Facta Philosophica 7, 2005: 121–144

The Belief-Desire Law

Christopher Gauker

1 Introduction

For a variety of reasons, philosophers have wanted to believe that there is a body of psychological laws governing intentional states such as beliefs and desires. One reason is that they think we rely on such laws when we explain people's behavior in terms of beliefs and desires. Another reason is that they think we can explain what beliefs and desires *are* by saying that they are essentially things that obey such laws.

If there really are such laws of belief and desire, then it ought to be possible to give some clear examples. This is not to say that the content of such laws should be common sense or discoverable by introspection. Nor is it to say that they must be folk psychological laws that everyone who speaks of beliefs and desires must already know in some way. If there are laws that govern belief and desire, then they may be discoverable only through empirical research. Perhaps we must have some inkling of the content of these laws insofar as we are able even now to give good explanations of behavior in terms of beliefs and desires, but such generalizations as we rely on might be only crude approximations to the truth.

Nonetheless, if there really are such laws, then it should be possible even now to think of some plausible examples. It should be possible to put forward some hypotheses that we cannot refute on the basis of common experience as soon as we think them up. If the content of these laws is an empirical question, then we should be able even now to think of some hypotheses that we cannot refute just by thinking about them, between which we can decide only by empirical research. Empirical research may suggest to us generalizations that we cannot think up just by exercising our imaginations, but if such research is not to be blind, it must be guided by some hypotheses that, as far as we know, might be true. In saying this, I am not singling out intentional psychology by imposing a requirement that we do not place on any

other science. Since the dawn of recorded history, people have been putting forward hypotheses about all kinds of things, hypotheses that were plausible from the points of view they occupied at the time and that they could not refute just by thinking about them.

In fact, the philosophical literature that posits such laws of belief and desire seldom makes any serious effort to articulate specific laws. One can find many journal articles in which philosophers discourse at length about the necessity and significance of such laws without ever giving a serious example. There is, however, one example that comes up time and again. We are often told that something like the following is a basic law of belief-desire psychology: *People do what they believe will satisfy their desires.* Below I will examine a wide variety of attempts at a more accurate formulation of this law. For now, let us take this crude formulation as a stand-in for whatever might turn out to be the more accurate formulation. Taking this formulation as merely a placeholder, I will call this principle *the belief-desire law*.

In this paper, I will argue that there is no such belief-desire law. Inasmuch as this is one of the few examples that philosophers have been able to come up with, my critique of the belief-desire law should cast doubt on the whole conception of beliefs and desires that rests on there being such laws. I will not, in this paper, attempt any more constructive, alternative account of explanation in terms of beliefs and desires. (Certainly one should not infer from my criticism any endorsement of what is called "simulation theory".)

I should emphasize that it is by no means my intention to defend eliminativism about beliefs and desires. My objective is not to show that beliefs and desires do not exist by denying that there are the sorts of constitutive psychological laws that there would have to be if beliefs and desires existed. My assumption is that *of course* beliefs and desires exist. Just try to imagine a human culture in which people do not talk about beliefs and desires! My point is, rather, that it is a mistake in the first place to think that there must be laws of intentional psychology just because beliefs and desires do exist.

My critique will not turn on a distinction between generalizations that are merely true and those that are in addition "law-like". In the next section I will explain what work we might expect a true principle of belief-desire psychology to do, and someone might argue that any generalization that can do that work must have a modal status stronger than ordinary truth. But that that is so will not be any part of my argument.

2 The Need For Laws Of Belief And Desire

One of the reasons why philosophers have supposed that there must be laws of intentional psychology is that they have conceived of explanation in such a way that there must be some. They start with the observation that we may sometimes explain what a person has done by citing his or her beliefs and desires. They then suppose that citing beliefs and desires can be explanatory only insofar as there are true generalizations relating beliefs and desires to action; anyone who offers such explanations must, at some level, in some way, be in cognitive touch with those generalizations. Call this the *explanatory rationale* for positing laws of intentional psychology. Philosophers who have accepted this explanatory rationale have been otherwise as diverse as Dennett (1987), Fodor (1975) and Churchland (1979). (In Churchland's case, the point was not that there really are such true generalizations but only that if explanation in terms of beliefs and desires were any good, and if beliefs and desires really did exist, then there would be some. Churchland famously questioned the truth of all such generalizations.)

