32 Autonomous Psychology and the Belief-Desire Thesis

Stephen P. Stich

A venerable view, still very much alive, holds that human action is to be explained at least in part in terms of beliefs and desires. Those who advocate the view expect that the psychological theory which explains human behavior will invoke the concepts of belief and desire in a substantive way. I will call this expectation *the belief-desire thesis*. Though there would surely be a quibble or a caveat here and there, the thesis would be endorsed by an exceptionally heterogeneous collection of psychologists and philosophers ranging from Freud and Hume, to Thomas Szasz and Richard Brandt. Indeed, a number of philosophers have contended that the thesis, or something like it, is embedded in our ordinary, workaday concept of action.[1] If they are right, and I think they are, then in so far as we use the concept of action we are *all* committed to the belief-desire thesis. My purpose in this paper is to explore the tension between the belief-desire thesis and a widely held assumption about the nature of explanatory psychological theories, an assumption that serves as a fundamental regulative principle for much of contemporary psychological theorizing. This assumption, which for want of a better term I will call the *principle of psychological autonomy*, will be the focus of the first of the sections below. In the second section I will elaborate a bit on how the belief-desire thesis is to be interpreted, and try to extract from it a principle that will serve as a premise in the argument to follow. In the third section I will set out an argument to the effect that large numbers of belief-desire explanations of action, indeed perhaps the bulk of such explanations, are incompatible with the principle of autonomy. Finally, in the last section, I will fend off a possible objection to my argument. In the process, I will try to make clear just why the argument works and what price we should have to pay if we were resolved to avoid its consequences.

From S. Stich, Autonomous psychology and the belief-desire thesis, *Monist* 61, 573–591 (1978).

The Principle of Psychological Autonomy

Perhaps the most vivid way of explaining the principle I have in mind is by invoking a type of science fiction example that has cropped up with some frequency in recent philosophical literature. Imagine that technology were available which would enable us to duplicate people. That is, we can built living human beings who are atom for atom and molecule for molecule replicas of some given human being (cf. Putnam 1973, 1975). Now suppose that we have before us a human being (or, for that matter, any sort of animal) and his exact replica. What the principle of autonomy claims is that these two humans will be psychologically identical, that any psychological property instantiated by one of these subjects will also be instantiated by the other.

Actually, a bit of hedging is needed to mark the boundaries of this claim to psychological identity. First, let me note that the organisms claimed to be psychologically identical include any pair of organisms, existing at the same time or at different times, who happen to be atom for atom replicas of each other. Moreover, it is inessential that one organism should have been built to be a replica of the other. Even if the replication is entirely accidental, the two organisms will still be psychologically identical.

A caveat of another sort is needed to clarify just what I mean by calling two organisms "psychologically identical." For consider the following objection: "The original organism and his replica do not share *all* of their psychological properties. The original may, for example, remember seeing the Watergate hearings on television, but the replica remembers no such thing. He may think he remembers it, or have an identical "memory trace", but if he was not created until long after the Watergate hearings, then he did not see the hearings on television, and thus he could not remember seeing them." The point being urged by my imagined critic is a reasonable one. There are many sorts of properties plausibly labeled "psychological" that might be instantiated by a person and not by his replica. Remembering that p is one example, knowing that p and seeing that p are others. These properties have a sort of "hybrid" character. They seem to be analyzable into a "purely psychological" property (like seeming to remember that p, or believing that p) along with one or more non-psychological properties and relations (like p being true, or the memory trace being caused in a certain way by the fact that p). But to insist that "hybrid" psychological properties are not psychological properties at all would be at best a rather high handed attempt at stipulative definition. Still, there is something a bit odd about these hybrid psychological properties, a fact which reflects itself in the intuitive distinction between "hybrids" and their underlying "purely psychological" components. What is odd about the hybrids, I think, is that we do not expect them to play any role in an explanatory psychological theory. Rather, we expect a psychological

theory which aims at explaining behavior to invoke only the "purely psychological" properties which are shared by a subject and its replicas. Thus, for example, we are inclined to insist it is Jones's *belief* that there is no greatest prime number that plays a role in the explanation of his answering the exam question. He may, in fact, have *known* that there is no greatest prime number. But even if he did not know it, if, for example, the source of his information had himself only been guessing, Jones's behavior would have been unaffected. What knowledge adds to belief is psychologically irrelevant. Similarly the difference between really remembering that p and merely seeming to remember that p makes no difference to the subject's behavior. In claiming that physical replicas are psychologically identical, the principle of psychological autonomy is to be understood as restricting itself to the properties that can play a role in explanatory psychological theory. Indeed, the principle is best viewed as a claim about what sorts of properties and relations may play a role in explanatory psychological theory. If the principle is to be observed, then the only properties and relations that may legitimately play a role in explanatory psychological theories are the properties and relations that a subject and its replica will share.

There is another way to explain the principle of psychological autonomy that does not appeal to the fanciful ideas of a replica. . . . Jaegwon Kim (1978) has explicated and explored the notion of one class of properties *supervening* upon another class of properties. Suppose S and W are two classes of properties, and that S# and W# are the sets of all properties constructible from the properties in S and W respectively. Then, following Kim, we will say that the family S of properties supervenes on the family W of properties (with respect to a domain D of objects) just in case, necessarily, any two objects in D which share all properties in W# will also share all properties in S#. A bit less formally, one class of properties supervenes on another if the presence or absence of properties in the former class is completely determined by the presence or absence of properties in the latter.[2] Now the principle of psychological autonomy states that the properties and relations to be invoked in an explanatory psychological theory must be supervenient upon the *current, internal physical* properties and relations of organisms (i.e., just those properties that an organism shares with all of its replicas).

