

## Postscript to *A Mental Representation*

**1. The language of thought hypothesis.** I now take more seriously than when I wrote this paper the idea that explanations of the behavior of cats and dogs via their beliefs and desires might be a second grade kind of explanation in Quine's sense: a dramatic idiom not to be taken seriously. If so, there is no reason to take a language of thought hypothesis seriously in the case of cats and dogs. We do of course need a serious psychological model of cats and dogs, but it is not implausible that some kind of connectionist model could be given, one that doesn't serve simply to implement a language of thought model or any other literally representational model but works in a wholly different way. The usual objection to connectionist models that aren't simply means of implementing representational models is that they can't account for certain systematicities in thought (Fodor and Pylyshyn 1988); but in the case of cats and dogs, it is by no means evident that the systematicities are there. Representational models also seem almost inevitable for creatures who show clear evidence of being able to evaluate their own reasoning processes; but again, this wouldn't seem to apply to cats and dogs. In the terminology used in the paper, what I'm now saying is that I was too quick to dismiss instrumentalism about belief- and desire- attributions for such creatures.

In the case of adult humans, on the other hand, instrumentalism seems virtually inconceivable: I know perfectly well that I believe that the Earth is not flat, by introspection. My introspective confidence that this is so very likely arises from an awareness of accepting the sentence 'the Earth is not flat'. Doubtless we can introspect beliefs and desires with finer contents than we know how to express in our language, but the kinds of beliefs and desires that seem indubitably more than instrumental devices are those that don't go too far beyond the expressive power of our language. If this is right, it is all the more reason for supposing (as I tentatively did in Sections IV and VI) that an adult human's language of thought is best viewed as just an extension of his or her public language.

I didn't really mean to rule out that there is representation of a nonlinguistic kind, e.g. pictorial representation; indeed, such representation would suit the ends of the paper, for it lends itself to a naturalistic (and more or less compositional) account of *how* the internal representations represent the

world. Linguistic representation seems *prima facie* a bit harder to explain naturalistically; that was the main reason for focus on it. It would not have altered my overall view much to have held that for lower animals, mental representation might be exclusively pictorial.

(The possibility of representational systems that are neither linguistic nor pictorial is discussed in the next chapter: see the (somewhat jocular) *Light bulb model*. Such a purely Boolean model seems clearly inadequate to explain the behavior of adult humans, for reasons there given, though I suppose something like it could conceivably work as the basis for a non-instrumental theory of intentionality for some lower animals.)

**2. Brentano's Problem (1).** I first formulate Brentano's problem as the problem of making naturalistic sense of belief, desire, and so forth, given that they are relations between organisms and intentional entities like propositions. In correspondence about this Chapter in the late 1970's David Lewis suggested comparing Brentano's problem to the following problem about numbers: *Many seemingly physical properties appear to relate physical things to non-physical entities called numbers. What kind of physical relation can a seven-gram stone bear to a non-physical abstract entity: the number seven?* The same comparison has been urged by many others, e.g. Churchland 1979, Dennett 1982, Stalnaker 1984. I believe the comparison to have some, though rather limited, use.

One solution in the number case is the nominalistic one: literally speaking there are no numbers, and so literally speaking there are no relations between physical objects and numbers. (See Field 1980; Yablo unpublished.) This allows talk of numbers, and relations between stones and numbers, as useful fictions, but if they are so regarded then there is no pressure to solve the analog of Brentano's problem. Similarly, we might regard propositions as useful fictions (or maybe even not-very-useful fictions). In that case there is no pressure to solve Brentano's problem *as I stated it*, i.e. in terms of propositions. But this wouldn't dissolve questions about how mental states can stand in referential relations to the world. This points up an important disanalogy between the two problems: there is more to Brentano's problem than the use of abstract entities. I will return to this.

Lewis of course did not have the nominalist solution in mind, and there is no need to resort to it

in dealing with the numerical analog of Brentano's problem: a solution to the problem has been found by mathematical work in the theory of measurement. (See Krantz et al, 1971.) The idea is that what is explanatorily basic is not relations like *x has mass-in-kg r* between physical objects and numbers; rather, what is explanatorily basic is certain intrinsic relations holding among physical objects themselves, e.g. *the mass of x is the sum of the masses of y and z*. This idea is made precise by laying down axioms governing these intrinsic relations that make no mention of numbers, and proving a representation theorem that shows that whenever these axioms are satisfied there is a mapping of objects into the real numbers that preserves structure. Consequently, assigning real numbers to physical objects can be viewed as just a convenient way of discussing the intrinsic mass-relations that those objects have. The numerical analog of Brentano's problem is then solved: this physical object bears to the number seven the relation of being mapped into it by the unique structure-preserving mapping that accords with the convention laid down to determine the scale (i.e. the convention that determines the standard gram).