In the same vein, laws of intentional psychology have figured into widely accepted conceptions of the attribution of beliefs and desires. Supposedly, we have some understanding of the sorts of beliefs a person will form when presented with various stimuli, and we have some understanding of what a person will do given that he or she has certain beliefs and desires. So, we can figure out what people believe and desire by observing what they do in response to what happens around them and then infer that they have just such beliefs and desires as might mediate in a law-governed way between their sensory inputs and their behavioral outputs. For an early example of a theory of interpretation along these lines, see David Lewis 1974. (For the expression of Lewis's commitment to something like the belief-desire law, see his discussion of the "rationalization principle," p. 337.)

The other important reason to suppose that there must be laws of intentional psychology is that in terms of such laws we might hope to explain the very nature of belief and desire. Call this the *ontological rationale* for positing laws of intentional psychology. Commonly the basic idea is expressed in terms of the method of "ramsification". (What we now call ramsification is actually due to David Lewis 1970; we call it ramsification because Lewis took his cue from an idea of F.P. Ramsey's.) If we have a theory T comprising all of the essential truths about a kind of thing designated k, so that the theory can be

represented as *T(k)*, then supposedly we can define that kind by putting a variable *x* in place of every occurrence of *k* and then writing:

The *k's* = the unique kind *x* such that *T(x)*.

Thus, if we had a sufficiently rich theory *P* of essential truths about beliefs, then we could put the variable *x* in place of "belief" throughout the theory and define beliefs as the unique kind *x* such that *P(x)*. Such an approach to defining beliefs and desires is endorsed by a large number of philosophers, including Lewis (1972), Loar (1981) and Rey (1997).

Actually, the idea that these authors are trying to express is probably not best formulated in terms of ramsification. Most of these philosophers have also wished to allow that beliefs and desires are *multiply realizable*. The belief that it is raining is something we could find in a Martian or a robot as well as in a human being. But the physical state that the Martian or robot is in when it believes that it is raining will not be the physical state that the human being is in when he or she believes that it is raining. The problem is that if beliefs are multiply realizable, then it is not so clear that we can say that belief is *the* kind of thing *x* such that …, where we then say everything about *x* that our theory says about belief. That seems to imply that there is just one kind of thing that can be a belief when we want to allow that many physical kinds may be beliefs.

For this reason I think that those who want to say that a body of intentional laws is constitutive of the very nature of beliefs and desires would do better to formulate their thesis using the logician's concept of a *model*. A *model* of a theory assigns objects and properties to the nonlogical vocabulary of the language in which the theory is formulated in such a way that the theory is true on that assignment. What proponents of the ontological rationale should say is something like this: Let us distinguish between the theoretical and the nontheoretical terms that occur in our theory *T*. The theoretical terms will include at least "believes" and "desires," and the nontheoretical terms might include terms such as "hand," and "eye" and "causes". Then we may define an *admissible interpretation* of the language of our theory as any interpretation that assigns to each nontheoretical term what it really refers to. For example, in any admissible interpretation, we would have to interpret the term "hand" as referring to the property of being a hand. Then we could say that for any kind of thinking thing, *H*, a property *B* qualifies as the state of *belief* for *H*'s if and only if the property *B* is the property that we interpret "believes" as referring to when we model

the theory *T* in terms of the properties of *H*'s. This allows the multiple realizability of belief inasmuch as the property that we interpret "believes" as referring to when we model the theory in terms of the properties of humans may be different from the properties we interpret "believes" as referring to when we model the theory in terms of the properties of Martians.

In view of the epistemological and ontological rationales for seeking laws of intentional psychology, we can make a couple of assumptions about the character of the requisite laws. The first is that the laws have to be fairly general. They have to have application across a wide variety of circumstances. This requirement is very vague, but not too vague to apply in some cases. So neither rationale will be vindicated if the only kinds of examples we can come up with are such as "The moon looks larger on the horizon than when it is high in the sky" or "People don't like to be insulted." Certainly the ontological rationale cannot make do with only such narrow generalizations, because it will not be plausible that such generalizations capture the essential nature of belief and desire. But even the explanatory rationale requires principles more general than these because explanations resting on principles no more general than these will not be very satisfying explanations.