Perhaps the best way to focus more sharply on what the autonomy principle states is to look at what it rules out. First, of course, if explanatory psychological properties and relations must supervene on *physical* properties, then at least some forms of dualism are false. The dualist who claims that there are psychological (or mental) properties which are not nomologically correlated with physical properties, but which nonetheless must be invoked in an explanation of the organism's behavior, is denying that explanatory psychological states supervene upon physical states, However, the autonomy principle is not inimical to all

forms of dualism. Those dualists, for example, who hold that mental and physical properties are nomologically correlated need have no quarrel with the doctrine of autonomy. However, the principle of autonomy is significantly stronger than the mere insistence that psychological states supervene on physical states.[3] For autonomy requires in addition that certain physical properties and relations are psychologically irrelevant in the sense that organisms which differ *only* with respect to those properties and relations are psychologically identical.[4] In specifying that only "current" physical properties are psychologically relevant, the autonomy principle decrees irrelevant all those properties that deal with the history of the organism, both past and future. It is entirely possible, for example, for two organisms to have quite different physical histories and yet, at a specific pair of moments, to be replicas of one another. But this sort of difference, according to the autonomy principle, can make no difference from the point of view of explanatory psychology. Thus remembering that p (as contrasted with having a memory trace that p) cannot be an explanatory psychological state. For the difference between a person who remembers that p and a person who only seems to remember that p is not dependent on their current physical state, but only on the history of these states. Similarly, in specifying that only *internal* properties and relations are relevant to explanatory psychological properties, the autonomy principle decrees that relations between an organism and its external environment are irrelevant to its current (explanatory) psychological state. The restriction also entails that properties and relations of external objects cannot be relevant to the organism's current (explanatory) psychological state. Thus neither my seeing that Jones is falling nor my knowing that Ouagadougou is the capital of Upper Volta can play a role in an explanatory psychological theory, since the former depends in part on my relation to Jones, and the latter depends in part on the relation between Ouagadougou and Upper Volta.

Before we leave our discussion of the principle of psychological autonomy, let us reflect briefly on the status of the principle. On Kim's view, the belief that one set of properties supervenes on another "is largely, and often, a combination of metaphysical convictions and methodological considerations" (Kim 1978). The description seems particularly apt for the principle of psychological autonomy. The autonomy principle serves a sort of regulative role in modern psychology, directing us to restrict the concepts we invoke in our explanatory theories in a very special way. When we act in accordance with the regulative stipulation of the principle we are giving witness to the tacit conviction that the best explanation of behavior will include a theory invoking properties supervenient upon the organism's current, internal physical state.[5] As Kim urges, this conviction is supported in part by the past success of theories which cleave to the principle's restrictions, and in part by some very fundamental metaphysical convictions. I think there

is much to be learned in trying to pick apart the various metaphysical views that support the autonomy principle, for some of them have implications in areas quite removed from psychology. But that is a project for a different paper.

The Belief-Desire Thesis

The belief-desire thesis maintains that human action is to be explained, at least in part, in terms of beliefs and desires. To sharpen the thesis we need to say more about the intended sense of *explain*, and more about what it would be to explain actions *in terms of beliefs and desires*. But before trying to pin down either of these notions, it will be useful to set out an example of the sort of informal belief-desire explanations that we commonly offer for our own actions and the actions of others.

Jones is watching television; from time to time he looks nervously at a lottery ticket grasped firmly in his hand. Suddenly he jumps up and rushes toward the phone. Why? It was because the TV announcer has just announced the winning lottery number, and it is the number on Jones's ticket. Jones believes that he has won the lottery. He also believes that to collect his winnings he must contact the lottery commission promptly. And, needless to say, he very much wants to collect his winnings.

Many theorists acknowledge that explanations like the one offered of Jones rushing toward the phone are often true (albeit incomplete) explanations of action. But this concession alone does not commit the theorist to the belief-desire thesis as I will interpret it here. There is considerable controversy over how we are to understand the 'because' in "Jones rushed for the phone because he believed he had won the lottery and he wanted . . ." Some writers are inclined to read the 'because' literally, as claiming that Jones's belief and his desire were the *causes* (or among the causes) of his action. Others offer a variety of noncausal accounts of the relation between beliefs and desires on the one hand and actions on the other.[6] However, it is the former, "literal," reading that is required by the belief-desire thesis as I am construing it.

To say that Jones's belief that he had won the lottery was among the causes of his rushing toward the phone is to say of one specific event that it had among its causes one specific state. There is much debate over how such "singular causal statements" are to be analyzed. Some philosophers hold that for a state or event S to be among the causes of an event E, there must be a law which somehow relates S and E. Other philosophers propose other accounts. Even among those who agree that singular causal statements must be subsumed by a law, there is debate over how this notion of subsumption is to be understood. At the heart of this controversy is the issue of how much difference there can be between the properties invoked in the law and those invoked in

the description of the event if the event is to be an instance of the law.[7] Given our current purposes, there is no need to take a stand on this quite general metaphysical issue. But we will have to take a stand on a special case of the relation between beliefs, desires, and the psychological laws that subsume them. The belief-desire thesis, as I am viewing it, takes seriously the idea of developing a psychological theory couched in terms of beliefs and desires. Thus, in addition to holding that Jones's action was caused by his belief that he had won the lottery and his desire to collect his winnings, it also holds that this singular causal statement is true in virtue of being subsumed by laws which specify nomological relations among beliefs, desires and action.[8]