It is initially plausible to suppose that Brentano's problem should be solved *somewhat* analogously. But two points need to be made.

The less important is that taking the analogy seriously seems to involve a commitment to strong assumptions about the internal structure of mental states. The obvious analog to the postulate of a rich array of intrinsic relations among physical objects in the number case is to postulate a rich array of intrinsic relations among internal states inside the believer in the psychological case: the relations among these states must be intrinsic in the sense that they are to be describable independent of propositions. The intrinsic relations and the assumptions about them must be powerful enough to prove a representation theorem that shows that whenever the assumptions are satisfied there is a mapping of internal states into propositions that preserves the kind of structure that is important to propositions. This includes at least Boolean structure, and very likely something more sentential. So the analogy seems to require us to ascribe psychological reality to at least the Boolean structure of propositions if not the sentential, and to think that this psychologically real structure can be described independent of propositions. Such an intrinsic structure, related to the system of propositions via a structure-preserving

mapping, would seem to be enough to constitute a system of internal representation.<sup>1</sup> (It is compatible with this argument that the intrinsic structure be only the sort of Boolean structure exemplified in the Light bulb model in the next chapter. However, for reasons given in that chapter, that kind of psychological structure seems far too impoverished to describe the psychology of creatures with a language.)

But there is a much more important point (related to the one made in connection with the nominalistic view). It is that not just any mapping of internal representations to propositions that preserves Boolean or sentential structure will serve the explanatory purposes to which we put propositions: for it is easy to see that we could preserve Boolean or sentential structure while assigning the proposition that grass is green to the state associated with  $\text{>Snow is white=}$ . For propositions, unlike for numbers, structure is not enough (unless  $\text{>structure=}$  is taken in an extraordinarily stretched sense, i.e. as involving a kind of encoding of the physical world not present in the numerical case). This suggests that a significant complication of the strategy used in the numerical case will be required. The analogy of propositions to numbers turns out to be rather limited. If the only point of the analogy is to show that there is nothing mysterious about entities playing a role in explanations other than *causal* role, it succeeds (though nothing in the Chapter suggested that the acausal role of propositions was mysterious in itself). But if it is taken as substantially illuminating the way that propositions play a role in explanations, it fails: the role of propositions is importantly different.

**3. Brentano's problem (2).** Horwich 1998 suggests a more drastic way of dismissing Brentano's problem. Horwich (unlike the possibly mythical proponent of the Orthographic accident view) grants that terms like  $\text{>means=}$  and  $\text{>believes=}$  stand for genuine relations between people and propositions; but he claims that there is no reason to suppose that a physicalistic account of meaning or believing (an account of  $\text{>what constitutes=}$  meaning something or believing something) would preserve this relational status. Indeed, he introduces the label  $\text{>the Constitution Fallacy=}$  for the supposition that an account of what constitutes relational facts must itself be relational. So Horwich's idea is that one can give an account of what constitutes *believing that* by finding one monadic physical property that constitutes *believing that snow is white*, another that constitutes *believing that Pope Pius X was the*

*brother of Malcolm X*, another that constitutes *believing that Michel Foucault invented the pendulum*, and so on; these distinct monadic properties need have nothing to do with each other, and they certainly don't need to involve a common physical relation.

Horwich's claims are in service of a deflationist approach to the theory of meaning with which I am now highly sympathetic; my own version of such a theory is given in Chapters 4 and 5. But trying to achieve deflationism by the means just sketched seems to me extremely ill-advised: I see nothing fallacious about *the Constitution Fallacy*, and think that Horwich's requirements on a physicalistic account are far too weak. Consider ordinary physical relations, like *x has the same temperature as y*. Surely we wouldn't count it as an acceptable answer to the question of what constitutes sameness of temperature to say that *having the same temperature as b* is constituted by one monadic property in the case of one object  $b_1$ , a different monadic property for a different object  $b_2$  (of different temperature), a third monadic property for  $b_3$ , etc., where these monadic properties have nothing in common. If the belief relation between people and propositions doesn't require a physicalistic account that meets the same standards as we would impose in the case of other relations, one must say why.

