is a sufficient condition for it.[5] The relation of supervenience implies, in other words, that there are necessary connections between the properties that it relates. In particular, if intentional properties supervene on physical properties, then every intentional property has a physical property that necessitates it.

Now, we may appreciate the role of a naturalized theory of meaning by observing that, in the absence of further comment, a relation of supervenience between sets of distinct and highly disparate properties is puzzling. How could there be a set of necessary connections between such properties as being a certain configuration of molecules and believing that *Lully was a better composer than Purcell*, given the admittedly highly divergent characters of the properties involved? We are entitled to be mystified.

It is the point of a naturalized theory of meaning to help remove this mystification. By supplying a property that is both *incontestably physicalistic* and *recognizably sufficient for the instantiation of an intentional property*, a naturalized theory of meaning seeks to render intelligible the existence of necessary connections between the physical and the intentional and, thereby, the existence of a supervenience relation between them. It attempts to purchase a right to believe in the supervenience thesis.

So, a naturalized theory of meaning is what you need, if the naturalistic conviction has you in its grip. In this paper, however, I want to argue that, the naturalistic conviction notwithstanding, a naturalized theory of meaning is precisely what you are not likely to get. Not, at any rate, if what you mean by a naturalized theory of meaning is an information-theoretic semantics.[6]

Informational Theories of Meaning and Type 1 Situations

The Basic Formula

Let's assume, for ease of exposition, that we think in a 'language of thought': having the concept *cow* involves having a mental symbol – "cow" as it may be – which means *cow*. Now, an informational theory of meaning is the idea that the meaning of such a mental symbol is determined by the information it carries. The root idea – and the basis for all further refinements – is supplied by the following *basic formula*:

> S-events (e.g. the tokenings of symbols) express the property P, if (it's a law that) Ps cause S-events.[7]

The prospects for such a theory depend entirely on whether this basic formula can be converted, via the imposition of appropriate non-semantic and non-intentional constraints, into a plausible theory of meaning.

Why does the basic formula need to be modified at all? Consider some mental symbol S and suppose it means *horse*. Now we all make mistakes: we are all prone, when conditions are sufficiently unfavorable, to misidentifying items that we are presented with. So it can happen that, when presented with a deceptively horsey looking cow, I misclassify it as a horse: I believe *falsely* of some cow that it is a horse. But that is just to say that, on a given occasion, it is the property of being a cow that is nomically sufficient to cause a tokening of a symbol which means *horse*.

The basic formula, however, can make no sense of this. For according to it, a symbol expresses whatever property is nomically sufficient to cause its tokening; it follows, therefore, that since tokenings of S are also being caused by cows, S cannot simply mean *horse*, but must mean *cow or horse*. In other words, what intuitively seemed to be a case of a non-disjunctive symbol being applied falsely to something not in its extension, the basic formula would have us describe as a case of a disjunctive symbol being applied correctly to something in its extension.

In general, then, since applying a symbol falsely involves applying it to something not in its extension and since, according to the basic formula, every property that can cause the tokening of a symbol is, *ipso facto*, in the extension of that symbol, the basic formula can make no sense of error.

Errors, however, aren't the only way in which a symbol might get caused by something that isn't in its extension. The thought that cows are mammals might get caused by the thought that platypi aren't; this would be a non-erroneous tokening of the expression "cow" yet, nevertheless, not a tokening of the expression that is caused by the property it expresses. Here again, and for much the same reason as before, the basic formula yields the wrong result: it has "cow" meaning *cow or platypus-thought*, whereas "cow" means *cow* and no more.

In sum: Because under ordinary circumstances it is possible for a symbol to be caused by something that it does not express – either through error or through its occurrence in a sequence of non-labelling thoughts – the basic formula appears bound to deliver the wrong verdicts about the meaning of practically every expression in the language. Following tradition, we may call this the *disjunction problem*.[8]

It is a condition of adequacy on a causal theory of meaning that it solve this problem. Solving it requires selecting, from among all the laws that govern the tokening of a symbol S, that law which is *meaning-determining*: a property M, specifiable without the use of semantic or intentional materials, must be defined so that: possession of M by a nomic cause of a tokening of S is necessary and sufficient for that cause to be in the extension of S.[9]

Those causes of S not possessing M would then not count as in the extension of S and would be free, therefore, to cause tokenings of S that are false.

Meaning-Determining Causal Laws and Type 1 Situations

Suppose that both P and Q are nomic causes of tokenings of S. And suppose that S means P. The causal theorist must somehow select the P --> S law as the one that fixes S's meaning. He must specify a property M that is possessed only by the P --> S relation. How is this trick to be turned? As Fodor notes, the standard attempts to turn it exhibit a common feature: they attempt to select the P --> S law by defining a *situation* in which only Ps can cause tokenings of S. In other words, their common strategy is to attempt to define a situation in which only the referent of a symbol can cause its tokening; they then proceed to identify the fact that the symbol expresses a given property with the fact that it is that property that is solely responsible for causing tokenings of the symbol under that sort of situation. Let us call a situation in which nothing but the referent of a symbol can cause its tokening, a situation of type 1; and let us call theories that specify meaning in terms of such situations, type 1 theories.

Teleological Type 1 Theories

The literature contains a number of proposals concerning the identity of situations of type 1, of which the most influential is the *teleological* proposal. The idea here is that there is a set of Normal conditions, specified by evolutionary biology – hence, one presumes, naturalistically – under which our cognitive mechanisms are functioning just as they are supposed to. The teleological proposal is that we equate type 1 situations – situations under which nothing but the referent of a symbol can cause its tokening – with Normal conditions so specified.

Its prima facie appeal notwithstanding, the teleological proposal suffers from a number of severe problems, some of which are outlined very effectively by Fodor.[10] First, it rests on an incredible conception of evolutionary selection assuming, for instance, that cognitive mechanisms are never selected for the purpose of hiding, rather than tracking, the truth. Second, it does not really solve the disjunction problem because facts about teleological function go soft at precisely the point at which the disjunction problem arises. And, finally, the root idea upon which the teleological proposal depends – the assumption that "when things go right" S will be tokened only in application to its referent – seems simply wrong. As Fodor observes, a glaring counterexample is provided by the humdrum phenomenon of one thought causing another. In the course of musing about horses I might be led to muse about cows. This would be a case of a

horse-thought causing a cow-thought, hence, not a case of a cow-thought being caused by its referent. But also, surely, not a case where anything has gone wrong from a teleological point of view. The point is that, even if evolutionary biology could help define a set of conditions which abstract from sources of *error*, it is hardly likely to help define a set of conditions which abstract from *thinking*. It is hardly likely, therefore, to deliver what a type 1 theory of meaning needs: a nonintentionally, nonsemantically specified situation in which only the referent of a symbol can cause its tokening.