There is one further point that needs to be made about my construal of the belief-desire thesis. If the thesis is right, then action is to be explained at least in part by appeal to laws detailing how beliefs, desires and other psychological states effect action. But how are we to recognize such laws? It is, after all, plainly not enough for a theory simply to invoke the terms 'belief' and 'desire' in its laws. If it were, then it would be possible to convert any theory into a belief-desire theory by the simple expedient of replacing a pair of its theoretical terms with the terms 'belief' and 'desire'. The point I am laboring is that the belief-desire thesis must be construed as the claim that psychological theory will be couched in terms of beliefs and desires *as we ordinarily conceive of them*. Thus to spell out the belief-desire thesis in detail would require that we explicate our intuitive concepts of belief and desire. Fortunately, we need not embark on that project here.[9] To feel the arguments I will develop in the following section, I will need only a single, intuitively plausible, premise about beliefs.

As a backdrop for the premise that I need, let me introduce some handy terminology. I believe that Ouagadougou is the capital of Upper Volta, and if you share my interest in atlases then it is likely that you have the same belief. Of course, there is also a perfectly coherent sense in which your belief is not the same as mine, since you could come to believe that Bobo Dioulasso is the capital of Upper Volta, while my belief remains unchanged. The point here is the obvious one that beliefs, like sentences, admit of a type-token distinction. I am inclined to view belief tokens as states of a person. And I take a state to be the instantiation of a property by an object during a time interval. Two belief states (or belief tokens) are of the same type if they are instantiations of the same property and they are of different types if they are instantiations of different properties.[10] In the example at hand, the property that both you and I instantiate is *believing that Ouagadougou is the capital of Upper Volta*.

Now the premise I need for my argument concerns the identity conditions for belief properties. Cast in its most intuitive form, the premise is simply that if a particular belief of yours is true and a particular belief of mine is false, then they are not the same belief. A

bit more precisely: If a belief token of one subject differs in truth value from a belief token of another subject, then the tokens are not of the same type. Given our recent account of belief states, this is equivalent to a sufficient condition for the non-identity of belief properties: If an instantiation of belief property p_1 differs in truth value from an instantiation of belief property p_2 then p_1 and p_2 are different properties. This premise hardly constitutes an analysis of our notion of sameness of belief, since we surely do not hold belief tokens to be of the same type if they merely have the same truth value. But no matter. There is no need here to explicate our intuitive notion of belief identity in any detail. What the premise does provide is a necessary condition on any state counting as a belief. If a pair of states can be type identical (i.e., can be instantiations of the same property) while differing in truth value, then the states are not beliefs as we ordinarily conceive of them.

Before putting my premise to work, it might be helpful to note how the premise can be derived from a quite traditional philosophical account of the nature of beliefs. According to this account, belief is a relation between a person and a proposition. Two persons have the same belief (instantiate the same belief property) if they are belief-related to the same proposition. And, finally, propositions are taken to be the vehicles of truth, so propositions with different truth values cannot be identical. Given this account of belief, it follows straightforwardly that belief tokens differing in truth value differ in type. But the entailment is not mutual, so those who, like me, have some suspicions about the account of belief as a relation between a person and a proposition are free to explore other accounts of belief without abandoning the intuitively sanctioned premise that differences in truth value entail difference in belief.

The Tension between Autonomy and the Belief-Desire Thesis

In this section I want to argue that a certain tension exists between the principle of psychological autonomy and the belief-desire thesis. The tension is not, strictly speaking a logical incompatibility. Rather, there is an incompatibility between the autonomy principle and some assumptions that are naturally and all but universally shared by advocates of the belief-desire thesis. The additional assumptions are that singular causal statements like the ones extractable from our little story about Jones and the lottery ticket are often true. Moreover, they are true because they are subsumed by laws which invoke the very properties which are invoked in the characterization of the beliefs and desires. A bit less abstractly, what I am assuming is that statements like "Jones's belief that he had won the lottery was among the causes of his rushing toward the phone" are often true; and that they are true in virtue of being subsumed by laws invoking properties like *believing that he had just won the lottery.* The burden of my argument is that if we accept the

principle of autonomy, then these assumptions must be rejected. More specifically, I will argue that if the autonomy principle is accepted then there are large numbers of belief properties that cannot play a role in an explanatory psychological theory. My strategy will be to examine four different cases, each representative of a large class. In each case we will consider a pair of subjects who, according to the autonomy principle, instantiate all the same explanatory psychological properties, but who have different beliefs. So if we accept the principle of psychological autonomy, then it follows that the belief properties our subjects instantiate cannot be explanatory psychological properties. After running through the examples, I will reflect briefly on the implications of the argument for the belief-desire thesis.