It wouldn't help Horwich to point out that the standards of reduction (or constitution) appropriate to relations between purely physical items can't be supposed to apply when one of the items involved is abstract. For in the first place, the measurement-theoretic solution of the numerical analog of Brentano's problem provides a genuine account of relations between objects and numbers; obviously it differs in important ways from a typical reduction of a relation between purely physical items, but nonetheless it *does* preserve the relational character. (As we've seen, this solution does not carry over to intentional relations. But unless we can give grounds for adopting lower standards on a physicalistic account in the intentional case than the numerical, that doesn't show that there is no need of a genuinely relational account in the intentional case, it only shows that the job is harder there.) In the second place, some of the mental relations for which Horwich wants to avoid giving a relational account hold between physical entities: e.g. *x has a belief about person p*. To repeat, I am inclined to think that Horwich is right that the present Chapter treats such mental relations as *too much like* ordinary physical relations, but we need an alternative model of what they are like (one that I try to supply in

Chapters 4 and 5). One shouldn't simply label as fallacious the problem of trying to provide such a model.

**4. Non-intentional explanations.** The two main topics of the paper are

- (i) What explanatory purpose (if any) is served by ascribing truth conditions to our inner representations?
- (ii) How (if at all) can the relation *r has the truth conditions that p* (where r ranges over inner representations) be naturalistically explained?

On (ii) (which is essentially Brentano's problem), I argue that by taking inner representations as having sentential structure, we can adapt the theory of truth in Chapter 1 to inner representations; also, I make explicit that such an approach naturalistically explains not just truth but truth-*conditions* (that is, the truth condition relation mentioned in (ii)). As succeeding chapters will make clear, I have become quite doubtful of the need for such an approach.

As for (i), a good bit of my discussion was directed toward showing that it is not entirely easy to find an explanatory role for truth conditions: for once one has the system of internal representations, one can do much of one's explaining simply in terms of it, with no reference to the truth-conditions that are assigned to the representations. I made a distinction between what I would now call *intentional explanations*, that make appeal to the truth conditions of states (or to related semantic features of the components of states, such as their referents), and *non-intentional* explanations that do not involve any such thing but simply rely on the causal laws governing those states. (By *laws* I don't mean anything very heavy duty: just macroscopic regularities that hold, within limits, for the system.)

Unfortunately, I used the term *narrow* for *non-intentional* and *broad* for *intentional*; this was unfortunate because shortly afterwards, Fodor 1980 popularized the use of similar terms for a very different distinction from mine, and many people failed to realize that the two distinctions were not at all the same. Fodor was concerned with the distinction between *wide explanations* that make appeal to things outside the agent and *narrow explanations* that do not; mine (to repeat) was between

*intentional explanations* that make appeal to the truth conditions of states (and to the referents of components of states) and *non-intentional* explanations that do not. The narrow-wide distinction and the intentional-nonintentional distinctions cut across each other. An explanation that appeals to external causes of our internal representations will be wide, even if the causes it appeals to have nothing to do with the truth conditions of the representations. Conversely, an explanation in which appeal is made to an internal symbol obeying the truth-table for disjunction is an intentional explanation even if the assignment of this truth-table to the symbol doesn't depend on factors external to the agent. In my view, the intentional v. non-intentional distinction is far more important than the distinction of wide v. narrow.

I don't think the paper makes clear enough how broad the scope of non-intentional explanations is. In the first place, I should have made clear that we can theorize about an agent at many different levels of detail and idealization, and that appeal to internal representations (described without mention of what they represent) can appear in theories at many different levels. At a highly idealized level, there is Bayesian decision theory construed as a psychological theory, using degrees of belief\* and desire\* (i.e. Attitudes toward@sentences instead of toward propositions); or rather, Bayesian decision theory supplemented with laws governing how sensory stimulations affect observation sentences and how decisions lead to motor outputs. We can introduce more and more complications, to get more realistic. If we give straightforward causal laws of how beliefs and desires and other representational states evolve over time (influenced by sensory stimulations and the like), their representational properties (truth conditions, reference etc.) will play no role in the laws; similarly, such properties will play no role in any laws of how such internal states lead to motor outputs. I'm not restricting my claim here to deterministic theories: it seems to me that except in the crudest of idealizations the psychological laws will mostly be indeterministic, but it is still hard to see how representational properties like reference or truth conditions could play a role in the laws.