Fodor's Theory: Assymmetric Dependence

Now, all this is, I think, pretty much decisive against teleological versions of type 1 theories. In fact, however, Fodor disapproves not merely of *teleological* versions of type 1 theories, but of *any* type 1 causal theory of meaning. He says:

> what's *really* wrong with teleological theories of content . . . is the idea that [in a certain kind of circumstance] the tokens of a symbol can have only one kind of cause – viz. the kind of cause that fixes meaning. . . . But surely this underestimates . . . the *robustness* of meaning: In actual fact, "cow" tokens get caused in *all sorts* of ways, and they all mean *cow* for all of that . . . If there's really going to be a causal theory of content, there has to be some way of picking out *semantically relevant* causal relations from all the other kinds of causal relations that the tokens of a symbol enter into. *And we'd better not do this by idealizing to contexts of etiological homogeneity.* [Final emphasis mine.][11]

And he commends his own view (to which we will come below)

> for having the desirable property of not assuming that there are such things as type 1 situations; in particular, it doesn't assume the possibility of specifying – naturalistically and otherwise non question beggingly – circumstances in which it's semantically necessary that only cows cause "cows."[12]

Well, how are we to naturalistically specify a symbol's meaning-determining causal relations without "assuming that there are such things as type 1 situations." The crux of Fodor's proposal is neatly summed up in the following passage:

> Cows cause "cow" tokens, and (let's suppose) cats cause "cow" tokens. But "cow" means *cow* and not *cat* or *cow or cat* because it's being a law that *cats cause "cow" tokens depends on it's being a law that cows cause "cow" tokens, but not the other way around.* "Cow" means *cow* because, as I shall henceforth put it, non cow-caused "cow" tokens are *assymmetrically dependent upon* cow-caused "cow" tokens.[13]

How, precisely, is this to be understood? It's not, I'm afraid, all that easy to say.

The trouble stems, on the one hand, from Fodor's insistence that the theory is best understood directly in terms of the nomic concepts in which it's couched; and, on the other, from the conspicuous absence of any account of those concepts. Fodor bemoans counterfactual analyses of law and possible worlds analyses of counterfactuals, castigating them as exemplifying the "sort of reductive move that is always blowing up in philosophers' faces." It is hard, in view of the less than spectacular achievements of the analytic enterprise, not to sympathize with this. But the fact remains that we need *some* handle on how the theory is to be understood and those accounts, warts and all, are the best we've got at the present

time. Having acknowledged Fodor's reservations, then, I propose henceforth to ignore them. I shall expound his theory in terms of the language of counterfactuals and possibilia of which he disapproves. In proceeding in this fashion, I am emboldened by the fact that, although I may be flouting Fodor's ideology, I am not flouting his actual practice:

> Having gotten all that off my chest, I shall join the crowd and talk counterfactuals from time to time, *faute de mieux*. And since it's widely supposed that talk about counterfactuals itself translates into talk about possibilia, I shall sometimes equate "there is a nomic dependence between the property of being a Y and the property of being a cause of Xs" with "Ys cause Xs in all (nearby? see below) nomologically possible worlds."[14]

What, then, is the proposed sufficient condition for S's meaning P? A first counterfactual approximation would have it that S means P if it satisfies the following condition. If P ceases being capable of causing S tokens, then every other cause of S is rendered similarly impotent; but not the other way around: any non-P could cease being capable of causing S tokens without this affecting P's ability to do so.

It may strike you, right off the bat, that this suggestion couldn't be right because it would have us thinking not about, say, cows but about their proximal projections. For it's presumably true that no cow would cause a "cow" token except as it produces some proximal projection. And it's plausible, moreover, that such proximal stimuli would be sufficient to cause "cow" tokens even if they weren't produced by cows. If all this is true, then it would appear that the cow-prompted "cow" tokens are asymmetrically dependent on the disjunction of proximal cow projections, and hence, on the terms of the theory on offer, that "cow" means not *cow* but that inner disjunction instead.

Fodor, however, won't admit any of this and for reasons to which we shall have occasion to return:

> It might still be said, however, that the dependence of cow thoughts on asymmetrically dependent on their dependence on *disjunctions* of proximal cow projections; distal cows wouldn't evoke COW tokens but that they project proximal whiffs or glimpses or snaps or crackles OR. . . . well, OR what? Since, after all, cow spotting can be mediated by theory to any extent you like; the barest whiff or glimpse of cow can do the job for an observer who is suitably attuned . . . To the extent that this is so, just about *any* proximal display might mediate the relation between cows and cow-thoughts for some cow-thinker on some or other cow-spotting occasion . . . So barring appeals to open disjunctions, it seems likely that there is just no way to specify an array of proximal stimulations upon which the dependence of cow-thoughts on cows is asymmetrically dependent.[15]

The bruited objection is no good, Fodor says, because it ignores the holistic and open-ended character of belief fixation. Since just about any theory can mediate the fixation of beliefs about cows, just about any proximal stimulation can cause a "cow" token; and open disjunctions of properties are not eligible to serve as the referents of our primitive expressions.

Now, there is room for doubting the effectiveness of this rejoinder, but I propose for the time being to accept it. The effect is to restrict the referents of our predicates to their *distal* causes; and so, on Fodor's view, S means P, if P is the

distal cause of S tokens on which all the other distal causes of S tokens depend. Henceforth, let it be so understood.

Another prima facie difficulty with the theory will help specify it further. So far we have it "cow" has cows it its extension but not cats, because if the cow --> "cow" connection were to break, then so would the cat --> "cow" connection, but not vice versa. Put in the language of possibilia, this seems best interpreted as suggesting that "cow" has cows in its extension and not cats, provided that, although there are worlds in which *cow* can cause a "cow" token but *cat* can't, there are no worlds in which *cat* can cause a "cow" token, but *cow* can't.

This, however, can't be precisely what's meant. The point is, of course, that even by the theory's own lights, there have to be *some* worlds in which the property of being a cat can cause "cow" tokens even if the property of being a cow can't, for there presumably are some worlds in which "cow" means *cat*. A more accurate statement of the theory, then, would have it that what's required

> is not that cows cause "cow"s in *every* nomologically possible world where Xs cause "cow"s. Rather, what's required is just that there be a world W such that (i) in W, cows cause "cow"s and non-cows don't; and (ii) W is nearer to our world than any in which some non-cows cause "cow"s and no cows do.[16]

Asymmetric Dependence and Type 1 Situations

A really plausible theory of content, Fodor says, would not be a type 1 theory: it would not commit itself to the existence of naturalistically specifiable circumstances in which a symbol can get caused only by its referent. As we shall see later on, there is much to be said in favor of this claim.

The trouble is, however, that in all essential respects Fodor's own theory would appear to be a type 1 theory and, hence, itself vulnerable to the attendant problems. Let me explain.