My other basic assumption is that the required principles must include principles that characterize sequences of events. They cannot cite only static relations among states of mind. An example of a principle governing only such a static relation would be: "If someone believes that if *p* then *q*, and he or she believes that *p*, then he or she does not believe that not-*q*." Such static principles may be part of what is needed, but they alone cannot suffice. They will not suffice for the explanatory rationale, because in terms of such principles alone we cannot explain what *happens*. They also will not suffice for the ontological rationale because the basic idea is that we can explain what beliefs and desires are by explaining what they *do*—where they come from and what they cause.

3 Initial Attempts

No one who believes in the belief-desire law will want to formulate it in just the way I formulated it above. People always want to improve upon this formulation by adding various qualifications. Furthermore, almost everybody acknowledges that we need to insert a *ceteris paribus* clause. I will discuss the *ceteris paribus* clause in some detail later on, but first let us see what goes wrong if we try to get by on other sorts of

qualifications. The general pattern we find is that either the theory rests on false presuppositions, or we can think of counterexamples, or it is too vague, or it is vacuous.

Here, again, is the simple formulation of the belief-desire law:

> *The simple formulation:* People do what they believe will satisfy their desires.

The problem with the simple formulation is that is rests on a false presupposition. There is never just one thing people desire; they always desire a lot of things. They cannot do everything they think will satisfy all of their desires, because they cannot do all of those things at once. Suppose that I am thirsty, and so desire a drink of water, and so am about to stoop over to drink from a water fountain. At that same moment, someone passes by with whom I wish to speak. Well, I cannot both stoop for the drink and stop the passerby. I have to choose.

Suppose, then, that we modify the formulation of the law to read as follows:

> *The strongest desire formulation:* People will do what they believe will satisfy their *strongest* desire.

The problem with this is that, taken in one way, it is just false, and taken another way it is too vague. One does not always act immediately on one's strongest desires. It may be more important to me to talk to the person who is passing by than to take a drink of water at that very moment; but I might nonetheless take a drink of water knowing that I can still catch up to the person a moment later. So the principle is false. In response, it might be said that the pertinent choice is between drinking now and talking later or talking now and drinking later, and even though talking soon is more important to me than drinking right now, I may prefer drinking now and talking later over talking now and drinking later. But the principle itself does not say any of this and does not carry with it any account of how the options between which a person has to choose are to be delineated, and so if it is not just false, then it is too vague.

To get around this, we might identify a person's "strongest desire" with whatever desire the person chooses to act on, and consequently rewrite the principle as follows:

> *The choice formulation:* People do what they believe will satisfy the desire that they have chosen to act on.

One problem with this formulation is that it considerably narrows the scope of the principle and to that extent weakens the force of the be-

lief-desire law as an illustration of the sort of principle required by the epistemological and ontological rationales. Even so, it is still false. People always have many desires that they have chosen to act on, but they cannot actively pursue the satisfaction of each one at every moment. We could try to answer that by writing, "Whenever a person performs an action *A*, he or she believes that action *A* will satisfy the desire that moves him or her to perform it." But this principle does not characterize a sequence of events, since the pertinent desire is identified only in terms of the action it moves. Consequently, this example does not assure us that we can have, as the epistemological and ontological rationales require, general principles on intentional psychology that describe sequences of events.

The ways in which attempts to formulate an adequately hedged belief-desire law may go wrong are not confined to just these however. I do not want to spend a lot of words taking apart in detail every actual proposal that has been made. But here, for the sake of illustration, is a representative example. Terence Horgan and James Woodward, referring to "folk psychology," write:

> The theory asserts, for example, that if someone desires that p, and this desire is not overridden by other desires, and he believes that an action of kind K will bring it about that p, and he believes that such an action is within his power, and he does not believe that some other kind of action is within his power and is a preferable way to bring it about that p, then *ceteris paribus*, the desire and the beliefs will cause him to perform an action of kind K. (Horgan and Woodward 1985, p. 197.)

(Other attempts may be found in Ayer, 1970, pp. 233–34, Loar 1981, p. 90, Newell 1990, pp. 48–49, Grice 1989, p. 285.) For the moment, let us ignore the *ceteris paribus* clause; I will come back to that. In this formulation, Horgan and Woodward insert two important hedges. The first is that the desire that *p* is not supposed to be "overridden" by any other desire. The second is that the agent is not supposed to know of any "preferable" way to bring it about that *p*. Both of these are in danger of rendering the whole statement either vacuous or false.