Case 1: Self-referential Beliefs[11]

Suppose, as we did earlier, that we have the technology for creating atom for atom replicas of people. Suppose, further, that a replica for me has just been created. I believe that I have tasted a bottle of Chateau d'Yquem, 1962. Were you to ask me whether I had ever tasted a d'Yquem, 1962, I would reply, "Yes, I have." An advocate of the belief-desire thesis would urge, plausibly enough, that my belief is among the causes of my utterance. Now if you were to ask my replica whether he had ever tasted a d'Yquem, 1962, he would likely also reply, "Yes, I have." And surely a belief-desire theorist will also count my replica's belief among the causes of *his* utterance. But the belief which is a cause of my replica's utterance must be of a different type from the one which is a cause of my utterance. For his belief is false; he has just been created and has never tasted a d'Yquem, nor any other wine. So by the premise we set out in section II, the belief property he instantiates is different from the one I instantiate. Yet since we are replicas, the autonomy principle entails that we share all our explanatory psychological properties. It follows that the property of believing that I have tasted a Chateau d'Yquem, 1962, cannot be one which plays a role in an explanatory psychological theory. In an obvious way, the example can be generalized to almost all beliefs about oneself. If we adhere to the principle of autonomy, then beliefs about ourselves can play no role in the explanation of our behavior.

Case 2: Beliefs about One's Spatial and Temporal Location

Imagine, to vary the science fiction example, that cryogenics, the art of freezing people, has been perfected to the point at which a person can be frozen, stored, then defrosted, and at the end of the ordeal be atom for atom identical with the way he was at the beginning of the freezing process. Now suppose that I submit myself to cryogenic preservation this afternoon, and, after being frozen, I am transported to Iceland where I am stored for a century or two, then defrosted. I now believe that it is the twentieth century and that there are many strawberry

farms nearby. It would be easy enough to tell stories which would incline the belief-desire theorists to say that each of these beliefs is serving as a cause of my actions. I will leave the details to the reader's imagination. On being defrosted, however, I would presumably still believe that it is the twentieth century and that there are many strawberry farms nearby. Since my current beliefs are both true and my future beliefs both false, they are not belief tokens of the same type, and do not instantiate the same belief property. But by hypothesis, I am, on defrosting, a replica of my current self. Thus the explanatory psychological properties that I instantiate cannot have changed. So the belief property I instantiate when I now believe that it is the twentieth century cannot play any role in an explanatory psychological theory. As in the previous case, the example generalizes to a large number of other beliefs involving a subject's temporal and spatial location.

Case 3: Beliefs about Other People

Hilary Putnam (1973, 1975) has made interesting use of the following fanciful hypothesis. Suppose that in some distant corner of the universe there is a planet very much like our own. Indeed, it is so much like our own that there is a person there who is my doppelganger. He is atom for atom identical with me and has led an entirely parallel life history. Like me, my doppelganger teaches in a philosophy department, and like me has heard a number of lectures on the subject of proper names delivered by a man called "Saul Kripke." However, his planet is not a complete physical replica of mine. For the philosopher called "Saul Kripke" on that planet, though strikingly similar to the one called by the same noun on our planet, was actually born in a state they call "South Dakota," which is to the north of a state they call "Nebraska." By contrast, our Saul Kripke was born in Nebraska—our Nebraska, of course, not theirs. But for reasons which need not be gone into here, many people on this distant planet, including my doppelganger, hold a belief which they express by saying "Saul Kripke was born in Nebraska." Now I also hold a belief which I express by saying "Saul Kripke was born in Nebraska." However, the belief I express with those words is very different from the belief my doppelganger expresses using the same words, so different, in fact, that his belief is false while mine is true. Yet since we are doppelgangers the autonomy principle dictates that we instantiate all the same explanatory psychological properties. Thus the belief property I instantiate in virtue of believing that Saul Kripke was born in Nebraska cannot be a property invoked in an explanatory psychological theory.

Case 4: Natural Kind Predicates

In Putnam's doppelganger planet stories, a crucial difference between our planet and the distant one is that on our planet the substance which we call "water," which fills our lakes, etc. is in fact H_2O, while on the

other planet the substance they call "water" which fills their lakes, etc. is in fact some complex chemical whose chemical formula we may abbreviate XYZ. Now imagine that we are in the year 1700, and that some ancestor of mine hears a story from a source he takes to be beyond reproach to the effect that when lizards are dipped in water, they dissolve. The story, let us further suppose, is false, a fact which my ancestor might discover to his dismay when attempting to dissolve a lizard. For the belief-desire theorist, the unsuccessful attempt has as one of its causes the belief that lizards dissolve in water. Now suppose that my ancestor has a doppelganger on the far off planet who is told an identical sounding story by an equally trustworthy raconteur. However, as it happens that story is true, for there are lizards that do dissolve in XYZ, though none will dissolve in H_2O. The pattern should by now be familiar. My ancestor's belief is false, his doppelganger's is true. Thus the belief tokens instantiate different belief properties. But since *ex hypothesi* the people holding the beliefs are physically identical, the belief properties they instantiate cannot function in an explanatory psychological theory.[12]

This completes my presentation of cases. Obviously, the sorts of examples we have looked at are not the only ones susceptible to the sort of arguments I have been using. But let us now reflect for a moment on just what these arguments show. To begin, we should note that they do *not* show the belief-desire thesis is false. The thesis, as I have constructed it here, holds that there are psychological laws which invoke various belief and desire properties and which have a substantive role to play in the explanation of behavior. Nothing we have said here would suffice to show that there are no such laws. At best, what we have shown is that, if we accept the principle of psychological autonomy, then a large class of belief properties cannot be invoked in an explanatory psychological theory. This, in turn, entails that many intuitively sanctioned singular causal statements which specify a belief as a cause of an action cannot be straightforwardly subsumed by a law. And it is just here, I think, that our argument may serve to undermine the belief-desire thesis. For the plausibility of the thesis rests, in large measure, on the plausibility of these singular causal statements. Indeed, I think the belief-desire thesis can be profitably viewed as the speculation that these intuitively sanctioned singular causal statements can be cashed out in a serious psychological theory couched in terms of beliefs and desires. In showing that large numbers of these singular causal statements cannot be cashed out in this way, we make the speculation embodied in the belief-desire thesis appear idle and unmotivated. In the section that follows, I will consider a way in which an advocate of the belief-desire thesis might try to deflect the impact of our arguments, and indicate the burden that this escape route imposes on the belief-desire theorist.