I have described the psychology Anarrowly@, i.e. as taking as inputs sensory stimulations rather than external conditions, and analogously at the output end. But nothing hangs on this: using Awide@ inputs and outputs, i.e. inputs and outputs described in external terms, is not enough to make the laws intentional. This should be obvious: no talk of reference or truth conditions is involved. Indeed, an input

law that says that coming to believe  $\text{>There is a rabbit in front of me=}$  is typically caused by the presence of rabbits can be factored into two components, one saying that coming to believe the sentence is typically caused by sensory stimulations in a certain class and the other saying that sensory stimulations in that class are typically caused by rabbits. (Factor may not be quite the right word: it suggests that the two resulting laws taken together are fully equivalent to the original law, but actually they introduce a slight improvement of it.) A wide input law is in effect just an existential quantification over conjunctions of narrow input laws and laws about what causes sensory stimulations. (Existential quantification is needed simply because we may not know exactly which class of sensory stimulations it is that is caused by the presence of rabbits and causes the belief in  $\text{>There is a rabbit in front of me=}$ .)<sup>2</sup>

The kind of non-intentional theories of an agent that I've been considering might be called computational theories, in a broad sense that allows features of the external world to appear in the computational theory if they regularly cause sensory stimulations or are regularly affected by motor outputs. These are the kind of theories of an agent that are used in functionalist models. One of the things I was arguing in the paper was that it is hard to see how we could ever give a functionalist account of belief and desire in the usual sense, i.e. belief and desire *as relations to propositions*: discussions of how to do this typically treat each belief and desire as a separate state, without getting the interconnections among (the potential infinity of) such states right. What we can hope to give a functionalist account of is of the relations belief\* and desire\*, which are relations *to inner representations*. The account of the relation of the inner representations to the propositions (or of the singular term components of inner representations to objects and the general term components of inner representation to sets or properties) has to be given separately from the functional story. This is part of what is argued in Section II of the Chapter. (There is a broad sense in which the relation between inner representations and propositions might be functional: it might be a second order relation, the relation of *standing in some first order relation or other that meets condition Y*. Indeed, in the paper I use the term *functional* in this broad sense, a terminological choice I now regret. But my point now is that the condition  $\Psi$  won't be the kind of broadly computational conditional that we are familiar with from standard functional theories. The question of what other kind of condition it

might be is one way to put the question of finding an explanatory role for propositional content or truth conditions.)

A small point: in explaining how there *can* be a functional theory of belief\* and desire\*, I offer a functionalist account of what it is to be an inner representation, and of the syntactic features of inner representations. This is still my preferred way of thinking of inner representations, but there is an alternative: we could think of the inner representations as abstract syntactic objects, and use them more or less as we might use numbers in formulating the computational theory. For abstract syntactic objects, the analogy to numbers suggested by Lewis, Churchland and others (mentioned in Section 2 of this postscript) would be quite a good one: by removing the representational aspect of the abstract objects, we have removed what makes the analogy problematic.<sup>3</sup>

I have said that computational theories (even in the broad sense that allow reference to external causes and effects of internal representations) don't employ representational properties like reference and truth conditions. I haven't said that they don't employ *semantic* properties, for in a broad sense of >semantic= there may be semantic properties for which something having the property is *constituted* by its having a certain computational role. For instance, the paper argues that the right moral to draw from Frege's Hesperus-Phosphorus problem is that yes, these coreferential names differ in some semantic property loosely construed, but (contrary to Frege) the difference is only that the names have different computational roles for most of us: one is more directly connected to beliefs involving the term >morning= and the other to beliefs involving the term >evening=. If this computational difference is counted as semantic, then obviously semantic features of representations are not left out of the computational story. But it is the other aspects of semantics, going beyond computational role even in the broad sense, that are left out.

And to reiterate, the kind of computational role that the functionalist story allows for includes what has been called *Along-armed conceptual role* (Harman 1982), which includes causal connections to the environment. It may be part of the conceptual role of >That is a rabbit= to be typically caused by nearby rabbits. Equally, it may have been part of the conceptual role of >Phlogiston is disappearing from

that flask= to have been typically caused by oxygen entering the flask. This latter example makes clear, I hope, that to bring in long armed conceptual roles is not to bring in representational properties.