A traditional type 1 theory has it that S means P because, in a favored naturalistically specified circumstance C, only P can cause S tokens. In such theories, then, the meaning of a symbol is specified directly in terms of *the causes that tokenings of that symbol may have in a certain type of situation*. Now, admittedly, Fodor's theory does not look like that. On his view, the meaning-determining cause is picked out not in terms of its relation to certain kinds of situation, but in terms of its relation to the symbol's other causes: that cause on which the other causes of the symbol asymmetrically depend, is the symbol's meaning. I propose to argue, however, that on the most favorable understanding of the theory on offer, if symbol S possesses asymmetric dependence base P, then there exists a world in which *only* P can serve as a cause of tokenings of S. I propose to argue, in other words, that even Fodor's theory is a type 1 theory, because even asymmetric dependence implies the existence of type 1 situations for symbols.

Suppose that among S's causes are properties P and Q. Now, clearly, a distribution of nearby worlds according to which there are some worlds in which S has only P as a cause, but no worlds in which it has only some non-Ps as a cause, would be *sufficient* to ground P's status as S's asymmetric dependence base; the question is whether such a distribution is *necessary*. Well, what *other* distribution

of worlds would also ground the fact in question? The only remotely plausible suggestion is that a distribution of nearby worlds which contains worlds in which both P and Q can cause S tokens, and worlds in which both P and R can, but no world in which only P can and no world in which only non-Ps can, would equally suffice to certify P's status as S's asymmetric dependence base.

And even *it* doesn't work. For there is nothing to distinguish the envisaged distribution from one in which the real asymmetric dependence base for S's causal relations is not P but, rather,

(P & R)v(P & Q).

Without a world in which only Ps can cause S, the most we would be entitled to claim is that it seems impossible to break any particular non-P causal relation without enforcing at least one of the others (in combination with P). And this falls far short of saying that it is precisely the P --> S connection that cannot be broken without breaking all the others.

Now, it might be objected that we are not *guaranteed* that (P & R)v(P & Q) will be a candidate asymmetric dependence base for S whenever P is. That is true, but irrelevant. It is true because there is, in general, no guarantee that whenever P --> S is a law and R --> S is a law that (P & R)v(P & Q) --> S will be a law also. It is irrelevant because the point at issue requires only that there be at least *one* case in which (P & R)v(P & Q) --> S is a law and, hence, (P & R)v(P & Q) a candidate asymmetric dependence base for S. Since Fodor's theory says that it is a *sufficient* condition of S's meaning X that X be S's asymmetric dependence base, any case in which S has both P and (P & R)v(P & Q) as candidate asymmetric dependence bases will be a case in which his theory yields either the wrong result or an indeterminiate one. Hence, any such case would constitute a straightforward counterexample to his theory. So, as I have said, all that's required is that there be at least one case in which (P & R)v(P & Q) --> S is a law whenever P --> S is and R --> S is; and it seems pretty obvious that there will be plenty of such cases.

To be sure to have eliminated this entire class of counterexamples, then, Fodor's theory must be read as follows: S means whatever properties are possible causes of S tokens in the closest world with the smallest subset of S-token causes in the actual world. Or, in the case of an expression with a *single* property P as referent, S means P if P is the sole cause of S tokens in the closest single-cause world. Or, to put this yet another way: Fodor's theory is a type 1 theory where the type 1 situation for S is specified as being: the closest world in which S can get tokened by only one sort of cause.

Now, Fodor has admitted that asymmetric dependence implies a commitment to the existence of circumstances in which a symbol can get caused only by its referent. But he has denied that this commits him to the sorts of situation envisaged by traditional type 1 theories.

> . . . real type 1 theories say that situations in which only Xs cause Ss are nomologically possible. Whereas my theory says only that the dependence of Ss on Xs is nomically dependent on their dependence on Xs. If you spell this out in terms of possible worlds, then what you get is that worlds in which Xs cause Ss but nonXs don't, are closer, ceteris paribus, than worlds in which nonXs cause Ss but Xs don't. But, notice, ALL THE WORLDS THAT FIGURE IN THIS RELATION ARE,

> BY ASSUMPTION, NOMOLOGICALLY IMPOSSIBLE. In particular, even the nearest of these worlds breaks the connection between nonXs and Ss, and that connection is, by assumption lawlike. . . . So, the difference between my story and a true type 1 story is that true type 1 stories hold that type 1 situations are nomologically possible; and my story doesn't.[17]

I must admit to being somewhat confused by this response, but for reasons that will detain us too long to look into. So I wish simply to grant the point: traditional type 1 situations are conceived as nomologically possible, whereas Fodor's type 1 situations aren't.[18] This particular difference won't matter to anything that's to follow.

For what follows is a set of objections to theories that are based on the existence of type 1 situations, regardless of their nomological status. I shall argue for two main claims. First, that there couldn't be naturalistically specifiable circumstances under which nothing but the referent of a symbol can cause its tokening. And, second, that even if there were such conditions, we could never be in a position to *certify* that there were. I'll end by returning to Fodor's theory and showing explicitly that it does not provide a convincing naturalistic sufficient condition for meaning.

Could there be Naturalistic Conditions of Type 1? – Verificationism and Natural Kind Terms

Causal Theories and Verificationism

I propose to begin by arguing that any attempt at specifying a type 1 situation for a symbol S, however refined, is bound to fail: *any* specified situation will contain possible causes of S tokens not in the extension of S.

The reason is painfully obvious: any situation in which X is a possible cause of my S-tokens is also a situation in which any other property Y, indistinguishable from X in all physically possible circumstances accessible to me, is *also* a possible cause of my S-tokens. Since there are no physically possible circumstances accessible to me in which X and Y can be told apart, any circumstance in which X can cause my S tokens is also a circumstance in which Y can. It follows, therefore, that on a type 1 theory, if S has X in its extension, then it also has all these other "X-equivalent" properties equally in its extension. But is this plausible? Is it really true that my having a symbol that means X but not Y depends on my being able to tell Xs and Ys apart? After all, it surely doesn't follow from the fact that Xs and Ys can only be told apart in worlds that are too far for *me* to get to, that being X and being Y are the same property. But, then if the difference between being X and being Y is real, then so too, presumably, is the difference between being X and being (X or Y). And if this difference is real, then why shouldn't we be able to think in ways that respect that difference?[19]

Natural Kind Terms

It will prove useful to present a concrete case of an expression that has X in its extension, but not all its X-equivalent cousins.