A Way Out and Its Costs

Perhaps the most tempting way to contain the damage done by the arguments of the previous section is to grant the conclusions while denying their relevance to the belief-desire thesis. I imagine a critic's objection going something like this: "Granted, if we accept the autonomy principle, then certain belief properties cannot be used in explanatory theories. But this does nothing to diminish the plausibility of the belief-desire thesis, because the properties you have shown incompatible with autonomy are the *wrong kind* of belief properties. All of the examples you consider are cases of *de re* beliefs, none of them are *de dicto* beliefs. But those theorists who take seriously the idea of constructing a belief-desire psychological theory have in mind a theory invoking *de dicto* beliefs and desires. *De re* beliefs are a sort of hybrid; a person has a *de re* belief if he has a suitable underlying *de dicto* belief, *and* if he is related to specific objects in a certain way. But it is only the underlying *de dicto* belief that will play a role in psychological explanation. Thus your arguments do not cast any serious doubt on the belief-desire thesis."[13]

Before assessing this attempt to protect the belief-desire thesis, a few remarks on the *de dicto/de re* distinction are in order. In the recent philosophical discussion of *de re* and *de dicto* beliefs, the focus has been on the logical relations among various sorts of belief attributions. Writers concerned with the issue have generally invoked a substitution criterion to mark the boundary between *de dicto* and *de re* belief attributions. Roughly, a belief attribution of the form

S believes that *p*

is *de re* if any name or other referring expression within *p* can be replaced with a co-designating term without risk of change of truth value; otherwise the attribution is *de dicto*.[14]

But now given this way of drawing the *de re/de dicto* distinction, my imagined critic is simply wrong in suggesting that all of the examples I used in my arguments are cases of *de re* belief. Indeed, just the opposite is true; I intend all of the belief attribution in my examples to be understood in the *de dicto* sense, and all my arguments work quite well when they are read in this way. Thus, for example, in Case 3 I attribute to myself the belief that Saul Kripke was born in Nebraska. But I intend this to be understood in such a way that

Stich believes 'Φ' was born in Nebraska

might well be false if 'Φ' were replaced by a term which, quite unbeknownst to me, in fact denotes Saul Kripke.

There is, however, another way the critic could press his attack that sidesteps my rejoinder. Recently, a number of writers have challenged

the substitutional account of the *de dicto/de re* distinction. The basic idea underlying their challenge is that the term *'de re'* should be used for all belief attributions which intend to ascribe a "real" relation of some sort between the believer and the object of his belief. The notion of a real relation is contrasted with the sort of relation that obtains between a person and an object when the object happens to satisfy some description that the person has in mind.[15] Burge, for example, holds that "a *de dicto* belief is a belief in which the believer is related only to a completely expressed proposition *(dictum),*" in contrast to a *de re* belief which is "a belief whose correct ascription places the believer in an appropriate, *nonconceptual, contextual* relation to the objects the belief is about."[16] Thus, if Brown believes that the most prosperous Oriental rug dealer in Los Angeles is an Armenian, and if he believes it simply because he believes all prosperous Oriental rug dealers are Armenian, but has no idea who the man may be, then his belief is *de dicto.* By contrast, if Brown is an intimate of the gentleman, he may have the *de re* belief that the most prosperous Oriental rug dealer in Los Angeles is an Armenian. The sentence

Brown believes that the most prosperous Oriental rug dealer in Los Angeles is an Armenian.

is thus ambiguous, since it may be used either in the *de re* sense to assert that Brown and the rug dealer stand in some "appropriate, nonconceptual, contextual relation" or in the *de dicto* sense which asserts merely that Brown endorses the proposition that the most prosperous rug dealer in Los Angeles (whoever he may be) is an Armenian.

The problem with the substitutional account of the *de dicto/de re* distinction is that it classifies as *de dicto* many belief attributions which impute a "real" relation between the believer and the object of his belief. In many belief attributions the names or definite descriptions that occur in the content sentence do a sort of double duty. First, they serve the function commonly served by names and descriptions; they indicate (or refer to) an object, in this case the object to which the believer is said to be related. The names or descriptions in the content sentence *also* may serve to indicate how the believer conceives of the object, or how he might characterize it. When a name or description serving both roles is replaced by a codesignating expression which does *not* indicate how the believer conceives of the object, then the altered attribution (interpreted in the "double duty" sense) will be false. Thus the substitutional account classifies the original attribution as *de dicto,* despite its imputation of a "real" relation between believer and object.[17]

Now if the *de dicto/de re* distinction is drawn by classifying as *de re* all those belief attributions which impute a "real" relation between believer and object, then the critic conjured in the first paragraph of this section is likely right in his contention that all of my arguments invoke examples of *de re* beliefs. Indeed, the strategy of my arguments is to cite an

example of a *de re* (i.e., "real relation") belief, then construct a second example in which the second believer is a physical replica of the first, but has no "real relation" to the object of the first believer's belief. However, to grant this much is not to grant that the critic has succeeded in blunting the point of my arguments.