To see this, consider just the first of these hedges. When does an agent have a desire that is not "overridden" by another desire? If this just means that the agent prefers no other outcome, then it is doubtful whether the condition will ever be satisfied. For every condition that might be the object of one's desire, one can imagine something even a little better. So on this interpretation the hypothesis of the generalization is never satisfied (so that the principle is merely vacuously true). Perhaps to say that the desire that *p* is not overridden by any other desire just means that the agent is not prevented by any other

desire from acting on the desire that *p*. No doubt such a condition might sometimes be fulfilled. But if someone prefers the outcome that *q* over the outcome that *p*, that does not mean that the desire that *q* *prevents* the agent from acting on the desire that *p*. So we could have a situation in which the hypothesis of the Horgan and Woodward formulation were fulfilled, because the agent was not prevented by any other desire from acting on the desire that *p*, and yet the agent might try to bring it about that *q* instead since that is what he or she preferred.

What all of these attempts to formulate a belief-desire law ignore is that, quite generally, deliberate choice involves the weighing of several options against one another with respect to the desirability of the various outcomes and the likelihood that each of those outcomes will result from the action chosen. No attempt to formulate a belief-desire law that ignores the comparison of outcomes with respect to desirability and probability can possibly be correct. Shortly I will examine some attempts to formulate a belief-desire law in a way that acknowledges this, but first I need to examine the role of the *ceteris paribus* clause.

4 The *Ceteris Paribus* Clause

I think most people who believe in some kind of belief-desire law will accept that we are not likely to formulate any exceptionless generalizations relating beliefs and desires to actions. One reason is that there are bound to be even more basic laws at the level of the purely physical description of reality. Since a collection of atoms will not realize a mental system with perfect reliability, we have to expect occasional breakdowns or misfires. A truly universal characterization of human behavior would be available only at the level of atoms, or even lower. Even at that level we cannot expect exceptionless generalizations specifying that if such and such happens, then, a moment later, such and such other thing will happen. Even the most basic laws of physics do not specify such regularities. The law of gravity does not say that two bodies of given masses will accelerate toward one another at a certain rate; for the law of gravity does not say that no other force will come along and prevent it.

The literal meaning of "*ceteris paribus*," viz., *other things being equal*, is plainly useless here. It is also useless to read it as meaning *unless not*, for that simply trivializes every proposition it attaches to. It is also not helpful to treat the *ceteris paribus* clause as introducing the following

qualification (in italics): If someone desires *x* and believes that by doing *y* he or she will obtain *x*, *and nothing prevents him or her from doing y*, then he or she will do *y*. If the only circumstance under which we are permitted to conclude that nothing prevented the agent from doing *y* is that the agent actually does *y*, then what this says is trivial. But if we substitute some independent account of the things that might prevent someone from doing *y*, such as a list of them, then again, the generalization is bound to have exceptions, since the list is bound to be incomplete, and so the generalization is bound to be false.

As an illustration of the trouble one is liable to get into in attempting to make sense of the *ceteris paribus* clause, consider the account in Paul Pietroski's *Causing Actions* (2000). (The pertinent chapter of Pietroski's book duplicates material from his 1995 article coauthored with Georges Rey.) According to Pietroski, a statement to the effect that *ceteris paribus* all F's are G's "is a true law only if its apparent exceptions are merely apparent, in that all instances of (F & ¬G) can be explained away by citing interference" (2000, p. 124). The problem is that the necessary condition that Pietroski thus places on the truth of a *ceteris paribus* law does not seem to be much of a constraint. Assuming that everything that has come to be can be explained somehow, any case of (F & ¬G) can be explained, and, for all that Pietroski says, that explanation might be described as "interference".

In reply it might be said that an "interferer" is in fact not just anything in terms of which we can explain an apparent counterexample. An interferer for a principle to the effect that *ceteris paribus* every F is G is some factor in terms of which we explain the fact that (Fa & ¬Ga) such that *if that interferer were not present*, then we would have (Fa & Ga). For example, why is it not true that *ceteris paribus* no window panes ever break? After all, for any broken window pane, there is presumably an explanation of how it came to be broken. The answer, it might be said, is that in some of those cases it will not be the case that if the factors we cite in explanation had not held, then the window pane would not have broken. Sometimes, a window pane is broken because a rock is thrown through it, and if that rock had not been thrown, or had not broken the window, then another one would have been thrown and would have broken the window. The problem with this reply is that it rules out too much. Consider some exception to the putative belief-desire law. Can we always maintain that if the factors we cite in explanation of its not being an instance had not been present then it would have been an instance? No, because, just as in the window case, if those factors had not prevented it from being an instance, there might have

been something else that would have prevented it from being an instance.[1]