I have chosen to focus on natural kind expressions. The choice is particularly apt, for three reasons. First, because natural kind terms are of incontestable importance. Second, because there is the fairly widespread conviction that recent philosophy of language has made significant progress in understanding their *intuitive semantics*. One of the many difficulties that bedevil discussion of the naturalization of semantic properties is the absence, in a significant range of cases, of an intuitively correct semantics to naturalize: we simply don't know what extensions to assign to various expressions under various counterfactual circumstances, and, hence, don't know whether particular naturalistic proposals capture those extensions correctly. Through the recent work of Kripke and Putnam, however, we have in hand the rough contours of a convincing and widely accepted description of the semantic functioning of natural kind expressions; and so a meaningful test becomes possible. Finally, and perhaps most importantly; because causal theorists themselves invariably illustrate their theories with the use of natural kind expressions. It is widely supposed by the proponents of such theories, that causal theories are at their strongest in application to natural kind expressions. It is easy to explain why. The point is that it is part of the moral of the Kripke/Putnam story about natural kind expressions, that they have an intuitively *externalist*, indeed *causal*, semantics: according to such views, it is true independently of any attachment to informational theories of meaning, that causal relations play an important part in fixing the reference of natural kind expressions. So what better candidate for naturalization via a causal theory than a natural kind expression?

In fact, I wish to argue, just the opposite is true; the semantics of such terms, properly understood, resists naturalization by a causal theory.

We need to start off with an account of their intuitive semantics. What, in rough outline, is the Kripke/Putnam account of the reference of such terms as "water," "cow," or "gold"? The essentials are neatly summarized in the following passage from Kripke. A natural kind concept, says Kripke, is the concept of

> *that kind of thing*, where the kind can be identified by paradigmatic instances. It is not something picked out by any qualitative dictionary definition . . .[20] The same observations hold for such a general term as 'gold.' If we imagine a hypothetical (admittedly somewhat artificial) baptism of the substance, we must imagine it picked out as by some such 'definition' as, 'Gold is the substance instantiated by the items over there, or at any rate, by almost all of them' . . . terms for natural kinds (e.g., animal, vegetable, and chemical kinds) get their reference fixed in this way; the substance is defined as the kind instantiated by (almost all of) a given sample.[21]

According to this plausible proposal, then, a natural kind expression S functions *as if* it had been introduced via a reference-fixing description of the following form: S names the naturally delineated kind exemplified by all, or most, of certain local exemplars.[22] Several aspects deserve comment.

First, there is the commitment that the expression name a naturally occurring kind, hence that some sort of basic naturally explanatory property unite all the things that are correctly said to fall in its extension.[23]

Which kind? The kind exemplified by all, or better *most*, of certain local samples. Not simply "all," for we need to allow for impurities in the local samples. "Gold" refers to the substance with atomic number 89, even though not

everything in the local samples has that atomic number: some of the samples are iron pyrites.

Finally, what if there is no dominant natural kind in the local samples? If investigation uncovers that there is no single hidden structure uniting the local samples, but that there are two (or so) well-defined such structures, then there appears to be a temptation to say that the expression has both of those structures in its extension. Thus, "jade," for example, applies to two minerals, jadeite and nephrite, which are chemically quite different. But if it turns out that nothing but a messy motley of basic properties unites the local samples, then the correct conclusion would appear to be that the term fails to refer: there simply is no naturally delineated kind for the term to refer to.

Does an information-theoretic semantics correctly capture these intuitive results? I shall argue that it does not, and could not.

Causal Theories and Natural Kind Terms

Consider "water." "Water" functions so as to name all and only the stuff bearing the same-liquid relation to all, or most, of such-and-so local samples. As it happens, that stuff is H_2O; so "water" refers to H_2O. It has all and only H_2O molecules in its extension. It does not have such possible substances as XYZ in its extension because XYZ is not of the same kind as H_2O, even though an ordinary speaker exposed to XYZ would call it "water."

Now, XYZ does not provide a straightforward counterexample to a type 1 informational theory of meaning, because it is part of the standard story about XYZ that there are physically accessible circumstances under which XYZ is distinguishable from H_2O. So there presumably are worlds where you would call H_2O "water" but not XYZ – namely, in those worlds where you could tell them apart.

But consider a substance – ABC – which is just like XYZ in being kind-distinct from H_2O, but yet which is distinguishable from H_2O only in circumstances that are physically inaccessible to humans. Let's suppose that this is because ABC would exhibit its distinctive characteristics only in gravitational fields of such intensity – characteristic of the interior of black holes – that nothing as complex as a human body could survive in them long enough to perform even the fleetest measurement; we needn't be too fussy about the physical details. ABC exists in abundance on Pluto, and hence is a potential, even if not actual, cause of "water" tokens.

Is such a substance coherently imaginable? Well, is there anything in physics that precludes the possibility of kind-distinct molecules that behave in precisely the same ways *in all circumstances physically accessible to human beings?* I know of no theorem of physics that precludes this. Which is not, if you think about it, all that surprising. For, remember, what's at issue is not the possibility of kind-distinct molecules that behave in exactly the same way in *all* physically possible worlds, but only the possibility of kind-distinct molecules that behave in exactly the same ways in all the physically possible worlds *that we can get to*, which is a rather different matter. And so physics couldn't preclude the possibility of such substances, because physics doesn't taxonomize molecules with reference

to our biological and medical limitations. So far as I can tell, then, there is absolutely no reason to believe that two molecules *couldn't* differ in respect of the sorts of property that make for a difference in kind, if they are distinguishable only in situations that are inaccessible to human beings.

Yet standard intuition would have it, I submit, that ABC is no more in the extension of "water" than XYZ is. "Water" is, as we have said, a natural kind term; it either denotes the non-disjunctive property exemplified by most paradigm local samples, or it denotes nothing. This kind, as it happens, is H_2O, not XYZ or ABC. Yet pure informational theories cannot respect this result. Since such theories equate what a term means with what could cause their tokening under a certain sort of physically accessible circumstance, and since there is no physically accessible circumstance under which H_2O could but ABC couldn't cause a tokening of "water," pure informational theories have to conclude – contrary to intuition – that both H_2O and ABC are in the extension of "water."

The problem, of course, is general. Provided only that it is consistent with the laws of physics that there could have been substances, kind-distinct from the ones that are actual, but which would manifest their distinctive features only in circumstances that are inaccessible to humans, it will follow, on a pure informational theory, that all our 'natural kind terms' have wildly disjunctive extensions, and hence are not really natural kind terms after all. And I would be very surprised if it weren't consistent with physics that there could be such substances.[24]

The Actual History Condition

Now, as I have already indicated, Fodor has expressed considerable sympathy with this sort of worry. Indeed, he has gone so far as to propose a modification to his own pure informational theory in an attempt to meet it.[25]

> [Pure informational] theories distinguish between concepts only if their tokenings are controlled by different laws. Hence only if different counterfactuals are true of their tokenings. Hence only if there are (possible) circumstances in which one concept would be caused to be tokened and the other would not . . . That is how you get from informational semantics to verificationism . . . Correspondingly, the way you avoid the verificationism is: You relax the demand that semantic relations be construed solely by reference to subjunctive conditionals; you let the actual histories of tokenings count too.[26]

Fodor's idea is to block the objectionable verificationist consequences of a pure informational theory by adding an 'actual history' condition: it is now to be required not merely that it be a law that Xs cause S-tokenings, if S is to mean X; it is further required that some S-tokenings *actually* be caused by Xs. And this would appear to rule out the sort of case we have just been contemplating: since it is true by assumption that ABC is only a *potential* cause of "water" tokens, it couldn't actually have caused any tokenings of "water," and, hence, doesn't satisfy the actual history condition. It is, therefore, ineligible to be in the extension of "water," just as intuition requires. So everything would appear to be OK.