Let me begin my rejoinder with a fussy point. The critic's contentions were two: first, that my examples all invoked *de re* belief properties; second, that *de re* belief properties are hybrids and are analyzable into *de dicto* belief properties. The fussy point is that even if both the critic's contentions are granted, the critic would not quite have met my arguments head on. The missing premise is that *de dicto* belief properties (construed now according to the "real relation" criterion) are in fact compatible with the principle of psychological autonomy. This premise may be true, but the notion of a "real" relation, on which the current account of *de dicto* belief properties depends, is sufficiently obscure that it is hard to tell. Fortunately, there is a simple way to finesse the problem. Let us introduce the term *autonomous beliefs* for those beliefs that a subject must share with all his replicas; and let us use the term *non-autonomous* for those beliefs which a subject need not share with his replica.[18] More generally, we can call any property which an organism must share with its replicas an *autonomous property*. We can now reconstrue the critic's claims as follows:

1. All the examples considered in section III invoke non-autonomous belief properties.
2. Non-autonomous belief properties are hybrids, analyzable into an underlying autonomous belief property (which can play a role in psychological explanation) plus some further relation(s) between the believer and the object of his belief.

On the first point I naturally have no quarrel, since a principle purpose of this paper is to show that a large class of belief properties are non-autonomous. On the second claim, however, I would balk, for I am skeptical that the proposed analysis can in fact be carried off. I must hasten to add that I know of *no argument* sufficient to show that the analysis is impossible. But, of course, my critic has no argument either. Behind my skepticism is the fact that no such analysis has ever been carried off. Moreover, the required analysis is considerably more demanding than the analysis of *de re* belief in terms of *de dicto* belief, when the distinction between the two is drawn by the substitutional criterion. For the class of autonomous beliefs is significantly smaller than the class of *de dicto* beliefs (characterized substitutionally).[19] And the most impressive attempts to reduce *de re* beliefs to *de dicto* plainly will not be of much help for the analysis my critic proposes.[20] But enough, I have already conceded that I cannot prove my critic's project is impossible. What I do hope to have established is that the critic's burden is the burden of the belief-desire theorist. If the reduction of non-autonomous

beliefs to autonomous beliefs cannot be carried off, then there is small prospect that a psychological theory couched in terms of beliefs and desires will succeed in explaining any substantial part of human behavior.

A final point. It might be argued that, however difficult the analysis of non-autonomous beliefs to autonomous ones may be, it must be possible to carry it off. For, the argument continues, a subject's non-autonomous beliefs are determined in part by the autonomous psychological properties he instantiates and in part by his various relations to the objects of the world. Were either of these components suitably altered, the subject's non-autonomous beliefs would be altered as well. And since non-autonomous beliefs are jointly determined by autonomous psychological properties and by other relations, there must be some analysis, however complex, which specifies how this joint determination works. Now this last claim is not one I would want to challenge. I am quite prepared to grant that non-autonomous beliefs admit of some analysis in terms of autonomous psychological properties plus other relations. But what seems much more doubtful to me is that the autonomous properties invoked in the analysis would be *belief properties*. To see the reasons for my doubt, let us reflect on the picture suggested by the examples in section III. In each case we had a pair of subjects who shared all their autonomous properties though their non-autonomous beliefs differed in truth value. The difference in truth value, in turn, was rooted in a difference in reference; the beliefs were simply about different persons, places or times. In short, the beliefs represented different states of affairs. If the non-autonomous belief properties of these examples are to be analyzed into autonomous psychological properties plus various historical or external relations, then it is plausible to suppose that the autonomous psychological properties do not determine a truth value, an appropriate reference or a represented state of affairs. So the state of exhibiting one (or more) of these autonomous properties itself has no truth value, is not referential, and does not represent anything. And this, I would urge, is more than enough reason to say that it is not a belief at all. None of this amounts to an *argument* that non-autonomous beliefs are not analyzable into autonomous ones. Those who seek such an analysis are still free to maintain that there will be at least one autonomous belief among the autonomous properties in the analysans of each non-autonomous belief property. But in the absence of an argument for this claim, I think few will find it particularly plausible. The ball is in the belief-desire theorist's court.[21]

Appendix

A bit more needs to be said about the premise urged at the end of section II. The premise, it will be recalled, was this:

If a belief token of one subject differs in truth value from a belief token of another subject, then the tokens are not of the same type.

A number of helpful critics have pointed out to me that we actually have a variety of intuitively sanctioned ways to decide when two belief tokens are of the same type. Moreover, some of these patently violate my premise. Thus, for example, if Jones and Smith each believes that he will win the next presidential election, there would be no intuitive oddness to the claim that Jones and Smith have the same belief. Though, of course, if Jones's belief is true, Smith's belief is false. It would be equally natural in this case to say that Jones and Smith have different beliefs. So I cannot rest my premise on our intuitive judgments; the intuitions will not bear the weight.

I think the best way of defending the premise is to make clear how it is related to a certain view (actually a category of views) about what beliefs are. The views I have in mind all share two features in common:

(i) they take belief to be a relation between a believer and a type of abstract object;

(ii) they take the abstract objects to be representational—that is, the abstract objects are taken to picture the world as being a certain way, or to claim that some state of affairs obtains. Thus the object, along with the actual state of the believer's world, determines a truth value.

For example, certain theorists take belief to be a relation between a person and a proposition; a proposition, in turn, determines a truth value for every possible world—truth for those worlds in which it is true and falsity for those worlds in which it is false. A person's belief is true if the proposition is true in his or her world. Rather more old fashioned is the theory which holds belief to be a relation between a person and an image or a mental picture. The belief is true if and only if the mental picture correctly depicts the believer's world.