Perhaps we may take the *ceteris paribus* clause as indicating that what is literally true is only a statistical generalization. So to say that *ceteris paribus* people do what they believe will satisfy their desires is just to say that there is some number *n*, which in principle we could calculate, such that what is strictly and literally true is that in *n* percent of all cases, people do what they believe will satisfy their desires. But this idea will not serve the purposes of the ontological rationale, because no mere statistical generalization can be constitutive of the very nature of beliefs and desires. We cannot maintain that the essence of belief and desire is to be such that in, say, 80% of all cases, people do what they believe will satisfy their desires. If we said that, then it would be literally impossible for there to be a world in which people literally had beliefs and desires, but in which they were a little worse on the whole as decision makers and in only 79% of all cases did what they believed would satisfy their desires.

I cannot show that a merely statistical generalization would not serve the purposes of the explanatory rationale. No doubt we are sometimes able to *predict* what people will do just because we know what people tend to do under circumstances like those present. But for two reasons I question whether such a generalization is really a basis for our explanations in terms of beliefs and desires. First, I doubt that we actually have the sort of statistical evidence that would support such a statistical generalization. Beliefs and desires, whatever they are, are not directly observable; so we could not just observe a reliable correlation, or allow, in Humean fashion, a constant conjunction to impress itself on us. We have to have some means of accessing a person's beliefs and desires, which is what the belief-desire law is supposed to provide us. Second, as we have already begun to see at the end of the last section,

1 Another attempt at a theory of *ceteris paribus* clauses is Fodor's 1991. I do not discuss it in the main text because it exhibits no common tendency in the literature. Fodor's basic idea is: *ceteris paribus* A's are B's only if: if R is a realization of A such that under no condition is R sufficient for B, then there are many other laws pertaining to A's such that, under specifiable conditions, R is not an exception to those. (See p. 27.) This proposal is no help to a defender of the belief-desire law, however. We are focusing on that just because it is our best hope for *a* law of intentional psychology that serves the explanatory and ontological rationale. If we cannot find a viable formulation of this law without first finding many others, then we will not be able to defend the explanatory and ontological rationales on the basis of this example.

our search for an adequate characterization of the rest of the content of the belief-desire law, apart from the *ceteris paribus* clause, will be guided by a conception of rational choice. We will not expect agents to act with perfect rationality, but our understanding of the ways in which people approximate to rational agents will not be well characterized as knowledge of the percentage of cases in which they act with perfect rationality.

Perhaps we might understand the *ceteris paribus* clause in the belief-desire law as telling us that what follows is only an approximation to the truth. So while it is only an approximation to the truth to say that people do what they believe will satisfy their desires, it is completely and precisely true to say that *ceteris paribus* people do what they believe will satisfy their desires, since what that means is just that it is approximately true to say that people do what they believe will satisfy their desires. This will not do either. If the *ceteris paribus* clause merely tells us that what follows is an approximation to the truth, then it should be possible in principle for us to state without the *ceteris paribus* clause the truth to which what follows it is an approximation. I do not assume that we must *know* the truth to which the belief-desire merely approximates, but we should be able to put forward some hypotheses that we cannot easily shoot down as soon as we think them up.

In response, someone might grant that we should be able to state the truth to which the belief-desire law merely approximates but might suppose that we will able to do this only after we are in possession of superior concepts. So what the belief-desire law, qualified by a *ceteris paribus* clause, really means is just that the belief-desire law is only an approximation to the real truth, which can only be formulated in terms of concepts that we do not yet possess. Of course, the superior concepts that we will employ in place of the concepts of belief and desire will have to be recognizable as successors to the concepts of belief and desire. So on this conception of the belief-desire law, there has to be some other true proposition that does *not* contain the *ceteris paribus* clause and may not be formulable *in terms* of beliefs and desires but is nonetheless in some way *about* beliefs and desires. I take for granted that, apart from a prior commitment to the belief-desire law, nobody has any reason to believe that there is any such proposition. That prior commitment is precisely what is now in question.