Unfortunately, matters are not quite so simple. For one thing, the actual history condition is purchased at a terrible cost. For another, it doesn't ultimately help

with the problem about verificationism. And, finally, even if we ignored the issue about verificationism, it still doesn't yield the right meaning ascriptions. All that, in any case, is what I propose to argue next.

It is one of the virtues of a pure informational theory that it can treat "unicorn" in just the way it treats every other syntactically primitive symbol. Since, according to such theories, all that semantic relations care about are nomic connections between properties, and since there can be nomic connections between uninstantiated properties, such theories would appear well-poised to explain – or at least not precluded from explaining – how it is possible for us to represent unicorns, even though there are actually no unicorns. And this seems highly desirable because, on the face of it anyway, there would appear to be no deep connection between being a primitive concept and being instantiated; there seems no reason to believe that every concept that has an empty extension in the *actual* world will turn out to be *complex*.

The actual history condition, however, incurs that implausible commitment. Since it requires that an instance of some property actually have caused a tokening of a symbol that has it in its extension, it must hold that all uninstantiated concepts are, appearances to the contrary notwithstanding, complex. But just what is the (non-question-begging) argument which shows that you couldn't get a primitive concept that was uninstantiated?

The second difficulty with the actual history condition is that it doesn't really help with the problem about verificationism.

Partly for heuristic reasons, I stipulated that ABC was to be found only on Pluto, and not on Earth. But this is, I think, completely inessential to the point it is meant to illustrate. The basic intuition behind the anti-verificationist objection, to repeat, is that there is no reason whatever – intuitive or otherwise – to believe that having a word S that means X but not Y depends on being able, in principle, to tell Xs apart from Ys. If the difference between being X and being (X or Y) is real, why shouldn't we be able to think in ways that respect that difference? But there is absolutely nothing in this intuition which depends on Y's being *remote*. The intuition persists even on the supposition that both of the substances in question are involved in the actual causal history of S-tokenings. Fodor says:

> The mixed theory is itself just a soupçon verificationist, but only in a way that might surely be considered untendentious. We used to have to say that "X"'s meaning X requires the nomological possibility of distinguishing X from any property that would cause "X"s *if it were instantiated* . . . Now all we require is that it be possible to distinguish X from any property that is *actually* instantiated in the causal history of "X"s.[27]

But I don't see that this is any less tendentious than the verificationism entrained by the original pure version. To be sure, the present proposal cuts down on the number of properties that must be distinguishable, if they are to be ineligible to count as in the extension of the term in question. But the end result is no more plausible: the theory continues to deliver verdicts that are at odds with the intuitive semantics of natural kind terms.

We may illustrate this by describing a situation in which ABC is instanced locally. Trace quantities of it, let us suppose, exist in our lakes and rivers and in

the atmosphere. It is sometimes, even if rarely, the cause of "water" tokenings. It remains, as before, however, indistinguishable from H_2O in any circumstance accessible to humans. Thus, there is no possible circumstance under which H_2O can cause "water" tokens and ABC can't.

Standard intuition would have it, I submit, that ABC is still not in the extension of "water." ABC remains a different *kind* of stuff from H_2O; and the intuitive semantics of "water" is such that, if it denotes anything at all, it denotes the kind exemplified by most of the paradigmatic local samples. ABC, however, is not exemplified by most of the local samples; by stipulation, it exists only in trace quantities. It stands to water like iron pyrites stand to gold: it is an impurity. Hence, it does not count as within the extension of "water."

Still, this is a judgment that the mixed theory cannot respect. ABC is an occasional cause of "water" tokens, so it satisfies the actual history condition. Since it is not physically distinguishable from H_2O, there is no physically possible circumstance in which H_2O can and ABC can't cause tokenings of "water." So, contrary to intuition, it counts as in the extension of "water."

In sum: an informational theory of meaning is necessarily verificationist and the implausibility of this verificationism can be vividly illustrated through the consideration of expressions – like natural kind terms – which intuitively possess non-disjunctive extensions but which are assigned wildly disjunctive extensions by such theories.[28]

Could you Recognize Naturalistic Conditions of Type 1?

I have been arguing that it won't prove possible to naturalistically specify a situation of type 1 because, even if you could naturalistically specify a situation in which all *verifiable* sources of error had been removed, you couldn't, in the nature of things, specify a situation in which all *unverifiable* sources of error had been removed. But since unverifiability doesn't correctness make, the most refined attempt at a naturalistic specification of a type 1 situation is bound to fail.

I now want to argue that, even if, *per impossibile*, there were naturalistic situations of type 1, we could never be in a position to recognize that there were. If, by a stroke of vastly implausible luck, we were to stumble onto a specification of such a condition, we could never be entitled to conclude that we had. If this is true, then there can be no point in attempting to develop such theories.

Type 1 Conditions and Belief Holism

The single most impressive reason for being skeptical about the existence of situations in which only the referent of a symbol can cause its tokening has to do with what we may call the *holistic character of belief fixation*.[29] I shall explain.

Under normal circumstances, belief fixation is mediated by background theory – what contents a thinker is prepared to judge depends upon what other contents he is prepared to judge.[30] And this dependence is again typically, arbitrarily robust: just about any stimulus can cause just about any belief, given a suitably mediating set of background assumptions.

So, for example, suppose you think, in response to some stimulus or other, "Lo, a magpie." Need there really be a magpie in respect of which this is thought? Of course not. It could be that you saw a currawong, but believing there to be no currawongs in Australia, and believing this to be Australia, you concluded it was a magpie. Or perhaps you heard a currawong call and you believe that that is just what magpies sound like. Or then, again, perhaps you believe that whatever the Pope says goes and the Pope says that this presented currawong is a magpie. Or . . . The point is that magpie beliefs can be mediated by theory to any extent you care to imagine. The thought that something is a magpie can get triggered by a currawong in any of an *indefinite* number of ways, corresponding to the potentially indefinite number of background beliefs which could mediate the transition.