(Admittedly, there is a heavy overlap between the causal connections to the environment that would be used in wide input laws and those that an advocate of a causal theory of representation would want to build into the truth condition relation for observation sentences. The explanation is obvious: most of those who believe in a substantive theory of truth conditions want it to be a consequence of their theory that observational beliefs tend to be true, and to guarantee this they suppose that by and large the factors involved in the causal input law are involved in constituting truth conditions. But this obviously does nothing to show that causal input laws are themselves intentional; and cases where it is natural to say that our observation reports are laden with a false theory makes this especially evident.)

The fact that representational properties are left out of the causal story should be even clearer in the case of other kinds of words. Consider proper names of long-dead people like Aristotle: the factors usually thought relevant to reference include causal chains extending through the past back to Aristotle, but it is a stretch to say that these are involved in the conceptual or functional role of the word >Aristotle=. (Causal connections to certain *portraits* might be part of the functional role of >Aristotle=, but this is so even if the portraits turn out to have someone else as their causal source.) And consider, once again, logical connectives like >not=. A representational semantics will take this as standing for the truth function that maps truth to falsehood and vice versa. Or if you don't like talk of >standing for truth functions=, the point can be put differently: the representational semantics will say that when attached to a true sentence, >not= yields a false sentence, and when attached to a false yields a true. But this isn't a fact about the conceptual or functional role of >not=. The conceptual role of >not= is relatively easy to specify: it is given in large part by the rules of deductive inference that we take to govern the term. But specifying the conceptual role in this way doesn't seem to say anything about truth functions.

Perhaps the last few paragraphs need a qualification. When I said that computational theories (even in the broad sense) don't employ representational properties like reference and truth conditions, maybe all I was entitled to say is that they don't employ representational *concepts*? Maybe conceptual role concepts and representational concepts are distinct concepts for the same property? I don't think

this is a plausible view even in cases like the logical connective case where it is plausible that there is some kind of supervenience of the representational property on the conceptual role property. (For one thing, the representational property seems to have a relational element that isn't part of the conceptual role property.) And in other cases, like the case of names of long-dead people, there doesn't even seem to be such supervenience; in which case the properties certainly can't be identical. But in any case, it wouldn't much matter to my point if we did qualify the previous paragraphs in the way suggested. For the important point is that a great deal of psychological theorizing can be done without representational *concepts*. And the question is, what additional advantages do we get by employing representational concepts?

**5. The explanatory role of intentional properties (1).** How then do representational concepts enter into psychological explanations? I will be especially concerned with explanations of *phenomena described without representational concepts*. Obviously we frequently use representational concepts (e.g. beliefs and desires) in explaining other beliefs and desires, or in explaining behavior described in an intentional way (>murdered his wife=) that presupposes representational concepts; but it seems to me that an illuminating account of the explanatory role of representational concepts should focus first and foremost on their use in explaining facts *not* so described. (>Why did her arm move in that way?=>Why did she end up in San Francisco?=>) Once we have motivated the use of representational concepts, it is no surprise that we want to explain facts described in terms of them in addition to facts described without them; but an illuminating account of the explanatory role of the practice requires that we show how it would be explanatorily useful for someone who didn't already have the practice to introduce it.

A natural view is that when we use representational concepts or properties to explain facts described in non-representational terms, the representational concepts or properties just code for conceptual or functional role properties: we specify the functional role property by specifying the representational property. This is certainly very plausible, if we don't take it as committed to there being any very uniform account of what conceptual role properties a representational property codes for, and if we don't suppose that even on a given occasion there is a very *precise* conceptual role property that

is coded for. But (as it stands anyway) it leaves unanswered the question of why it is *useful* to code conceptual role properties by representational properties (or by representational concepts).

Additionally, even if it is granted that *one* function of representational properties or concepts is to code for conceptual role properties, the question arises whether there might be other functions in addition.

One obvious fact that needs emphasis is that typically when one offers an explanation in terms of beliefs and desires, one is not in a position to offer a complete explanation in computational terms. In itself this doesn't show much, for the same holds in explaining anything: if we explain the center's goal by citing the wing's perfectly placed pass, we certainly aren't in a position to cite precise laws of physics (or regularities of hockey, or of the system consisting of this particular center and this particular goalie) which would determine probabilities for the goal under the conditions present with and without the perfectly placed pass. What we call "explanations" are only explanation sketches (indeed, only sketches of *partial* explanations): sometimes very incomplete sketches that consist of nothing more than mentioning a salient part of the overall cause. No reason so far for supposing the need of any theory in psychology beyond the computational.