5 Decision Theory

Many people who believe in the belief-desire law have supposed that ultimately decision theory would yield the proper formulation (see, for example, Dennett 1971, Fodor 1975, pp. 28–29, Rey 1997, pp. 216–17). We make decisions on the basis of our beliefs and our desires. Decision theory, they have supposed, tells how to make those decisions rationally. People may be presumed to be more or less rational, that is, to make decisions more or less in accordance with the norms that decision theory articulates, and that is all the belief-desire law says. There may be no purely descriptive formulation of the belief-desire law, because there may be indefinitely many ways in which believing and desiring agents fall short of the ideal. Nonetheless, they may all approximate to the same ideal, and so we can use that ideal in formulating a version of the belief-desire law.

One virtue of this approach is that it allows us to make good sense of the *ceteris paribus* clause in the belief-desire law. On this view, what the *ceteris paribus* clause really says is just that what follows it holds only in ideal conditions. The belief-desire law will say that *ceteris paribus* an agent's beliefs and desires will be related to his or her actions in such and such manner, and the meaning of the *ceteris paribus* clause will be that we should expect this relation to hold only in ideally rational agents. Similarly, if we attach a *ceteris paribus* clause to Galileo's law of free fall, then that might indicate that the law is supposed to hold only under conditions of no friction and no lift. The belief-desire law might still be useful in explanation and prediction, because we can expect that people will often approximate to the ideal. A law that describes only ideally rational agents might still be constitutive of the very nature of belief and desire, for we might concede that any less than perfectly rational agent only imperfectly possesses the properties of believing and desiring. Alternatively, we might say that an agent's mental states qualify as full-fledged beliefs and desires if they conform closely enough to the ideal.

In taking this approach, a proponent of the belief-desire law might have to concede that "belief" and "desire" are not exactly the right terms to use in formulating the belief-desire law. Instead we might have to formulate the law in terms of an agent's subjective probability assignments and the agent's preference ranking of possible outcomes. But that concession can be taken in stride. If we can understand the explanation of behavior in terms of subjective probabilities and preferences and can explain the nature of subjective probability and preference in terms of constitutive laws, then we might hold that talk "belief" and

"desire" *per se* is merely loose talk for what is more properly described in these other terms.

Unfortunately, this confidence in decision theory to provide the content of the belief-desire law is entirely misplaced. It rests on the assumption that decision theory does in fact describe an ideal of rational decision-making that decision makers can use in making decisions and to which we can expect people to conform, to the extent that they are rational. Such an assumption is quite explicit in the writings of some authors whose philosophy of mind requires that there be laws of intentional psychology. Consider this passage from Georges Rey:

> Creatures don't only reason about how the world is; they also reason about what they *ought to do*. [...] Contemporary *decision theory* provides a framework for beginning to understand that process. For purposes here, something like the following description will serve as a good first-approximation:
> (i) The agent judges herself to be in a certain situation S;
> (ii) She judges that a certain set of exclusive and exhaustive basic acts—A_1, A_2, ... A_n—are live options for her to perform in S;
> (iii) She predicts the probable consequences, C_1, . . . C_n, of performing each of A_1 through A_n; [...]
> (iv) A preference ordering is assigned C_1 through C_n;
> (v) The agent selects one of the acts as a decision-theoretic function (e.g., maximize expected utility) of the probabilities of (iii) and the preference ordering of (iv). (1997, pp. 216–17, footnotes omitted)

In a footnote in this passage, Rey cites Fodor 1975, pp. 28–29, as his source for this characterization of the methods of decision theory, and, indeed, in that passage, Fodor says almost exactly same thing.

Unfortunately, Rey and Fodor are just wrong about what contemporary decision theory actually says. In fact, *there is no* decision-theoretic function, *no* many-to-one mapping, from probability assignments and preference orderings to acts such as Rey describes. In attempting to formulate an ideal of rational decision-making we cannot get by on preferences and subjective probability assignments alone; we must also appeal to the *distances* on a preference scale. There is no decision theoretic function such as Rey describes because the same preference rankings and the same beliefs are compatible with different choices. The correct choice depends also on relative *distance* on the preference scale, and Rey takes no account of that at all.