We have had occasion to remark on this doctrine before and to note its importance for Fodor's purposes. The point surfaces, as we saw, in the course of showing that it doesn't follow from the asymmetric dependence story about meaning that the intentional content of "cow" is not *cow*, but rather some disjunction of proximal cow projections. Just about any cow-projection can cause a cow-thought, Fodor argues, since just about any theory can mediate the fixation of beliefs about cows. That is why cow-projections are not eligible to serve as the referents of our expressions. But by the same token, just about any *non-cow* can cause a cow-thought too, if just about any theory can mediate the fixation of beliefs about cows. Indeed, it follows, and is in any case independently plausible, that just about any property can cause just about *any* thought, given a suitable set of background assumptions.

But now we are in a position to see why the existence of naturalistically specifiable situations of type 1 should appear so wild. A type 1 situation is, by definition, a situation in which, if the concept *magpie* gets tokened, then it's nomologically necessary that that tokening was caused by a magpie. Specifying a situation of type 1 naturalistically will involve, therefore, specifying a situation, in non-semantic and non-intentional terms, in which one can think "Lo, a magpie" only in application to magpies.

But the point about the holistic character of belief fixation suggests that *anything* can cause the thought "Lo, a magpie" in just about *any* type of situation provided only that the appropriate background beliefs are present. And, as I have argued, these background beliefs could be just about anything. Consequently, specifying a situation in which no non-magpie can cause the thought "Lo, a magpie" involves, at a minimum, specifying a situation that ensures the absence of all the beliefs which could potentially mediate the transition from non-magpies to magpies. Since, however, there looks to be a potential infinity of such mediating background clusters of belief, a non-semantically specified situation of type 1 would consist in the specification of a situation in which it has somehow been non-semantically guaranteed that none of this potential infinity of background clusters of belief is present. And it appears utterly incredible that there should be such a specification.

The impact of this observation on various *particular* proposals is transparent. For example, the observation that type 1 conditions must exclude a potentially indefinite array of frustrating background beliefs provides us with one more way of saying what is wrong with a teleological causal theory: it is simply not plausible

to suppose that our cognitive mechanisms' functioning as they are supposed to is sufficient to ensure the absence of any untoward background theory. Hence, it is not plausible to claim, as a teleological theory does, that when they are so functioning, it will prove impossible to think "Lo, a magpie," except in application to magpies. But does the observation provide us with a more general reason for doubting the feasibility of a type 1 style causal theory of meaning? I think it does.

Suppose you come up with a naturalistic specification of some situation C, which you conjecture to be a type 1 condition: under C, nothing but the referents of symbols can cause their tokenings. Could you show that your conjecture was in fact true? I want to argue that you could never non-question-beggingly certify that C is a condition of the required kind, even if it is one.

The worry that needs allaying is that the specified condition is consistent with the presence of background beliefs which would frustrate the covariation between symbol tokenings and their referents – magpie beliefs and magpies, as it may be. Putting this worry to rest requires showing that the situation's being C is enough to ensure the truth of the following open conjunction:

(*) $-Bel_1$ & $-Bel_2$ & $-Bel_3$ & . . .

where the Bel_i stand for the various clusters of background beliefs which could potentially frustrate the connection between being a magpie and the tokening of an expression which refers to it. Well, could C ever be recognized as sufficient for the truth of (*)?

Recognizing that a particular naturalistic condition ensures the absence of a set of beliefs, requires possession of naturalistic *necessary* conditions for those beliefs; we need to be able to tell that nothing in the naturalistic condition in question adds up to one of the beliefs in question. Now, we may, for present purposes, assume that we possess naturalistic necessary conditions for something's being a *belief*. But possession of naturalistic necessary conditions for being the *belief that p* requires possession of naturalistic necessary conditions for something's *meaning p*.

Now, suppose that the envisaged C condition is in fact a type 1 condition for *magpie*. Under the terms of this assumption, there would be naturalistic *sufficient* conditions for meaning: type 1 causal theories of meaning, remember, are advertised as providing no more than sufficient conditions for meaning. However, what's needed to certify that C is a type 1 condition for *magpie* is a certification that it is sufficient for the truth of (*); and what's needed to certify that it is sufficient for the truth of (*) is a set of naturalistic *necessary* conditions for meaning. So, we can't certify that C is a type 1 condition for *magpie*, even on the assumption that it is one. So, we can't certify that it is one.

Suppose, however, that the causal theorist sheds his modesty and promotes his theory as providing both necessary and sufficient conditions for meaning. As before, we have a specification of C, but now it is a clause in a *necessary and sufficient* condition for S meaning M. Could we now certify that C is a type 1 condition for *magpie*, on the assumption that it is one?

Well, on the assumption that it is one, there would be a set of necessary and sufficient conditions for an expression's expressing a certain property. Unfortunately, even so, we would still not be in a position to certify that C is a type 1

condition. The trouble is that proposition (*) is not finitely stateable: there is no finite way to state what beliefs the causal theorist must exclude before he may be assured of the desired concomitance of magpie beliefs and magpies. Literally any belief can frustrate the desired connection. So, there is no way to certify that C is sufficient for the truth of (*), even granted a set of naturalistic necessary and sufficient conditions for meaning. So, there is no way to certify that C is a type 1 condition for *magpie*.

If these considerations are correct, then there can be no point in constructing type 1 naturalistic theories of meaning. Even if, *per impossibile*, we were somehow to stumble onto such a theory, we could never convince ourselves that we had. For perfectly general reasons, we could never non-question-beggingly certify that a particular naturalistic condition was in fact a type 1 situation.

Asymmetric Dependence Revisited

If any of this is true, then Fodor's asymmetric dependence proposal could not constitute a convincing naturalistic sufficient condition for meaning. For as we have seen, Fodor's theory is a type 1 theory in disguise: to say that P is an asymmetric dependence base for S is simply to say that P is the sole cause of S tokens in the closest world where S has a single cause. And so, since it is the burden of the preceding considerations that there can be no convincing naturalistic specification of a type 1 situation, either Fodor's criterion yields the wrong meaning ascriptions, or it is not naturalistic in character.

Nevertheless, since arguments at this level of generality can seem disengaged from their targets, and since it may in any event prove independently edifying, I now propose to run through the relevant dialectic with explicit reference to the details of Fodor's particular proposal. I propose to show directly, in other words, that Fodor's asymmetric dependence condition for S's meaning P is either not naturalistic, or not sufficient for S's meaning P.

Let's go back to H_2O and XYZ. As before, "water" is undoubtedly a natural kind term, so it has H_2O in its extension but not XYZ; and, as before, it is true by stipulation that both H_2O and XYZ are nomically sufficient to cause tokenings of "water." What makes it true, according to Fodor, that "water" means H_2O and not XYZ, is that the closest world in which "water" tokens can get caused by one and not the other of these properties, is a world in which it is H_2O that can do the causing, and not XYZ.

But *closer* with respect to what? Clearly, everything depends on whether the relevant similarity relation can be specified non-question-beggingly – without the benefit of sidelong looks at the meanings of the expressions in question. What the success of Fodor's theory depends on, in other words, is that when nearness of worlds is judged from a purely non-semantic and non-intentional – for our purposes, therefore, from a purely physical – point of view, the H_2O-only world always turns out to be closer than the XYZ-only world. Will this be true?