Still, this point doesn't seem to be enough to account for the fact that in the intentional case, the ideology that we use in our explanation sketches is so far removed from the ideology of the computational theories that provide the only kinds of laws we have so far found hope of finding. If there aren't even very rough laws at the intentional level (and from which the intentionality isn't easily eliminable), why is it that the policy of explaining in intentional terms is so useful?<sup>4</sup>

One answer that I briefly consider (and deem insufficient) in Section V is that intentional explanations of another person's behavior are "projective" in that they involve reference to the explainer's language even though that is not causally relevant to the behavior. When explaining a person's behavior (say the raising of his gun) in terms of his belief that there is a rabbit nearby, what I am in effect doing is explaining the behavior in terms of his believing\* a representation that plays a role in his psychology rather similar to the role that "There are rabbits nearby" (or the mental representation associated with it) plays in mine. (This has some similarity to the "second grade" explanations

mentioned in section 1 of this postscript, but unlike those, this takes the computational belief-desire psychology seriously.) Such an explanation is still basically non-intentional: truth conditions play no real explanatory role. Of course, there is a sense in which my *sentence* >There are rabbits= plays an explanatory role here: obviously not as a causal factor in the explanation, but as a device we use in picking out the agent=s internal representation (which *is* a causal factor). Moreover, my sentence, to which his sentence is compared, has truth conditions, and I know what these truth conditions are. But it is my *sentence*, not its truth conditions, that is playing the direct role in picking out the agent=s internal representation. Moreover, even my sentence is playing a very indirect role in explaining the agent=s behavior: the agent=s internal representation is what is centrally involved in that. So truth conditions are playing a *doubly* indirect role in the explanation. (And note that it is the assignment of truth conditions to *my* sentence that is playing this doubly indirect role. If, as I assumed in the paper, the assignment of truth conditions to the agent=s representations is assumed autonomous, i.e. not constituted by their translatability into a sentence of mine with those truth conditions, then the assignment of truth conditions to the agent=s states doesn't even play an indirect role in the explanation.)

We still have no example where truth conditions play a serious explanatory role. And at this point we might well wonder whether there *are* such examples. If not, the motivation for a serious theory of the truth conditions of internal representations may be called into question: the assignment of truth conditions to the representations seems explanatorily idle.

**6. The explanatory role of intentional properties (2).** At this point in Section V, I introduced an idea from Chapter 1: that we assign truth conditions to the states of others as a means to use their states to find out about the world. Given that they accept >It=s raining here= or a sentence that would normally be translated that way, we have reason to think that it is raining there. (I should have added the converse: given that it is raining there, and that they are outdoors etc., we have reason to think that they accept >It is raining here= or a corresponding sentence in their language.) I said that in making such inferences we implicitly employ a *reliability theory* of agents, and that this should be viewed as an extension of our non-intentional psychology of the agents.<sup>5</sup>

Stephen Stich (1983, pp. 200-204) argues against the idea of such a reliability theory. He first notes that we would need different reliability theories for different people, since we use idiosyncracies of different people in predicting and explaining their behavior. (I have to admit that the end of Section V suggested the contrary; this was a silly mistake.) He then says it is implausible that we could carry around so many different reliability theories in our head. But of course there is an obvious story to tell here, namely that we carry around a lot of general assumptions about what things people tend to be reliable about, and that for people we know we carry around information about how the general assumptions are to be expanded and/or modified in their case. (Not just for individual people actually: we also have general assumptions about certain categories of people, like doctors and preachers and politicians.) Of course, the general views about people, as well as the assumptions about specific individuals and specific categories, can evolve over time. Looked at this way, the idea that we have reliability theories that vary from person to person doesn't seem as preposterous as Stich tries to make it sound. And since our computational theories would presumably vary from person to person also—even very idealized computational theories such as Bayesian theories (where various caution parameters and other aspects of inductive procedures would surely vary from one person to the next)—it's hard to see why Stich thinks that there is a special problem about reliability theories.