Now on views such as these which take belief to be a relation between a person and an abstract object, the most natural way of determining when a pair of belief tokens are of the same type is by appeal to the abstract objects. A pair of subjects' belief tokens are of the same type when the subjects are related to the same abstract object. Thus when subjects are in the same possible world, their belief tokens are of the same type only if they are identical in truth value. And this, in effect, was the premise advanced in section II. The thesis of this paper is best taken to be that the principle of psychological autonomy is in conflict with the belief-desire thesis, *when beliefs are construed as in (i) and (ii).* Let me add a final observation. A number of theorists have taken belief to be a relation between a person and a sentence or sentence-like object. For example, Jerry Fodor (1975) holds that belief is a relation between a person and a sentence in "the language of thought." It is interesting

to ask whether a theory like Fodor's is at odds with the principle of psychological autonomy. The answer, I think, turns on whether the sentences in the language of thought are taken to have truth values, and whether their referring expressions are taken to determine a referent in a given world, independent of the head in which they happen to be inscribed. If sentences in the language of thought are taken to be analogous to Quine's eternal sentences, true or false in a given world regardless of who utters them or where they may be inscribed, then Fodor's view will satisfy (i) and (ii) and will run head on into the principle of psychological autonomy. For Fodor, I suspect, this would be argument enough to show that the sentences in the language of thought are not eternal.

Notes

1. The clearest and most detailed elaboration of this view that I know of is to be found in Goldman 1970. The view is also argued in Brandt and Kim 1963, and in Davidson 1963. However, Davidson does not advocate the belief-desire thesis as it will be construed below (cf. n. 8).

2. Kim's account of supervenience is intentionally non-committal on the sort of necessity invoked in the definition. Different notions of necessity will yield different, though parallel, concepts of supervenience.

3. This weaker principle is discussed at some length in Kim 1977.

4. Note, however, that physical properties that are irrelevant in this sense may nonetheless be *causally* related to those physical properties upon which psychological properties supervene. Thus they may be "psychologically relevant" in the sense that they may play a role in the explanation of how the organism comes to have some psychological property.

5. It has been my experience that psychologists who agree on little else readily endorse the autonomy principle. Indeed, I have yet to find a psychologist who did not take the principle to be obviously true. Some of these same psychologists also favored the sort of belief-desire explanations of action that I will later argue are at odds with the autonomy principle. None, however, was aware of the incompatibility, and a number of them vigorously resisted the contention that the incompatibility is there.

6. For a critique of these views, cf. Goldman 1970, chap. 3; Alston 1967b.

7. For discussion of these matters, see Kim 1973. Kim defends the view that the property invoked in the description must be identical with the one invoked in the law. For a much more liberal view see Davidson 1967.

8. Thus Davidson is not an advocate of the belief-desire thesis as I am construing it. For on his view, though beliefs and desires may be among the causes of actions, the general laws supporting the causal claims are not themselves couched in terms of beliefs and desires (cf. Davidson 1970). But Davidson's view, though not without interest, is plainly idiosyncratic. Generally, philosophers who hold that beliefs and desires are among the causes of behavior also think that there are psychological laws to be found (most likely

probabilistic ones) which are stated in terms of beliefs and desires (cf. Hempel 1965, 463–487; Alston 1967a, 1967b; Goldman 1970, chaps. 3 and 4).

We should also note that much of recent psychology can be viewed as a quest for psychological laws couched in terms of beliefs and/or desires. There is, for example, an enormous and varied literature on problem solving (cf. Newell and Simon 1972) and on informal inference (cf. Nisbett and Ross 1980) which explores the mechanisms and environmental determinants of belief formation. Also, much of the literature on motivation is concerned with uncovering the laws governing the formation and strength of desires (cf. Atkinson 1964).

9. For an attempt to explicate our informal concepts of belief and desire in some detail, see Stich (1983).

10. For more on this way of viewing states and events, cf. Kim 1969 and 1976. I think that most everything I say in this paper can be said as well, though not as briefly, without presupposing this account of states and events.

11. The examples in Case 1 and Case 2, along with my thinking on these matters, have been influenced by a pair of important papers by Castañeda 1966 and 1967.

12. We should note that this example and others invoking natural kind words work only if the extension of my ancestor's word 'water' is different from the extension of the word 'water' as used by my ancestor's doppelganger. I am inclined to agree with Putnam that the extensions are different. But the matter is controversial. For some support of Putnam's view, see Kripke 1972 and Teller 1977; for an opposing view cf. Zemach 1976. Incidentally, one critic has expressed doubt that my doppelganger and I could be physically identical if the stuff called 'water' on the far off planet is actually XYZ. Those who find the point troubling are urged to construct a parallel example using kinds of material not generally occurring within people.

13. The idea that *de dicto* beliefs are psychologically more basic is widespread. For a particularly clear example, see Armstrong 1973, 25–31. Of the various attempts to analyze *de re* beliefs in terms of *de dicto* beliefs, perhaps the best known are to be found in Kaplan 1968 and Chisholm 1976.

14. The substitutional account of *de re*/*de dicto* distinction has a curious consequence that has been little noted. Though most belief sentences of the form

S believes that Fa

can be used to make either *de re or de dicto* attributions, the substitutional account entails that some can only be used to make *de re* attributions. Consider, for example.

(i) Quine believes that the Queen of England is a turtle.