Now, Fodor thinks that the results are bound to turn out as desired because:

> . . . to get to a world where we can [infallibily] tell XYZ from H_2O, you have to either change us (provide us with instruments of observation we don't now have, for

> example) or change H_2O/XYZ (make their molecular structure visible to the naked eye, for example) or do both. But now, having gotten to such a world, if you want to also make it the case that our "water" tokens track XYZ rather than H_2O (or XYZ and H_2O) you would also have to change something about us which corresponds to the disposition which, I take it, everybody agrees underlies our use of "water" in THIS world, viz., our disposition to use "water" only of things that bear the same kind relation to our local samples.
>
> So you'd have to change more things to get to a world where XYZ is distinguishable from H_2O and our use of "water" tracks XYZ than you would to get to a world where XYZ is distinguishable from H_2O and our use of "water" tracks H_2O . . . of course, this extra thing that you'd have to change IS an intention to use "water" as a kind term; and that intention may or may not be a physical state (depending on whether physicalism is true of our intentions).[31]

To get from our world to an XYZ-only world, argues Fodor, you have to make H_2O infallibly distinguishable from XYZ *and* you have to change whatever intentional facts ground the fact that in this world "water" means H_2O; whereas, to get to an H_2O-only world, you have only to make XYZ infallibly distinguishable from H_2O. Hence, any XYZ-only world is bound to be further than any H_2O-only world.

To begin to get a sense for what's wrong with this argument we may start with the observation that the purely physical changes involved need not be on a par. From a purely physical standpoint, it may be much easier to get to a world where you will apply "water" only to XYZ, if you want to, than to get to a world where you will apply "water" only to H_2O, if you want to. This ought to be obvious. To bring it about that H_2O never gets confused for XYZ (or anything else for that matter), you only have to bring it about that H_2O has some unique distinguishing and infallibly detectable property; you don't also have to bring it about that XYZ is never confused for anything else. Vice versa for bringing it about that XYZ is never confused for H_2O. But bringing it about that some substance has some unique infallibly detectable property is a function not only of our detecting capacities but of the substance's chemistry as well. And so, it seems inevitable that it will be easier to render some substances infallibly detectable than others.

Imagine, in fact, that the chemistry of XYZ is such that, by the merest alteration of some tissue in our nostrils, we are able to *smell* its presence wherever it may be. In a world with the contemplated nasal alteration, XYZ is uniquely and infallibly detectable by the foul scent it gives off. In that world, then, there is no problem ensuring that "water" gets applied only to XYZ (if that's what you want to do): all you have to do is ensure that it gets applied only to substances that give off the distinctive scent. It doesn't follow, of course, that this foul-smelling world is equally optimal for the infallible detection of H_2O – H_2O need give off no distinctive scent of its own in this world, andy may remain confusable with substances other than XYZ. Nor does it follow that it will be *equally* easy to get to a world in which H_2O is in like manner infallibly detectable: H_2O's chemistry may not allow for that. Getting to a world in which you apply "water" only to H_2O, if you want to, may involve making many more changes in us and our surroundings than it took getting to the foul-smelling XYZ world.

Suppose in fact that that's precisely the way things are. To get, then, from our world – in which "water" means H_2O – to a world in which "water" gets applied only to H_2O, you have to make a big physical change; to get to a world in which

it gets applied only to XYZ, you have to make a small physical change and a small intentional change. Now: which world is closer to ours?

Well, if the distance measure were specified as follows, the answer would be clear: all physical changes are on a par, and every intentional change counts for as much as every physical change. But, of course, this specification would be entirely question-begging in the present context. It is not allowed to specify the similarity relation in intentional terms. And yet, on the other hand, I don't see that any other specification will yield anything like the desired results.

Fodor says:

> . . . of course, this extra thing that you'd have to change IS an intention to use "water" as a kind term; and that intention may or may not be a physical state (depending on whether physicalism is true of our intentions). But you don't . . . have to MENTION its being an intention to say what I just did say . . . viz., that you have to change more things to get to worlds where H_2O is distinguishable from XYZ AND our "water" tokens track XYZ . . . than you have to change to get to a world where H_2O is distinguishable from XYZ and our "water" tokens track just H_2O. So it doesn't look to me as though the required notion of distance is question-begging.[32]

This argument depends on the claim that getting to a world in which H_2O is infallibly detectable involves exactly as much (or as little) as getting to a world in which XYZ is. Thus, the detecting changes drop out, leaving the intentional change to settle the matter. But as I have just argued, there is absolutely no reason to believe this.

If we correct for the false assumption, it is perfectly plain, I think, that there will be no non-question-begging way of getting the H_2O-only world to come out closer than the XYZ-only world. Judged from a purely physical standpoint, the XYZ-only world may well come out closer. And no other standpoint is acceptable in the present context. Hence, Fodor's theory is either false, or it's not naturalistic.

NOTES

I want to express a special debt of gratitude to Jerry Fodor for sharing and discussing his work with me over several years. For much helpful discussion I am also grateful to Barry Loewer, Stephen Schiffer, Steve Yablo, David Velleman, Larry Sklar, Allan Gibbard, Jennifer Church and Jaegwon Kim.

1 *Psychosemantics*, 1987, MIT Press, p. 97.

2 A set of properties A *weakly* supervenes on a set B, if no two objects in a given world could differ in their A properties without differing in their B properties. On the other hand, a set of properties A *strongly* supervenes on set B, if no two objects drawn from any two worlds could differ in their A properties without differing in their B properties.

3 I think, actually, that considerations based upon the problem of mental causation may provide such an explanation; but here is not the place to go into why. For useful discussion see Stephen Yablo, "Mental causation," (forthcoming) and Brian P. McLaughlin, "Type epiphenomenalism, type dualism, and the causal priority of the physical," in *Philosophical Perspectives* 3, Ridgeview, 1989.

4 *Psychosemantics*, p. 98.

5 See Jaegwon Kim, "Concepts of supervenience," *Philosophy and Phenomenological Research*, 1984.

6 Which is not, perhaps, as big an "if" as it may at first appear. For I am inclined to believe that if there is to be a naturalized theory of meaning at all, it is likely to be a causal theory: i.e., it is likely to attempt to reconstruct the relation between a predicate and the property it expresses in terms of the *causal* relations between that predicate and that property. (What *other* sort of possible naturalistic relation between a predicate and a property looks even remotely semantically relevant?) Furthermore, any such theory is likely to be framed, not just in terms of the *actual*, but in terms of the *counterfactual*, causal relations that may obtain between a predicate and a property. In contrast with the case of the reference of proper names – where actual causal history seems paramount – it seems wrong to ignore counterfactuals in determining the meaning of predicates. But a causal theory of meaning couched in terms of truths about a symbol's counterfactual causal history is precisely what an information-theoretic semantics is all about.