Stich's objection isn't simply to the question of whether our reliability assumptions are sufficiently systematic to deserve to be called a theory; he disagrees with the very idea that we use the notion of truth in learning from others. But as he admits, he has no serious alternative account of how we learn from others on offer, and I can't imagine how there could be an alternative story that doesn't involve a notion of truth. But Stich's larger point was to defend his "syntactic theory of the mind", and it now seems to me that the defense of this does not require his radical line that truth plays no role in learning from others. For it is hard to see why, in an account of how we learn from others, anything more than a purely disquotational truth predicate is needed. (This will be discussed in Section 11 of Chapter 4.) Finding an *explanatory* role for the assignment of truth conditions to mental states might give reason for believing in the sort of heavy duty account of truth conditions that the paper suggests, but saying that we need to mention truth or truth conditions in an account of how we learn from others

isn't to give truth conditions an explanatory role.

But at this point there is a new wrinkle to consider: Stephen Schiffer (1981) argues that the kind of reliability assumptions that we use in learning from others also play a role in explanations. If we were in a position to give a full computational explanation of an agent's behavior, we wouldn't need to use truth-conditions, but of course we are never really in a position to do that; and (he claims) in explanations given in ignorance of such details, *truth conditions serve a more central role than they have according to the projectivist story of the previous Section.*

To see what Schiffer had in mind, consider an example (a variation on his). Smith is in a fairly large nearby room, into which I can't see; Oscar runs by with gun in hand, yelling "I'll kill that bastard Smith!", seeming deadly serious, and he goes into the room that Smith is in; I hear a shot. Now, I don't know where in the room Smith is: I just have a broad probability distribution. I also have no idea where in the room Oscar is aiming his gun: I just have a broad probability distribution for that too. If these two probability distributions were independent, I wouldn't attribute a very high probability to Smith's having been shot. But of course they are not independent: I think that Oscar is likely to have pointed the gun in Smith's direction. And it seems that the justification for the correlation in the probabilities is that I think that Oscar is likely to have had the *correct* belief about where in the room Smith is.

Correctness **B** truth **B** is playing an explanatory role.

Of course there are other cases where we will introduce correlations between belief states and external states of affairs not based on truth. (I read Patrick Buchanan's latest rantings, and know that the very same sentences will appear in my uncle's belief box. I see a 6'4" person in front of my mother, and know she will believe the sentence "He is at least 6'7". I see the lightning in front of a Zeus-worshipper, and know he will believe "Zeus is throwing thunderbolts".) Still, we might have a hard time describing the correlations in the Oscar case without using a notion of truth.

Actually this is not entirely clear: after all, we have the causal input laws. Doesn't it follow from the causal input laws that (with high probability) Oscar will believe a sentence of the form 'Smith is at angle  $\alpha$  from me' iff Smith is at angle  $\alpha$  from him? If so, there is no need of the notion of truth here.

This is of course a reliability law of sorts, but that is because all causal input laws are in effect reliability laws of sorts; the point is that *this sort of* reliability law doesn't go beyond the broad computational theory.

Still, once one appreciates the idea of using reliability assumptions in explanation, there is no obvious reason to confine them to the causal input laws. By our earlier discussion, we have rather wide-ranging beliefs about the reliability of different people, and different sorts of people, on particular issues: we expect, for instance, that the adults around us will have true beliefs about who is President and what city they are in and whether they are homeless and what they need to do if they want a meal in a restaurant. (That is: we expect that they will believe\* a sentence that is appropriate to translate as a true sentence of the form  $\exists x$  is President $\neg$ , etc.) There is no reason that these reliability assumptions can't be used in explanations, and these do go beyond computational psychology even in the broad sense explained earlier. This is a serious use of truth in explanations, going beyond computational psychology. (I will have more to say about this in particular, about the relation between using truth in this way and projective explanations near the end of the Postscript to Chapter 4.)

But this use of the notion of truth in explanations is not enough to motivate the assumptions about the need for a substantive theory of truth that I made in this Chapter. (In this I agree with Schiffer: see the final section of his paper.) As I see it, the crucial point is that in the reliability laws and reliability assumptions, truth is serving simply as a device of generalization. For each given angle  $\alpha_0$ , we can explain, from Smith's being at angle  $\alpha_0$  from Oscar and Oscar's believing  $\exists$ Smith is at angle  $\alpha_0$  from me $\neg$ , how the shooting took place. The role of truth is simply to generalize on these explanations. (Similarly for explanations involving reliability assumptions that go beyond the causal input laws.) And as noted in the Postscript to the previous chapter, the use of truth as a device of generalization is not enough to motivate the kind of heavy duty truth theory that this Chapter and the previous one assume.