The claim of course, is false. Indeed, it is *so* false that it could not be used to make a *de dicto* belief attribution. For in all likelihood, there is *no* name or definite description Φ denoting Elizabeth II such that

Quine believes that Φ is a turtle

is true. Thus 'Quine believes that the Queen of England is a turtle' is false and cannot be turned into a truth by the replacement of 'the Queen of England' by a codesignating expression. So on the substitutional account, this sentence can be used to make only *de re* attributions. A parallel problem besets Quine's well known substitutional account of a *purely referential position* (Quine 1960, 142 ff.) In (i), the position occupied by 'the Queen of England' can only be regarded as purely referential.

15. For more on the distinction between "real" relations and mere "satisfaction" relations, cf. Kim 1977.

16. Burge 1977, 345 and 346; last emphasis added.

17. For more on this "double duty" view of the role of names and descriptions in content sentences, see Loar 1972.

18. Of course when the notion of a "real relation" has been suitably sharpened it might well turn out that the autonomous/non-autonomous distinction coincides with the "real relation" version of the *de dicto/de re* distinction.

19. For example, when I say, "I believe that Kripke was born in Nebraska," I am attributing to myself a belief which is substitutionally *de dicto*, but not autonomous.

20. Kaplan's strategy, for example, will be of no help, since his analysans are, for the most part, non-autonomous substitutionally *de dicto* belief sentences (cf. Kaplan 1968; Burge 1977, 350 ff.).

21. I am indebted to Robert Cummins, Jaegwon Kim, William Alston and John Bennett for their helpful comments on the topics discussed in this paper. After completing this paper, I was delighted to discover a very similar view in Perry 1979. Fodor 1980 defends a version of the principle of psychological autonomy.

References

Alston, W. P. (1967a). Motives and motivation. *The encyclopedia of philosophy*. New York: MacMillan.

Alston, W. P. (1967b) "Wants, Actions and causal explanations." In H. N. Castañeda, ed., *Intentionality, minds and perception*. Detroit: Wayne State University Press.

Armstrong, D. M. (1973). *Belief, truth and knowledge*. Cambridge: Cambridge University Press.

Atkinson, J. W. (1964). *An introduction to motivation*. New York: Van Nostrand.

Brandt, R. B., and Jaegwon Kim (1963). Wants as explanations of actions. *Journal of Philosophy* 60, 425–435.

Burge, T. (1977). Belief de re. *Journal of Philosophy* 74, 338–362.

Castañeda, H. N. (1966). 'He': A study in the logic of self-consciousness. *Ratio* 8, 130–157.

Castañeda, H. N. (1967). Indicators and quasi-indicators. *American Philosophical Quarterly* 4, 85–100.

Chisholm, R. (1976). *Person and object*. LaSalle, IL: Open Court.

Davidson, D. (1963). Actions, reasons and causes. *Journal of Philosophy* 60, 685–700.

Davidson, D. (1967). Causal relations. *Journal of Philosophy* 64, 691–703.

Davidson, D. (1970). Mental events. in L. Foster and J. W. Swanson, eds., *Experience and Theory*. Amherst: University of Massachusetts Press.

Fodor, J. (1975). *The Language of Thought*. New York: Crowell.

Fodor, J. (1980). Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and Brain Sciences* 3, 63–73.

Goldman A. (1970). *A Theory of Human Action*. Englewood Cliffs, NJ: Prentice-Hall.

Hempel, C. G. (1965). *Aspects of Scientific Explanation*. New York: Free Press.

Kaplan, D. (1968). Quantifying in. *Synthese* 19, 178–214.

Kim, J. (1969). Events and their descriptions: Some considerations. In N. Rescher et al., eds., *Essays in Honor of C. G. Hempel*. Dordrecht, Holland: Reidel.

Kim, J. (1973). "Causation, nomic subsumption and the concept of event." *Journal of Philosophy*, 70, 217–236.

Kim, J. (1976). Events as property-exemplifications. In M. Brand and D. Walton, eds., *Action Theory*. Dordrecht, Holland: Reidel.

Kim, J. (1977). Perception and reference without causality. *Journal of Philosophy* 74, 606–620.

Kim, J. (1978). Supervenience and nomological incommensurables. *American Philosophical Quarterly* 15, 2, 149–156.

Kripke, S. (1972). Naming and necessity. In D. Davidson and G. Harman, eds., *Semantics and Natural Language*. Dordrecht, Holland: Reidel.

Loar, B. (1972). Reference and propositional attitudes. *Philosophical Review* 80, 43–62.

Newell, A., and H. A. Simon (1972). *Human Problem Solving*, Englewood Cliffs: Prentice-Hall.

Nisbett, R., and L. Ross (1980). *Human Inference: Strategies and Shortcomings of Social Judgment*, Englewood Cliffs, NJ: Prentice-Hall.

Perry, J. (1979). The problem of the essential indexical. *Noûs* 13, 3–21.

Putnam, H. (1973). Meaning and reference. *Journal of Philosophy* 70, 699–711.

Putnam, H. (1975). The meaning of 'meaning'. In K. Gunderson, ed., *Language, Mind and Knowledge*. Minneapolis: University of Minnesota Press.

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

Stich, S. (1983). *From Folk Psychology to Cognitive Science*. Cambridge, MA: Bradford Books/ MIT Press.

Teller, P. (1977). Indicative introduction. *Philosophical Studies* 31, 173–195.

Zemach, E. (1976). Putnam's theory on the reference of substance terms. *Journal of Philosophy* 83, 116–127.