7 For theories based on this basic formula see Fred Dretske, *Knowledge and the Flow of Information*, MIT Press, 1981; Dennis Stampe, "Towards a causal theory of linguistic representation," in French, Uehling, and Wettstein (eds), vol. 2, *Midwest Studies in Philosophy*, University of Minnesota Press, 1977; R. Stalnaker, *Inquiry*, MIT Press, 1984, Jerry Fodor, op. cit., and "A Theory of Content," Parts I and II, (in press-f) (henceforth *TOC*). It is with this latter piece that the present paper is most concerned.

In the interests of presenting the strongest possible version of a causal theory, I am going to suppress many qualms and follow Fodor in two important respects. First, I am going to allow that we can talk about genuine (obviously *ceteris paribus*) laws here – despite the fact that what's in question are such "laws" as that presidents cause "president" tokens and that sopranos cause "soprano" tokens. And, second, I will allow that "it's bedrock that the world contains properties and their nomic relations" (p. 41 of *TOC*), so that it counts as a perfectly objective matter what properties are causally responsible for what effects.

Notice by the way, that it can be a law that Ps cause S-events, even if there are no Ps: all that's required is that Ps would cause S-events, if there were any.

8 Causal theories of meaning face, I believe, another serious problem, that of accounting for the *normative* character of the notion of meaning. For a discussion of this point see my "The rule-following considerations," *Mind*, October 1989.

9 Strictly speaking, of course, it is instances of a property, rather than the property itself, that are said to be in extensions. To avoid prolixity, however, I shall continue to talk of properties as being in extensions; it is hardly likely to cause confusion.

It is sometimes suggested that it would be enough if possession of M were *sufficient* for being in the extension of S. But that is not right. If only sufficiency were required, we would not know S's meaning simply as a result of a definition of M. For although we would know what properties were definitely in the extension of S, we would not know if we had them all. Thus, a definition of M would not even constitute a sufficient condition for a symbol possessing a given meaning, which is the very least required from a (naturalistic) theory meaning.

10 In *TOC*, Part I.

11 ibid., p. 38 (of the typescript).

12 ibid., p. 39.

13 ibid., p. 39.

14 ibid., p. 43.

15 ibid., p. 56.

16 ibid., p. 59.

17 This is cited from Fodor's reply to a talk I gave on asymmetric dependence at the Eastern Division of the APA in Washington D.C. in December 1988. I shall refer to this reply henceforth as "D.C. Notes." The passage is from p. 9.

18 Of course, both types of situation are *physically* possible – that is, possible relative to the laws of physics, even if not possible relative to the laws of psychology – which is all that will matter to the discussion that follows.

19 cf. *TOC*, p. 65.

20 Saul Kripke, *Naming and Necessity*, p. 122.

21 ibid., pp. 135–6.

22 I stress the "as if." What we are after here is a description of the intuitive semantics of natural kind expressions, not a theory of the facts in virtue of which they come to have those semantics. What we need to know is what the intuitively correct assignments would be under a variety of counterfactual circumstances, so that we can assess theories that purport to capture those assignments; we don't also need to commit ourselves to a conception of the mechanisms by which they come to have those assignments. I emphasize this because causal theories are restricted to primitive symbols, which for our purposes means "syntactically primitive symbols that are not introduced by definitions." And so, I wish to emphasize that accepting the Kripke/Putnam account of the functioning of natural kind expressions does not imply that these expressions had to have been introduced by definitions. (As Kripke himself notes, the idea of such introductory baptisms is in any case artificial.)

23 What counts as a basic explanatory property? That's a good question to which I have no decisive answer. Information-theoretic semantics is in any event committed to hyperrealism about which properties are suitable for entering into nomic relations; it is enough for our purposes if we say that the basic explanatory properties will be some proper subset of these.

24 In the interests of keeping matters as simple as possible, I have described the case directly in terms of molecules; but it is probably more intuitive to do it first at the level of elementary particles and then work up to the molecular level.

Thus, consider "T." "T" is a name for a certain kind of elementary particle characterized by a particular array of fundamental features – charge, mass, baryon number, charm, what-have-you. It has been frequently observed and its fundamental features have been accurately recorded. It is often the cause of "T" tokens. Now, suppose that there is a physically possible particle T* that, although possessing some fundamental features distinct from T – and, hence, although constituting a different kind of particle than T, according to the basic principles of elementary particle taxonomy – nevertheless would behave just like T in all circumstances accessible to humans, or to appropriate idealizations thereof.

Now, standard intuition would have it, I submit, that T* is not in the extension of "T." "T" is a term in a basic science; it is intended to denote a natural kind: the kind defined by the non-disjunctive property exemplified by paradigm local samples. This kind, as it happens, is T. T* is not of the same kind as T. That much can be incontrovertibly clear: simply let T and T* possess different values for the basic parameters in terms of which physics taxonomizes elementary particles. So T* cannot count as within the extension of "T" and, hence, the use of "T" in application to T* is false.

If there can be kind-distinct *particles* that are indistinguishable, it seems overwhelmingly plausible that there could be kind-distinct *molecules* that are indistinguishable. Not to say that this is an automatic inference. But what with molecules being individuated partly in terms of the atoms that constitute them, and what with atoms being individuated partly in terms of the particles that constitute them, it would be very surprising, to say the least, if it were simply inconsistent with physics that there should be kind-distinct molecules between which no humanly possible experiment could distinguish.

25 I first expressed this objection in the APA paper. The modification found its way into

later drafts of *TOC*.

26 *TOC*, p. 66.

27 ibid., p. 68.

28 That's one side of the coin; the other side is that such theories would assign disjunctive extensions to expressions even when the correct result is that they have *no* extensions, because they fail to refer.

Thus, it would appear to follow from the account of natural kind expressions that we have been working with, that if there were no natural kind uniting the local paradigm samples used to fix the referent of S – if, for example, the only thing the chosen exemplars had in common was a messy motley of basic properties – that S would fail to refer. But informational theories cannot respect this result.

Again, the easiest way to see this is by emphasizing the verificationism entrained by such theories. Just imagine that the samples used to fix the extension of "water" contain a myriad distinct but indistinguishable substances. Then, all those substances will be in the extension of the term "water," even though the intuitive result is that none should be. (A myriad kind-distinct but humanly indistinguishable water-like substances may seem far-fetched; but that's no objection to its use in the present context.)

29 A version of this argument was first given in my *Essays on Meaning and Belief* (Princeton Ph.D., 1986).

30 Note that this is *psychological* holism, not the controversial doctrine of semantic holism. Fodor, for example, rejects the latter; but, as we have seen, it is very important for his purposes that he accept the former.

31 D.C. Notes, p. 10.

32 ibid.