Does what I've just said depend essentially on supposing that Oscar speaks my language? At one point I seem to have thought so, but now I don't see that it does. If Oscar speaks a different language, the situation is that for each given angle  $\alpha_0$ , we know how to explain, from Smith's being at

angle  $\alpha_0$  from Oscar and Oscar's believing a sentence *that we translate as*  $\lambda$ Smith is at angle  $\alpha_0$  from me, how the shooting took place. Truth *under this translation scheme* can then serve as a device of generalization in this case. We don't even need to assume that the translation scheme has any privileged role in relating his language to ours: we might employ different translation schemes in different contexts, for different kinds of generalization that proved useful to us. If there is an argument for a heavy duty notion of truth or truth conditions, we have yet to find it.

## Notes to Postscript

---

1. I put this tentatively, in part because of the vagueness of ~~system of internal representation~~ and in part because one might demand only a slightly weaker representation theorem. Dennett 1982 (note 2) observes in effect that there is an alternate version of the representation theorem, where numbers are assigned not to physical objects directly but to monadic mass properties, properties of having a particular mass; and that the analog of this in the intentional case invokes not *objects* or *inner occurrences* with a Boolean or sentential structure but *properties* with such a structure. Does this give a way to commit ourselves to less inner structure? If we follow the numerical analogy closely, by supposing that the properties in question are ones that a given object can have at most one of at a given time, then the answer is clearly ~~no~~: we would need a system of inner objects or inner occurrences to instantiate the properties. But maybe what Dennett had in mind is that the properties are ones that apply to organisms as a whole, and which the organism can have a large number of simultaneously. In that case, we need only intrinsic structural relations of the sort appropriate to propositions *among the simultaneous states of the organism as a whole*; and this may not be what most people have in mind when they speak of a ~~system of inner representation~~.

2. In this paragraph I've stated the sample input law very vaguely and crudely, but I believe that the basic point could be made with more precise and realistic laws.

3. Whichever route one prefers, a point made in the Chapter (note 32) bears repeating: there is no evident need in the functional theory for the idea of two ~~internal word-tokens~~ or ~~internal sentence-tokens~~ *employed by different thinkers* being instances of the same word or the same sentence; an *intrapersonal* notion of type-identity is sufficient. I believe that Stich 1983 gets his ~~syntactic theory of the mind~~ into needless trouble by failing to realize this: for instance, in his Chapter 4. (If anyone

---

thinks that the use of abstract syntactic objects requires interpersonal identifications, see the discussion of local abstract objects in Chapter 5, near the end of Section 6.) Of course, disallowing talk of interpersonal sameness of internal words doesn't rule out semantic comparisons: one might say that a word in you and a word in me both refer to Aristotle without regarding as meaningful the question of whether they are the same word. Also, when speakers (by behavioral criteria) share a public language in which a certain word is used with no intrapersonal ambiguity, this sets up a natural way to correlate those of their internal symbols that are causally associated with the word.

4. Stephen Schiffer (in conversation) has raised a second question about the explanation sketch view: if all we are doing in ordinary explanation is mentioning a cause of the behavior, why insist on causes of the form 'He believes (or desires etc.) some sentence that roughly translates as 'p''? Why not other causes or cause-description, eg 'such & such a neuron was firing'? Maybe you could rule out that particular alternative cause-description, perhaps a bit *ad hocly*, by demanding that the causes described be ones that serve as the agent's reasons as well. But even so, that wouldn't seem to handle 'He is in some belief state that he'd express by S' (S a sentence of his language that we don't understand); that may give his reason (as he would express it), but we wouldn't find it terribly helpful to us.

I think that the answer to Schiffer's question involves a connection between explanation and prediction. One of the main reasons that explaining the behavior of individuals is important to us is that it aids us in predicting their behavior on *other* occasions. (Being a good explainer of one person also improves our ability to predict other people.) And *prediction goes better if we can relate the state of the agent in question to one of our sentences than to one of his (or to the firing of one of his neurons)*. The reason for that is a combination of two factors. Probably the more important is that there is something right in the simulationist account of prediction (Gordon 1986), according to which we typically predict the state of others by imagining that we accept and desire what

---

they do and making an off-line decision@ourselves: we need to have beliefs and desires *described in our language* in order to do this. The other is the role of Reliability assumptions@in explanation: this component (first clearly brought to my attention by Schiffer 1981) is discussed below.

5. I think what I had in mind was not simply a theory about when the agents are reliable but also how their reliability contributes to their success; but the inclusion of the latter was certainly not explicit.