For example, suppose that someone thinks that the likelihood of getting heads on a flip of a certain coin is the same as the likelihood of getting tails on a flip of that coin, and also prefers a turkey sandwich to a baloney sandwich and prefers the baloney sandwich to a peanut butter sandwich. Suppose that that person is given a choice between two

bets. On Bet *A*, the person receives a turkey sandwich if the coin comes up heads and a peanut butter sandwich if it comes up tails. Bet *B* is not really a bet; the person just gets the baloney sandwich (the middle ranked outcome) straight out. If the person is rational, the choice between *A* and *B* will depend not only on the preference ranking over sandwiches but also on whether the baloney sandwich is closer in preference to the turkey sandwich or closer in preference to the peanut butter sandwich. (If baloney is closer to turkey, then the agent should prefer the unconditional offer of baloney; if baloney is closer to peanut butter, then the agent should prefer the wager.)

For this sort of reason, decision theory needs not only the agent's subjective probabilities and preferences over basic outcomes, but also a numeric measure of the *utility* of each of these outcomes. Given the utility of each of the possible outcomes of a decision, we can define the *expected utility* of an action *a* as follows: Where o_1, o_2, …, o_n are all the possible basic outcomes of an action *a*, the expected utility of *a* = (prob(o_1) x util(o_1)) + (prob(o_2) x util(o_2)) + … + (prob(o_n) x util(o_n)). And then we can make the following positive recommendation: An agent should prefer that action, of those that are available, that has the highest expected utility.

Offhand, this might look like the principle of decision theory Rey is looking for. After all, while not acknowledging the difference between a preference *ranking* and a utility *scale*, Rey does also suggest that his decision theoretic function might be described as maximization of expected utility. So, taking this correction on board, Rey and Fodor might put forward the following as their candidate for the belief-desire law:

> *The expected utility formulation:* To the extent that they are rational, people maximize expected utility.

Before we conclude that this is just the formulation of the belief-desire law that we have been looking for, though, we ought to consider more carefully what is meant by *utility*.

In decision theory utility is usually conceived as merely a measure of preference. One supposes that for any given decision problem, there is a finite array of *basic outcomes*. A *lottery* for a decision problem is a probability distribution over the basic outcomes for that decision problem. To choose is, in effect, to choose a lottery, namely, the lottery associated with the action that one chooses. A *utility scale*, conceived merely as a measure of preference, pertains exclusively to a given decision problem, and it *exists* if and only if the person's preferences over all lotteries for that problem conform to certain so-called rationality

conditions.[2] One of these, which I state here just for the sake of illustration, is the continuity axiom. It says that if *X* is preferred to *Y* and *Y* is preferred to *Z*, then there must be a real number *n* strictly between 0 and 1 such that the agent is indifferent between, on the one hand, *Y* and, on the other hand, a lottery in which the agent has a probability *n* of receiving *X* and a probability 1 – *n* of receiving *Z*. These conditions do not say anything about utility, but when they are satisfied a utility scale can be constructed. Since utility is relative to preferences in this way, as soon as a person's preferences change in any way (first she was hungry, then she ate and was not) an entirely new utility scale, utterly incomparable to the previous one, is called for.

A common misconception is that even when utility is conceived as merely a measure of preference in this way, the calculation of expected utility can serve as a method for making decisions. That is quite wrong. No utility scale even exists except insofar as *all* lotteries over basic outcomes have been ranked in accordance with the rationality conditions, which do not say anything about utility. Once the lotteries have been ranked, it is entirely redundant to calculate expected utility. One can simply choose the highest ranked lottery of those that are available. Indeed, preferences conform to the standard rationality conditions if and only if the ranking of lotteries by preference corresponds to the ranking by expected utility.

Since no utility scale exists at all unless the ranking of all lotteries satisfies the rationality conditions, and a ranking that satisfies the conditions is identical to the ranking by expected utility, the proposition that a rational person will prefer the action, i.e., lottery, with the highest expected utility is *strictly equivalent* to the proposition that a rational person's preferences will conform to the rationality conditions. This means that the expected utility formulation of the belief-desire law is strictly equivalent to the following formulation:

> *The rationality conditions formulation:* To the extent that they are rational, people's preferences over the lotteries for a given decision problem conform to the rationality conditions.

In other words, the claim that, *ceteris paribus*, people will decide in the manner that decision theory recommends is nothing more than the claim that, *ceteris paribus*, people's preferences will conform to the rationality conditions.

2 For instance, the Von Neumann-Morgenstern rationality conditions. For an elementary exposition, see Resnik 1987.

So decision theory does not tell us what a person will do given his or her beliefs and desires. It does not tell us how a person will rank lotteries given his or her beliefs and desires. It does not tell us how a person who has ranked some but not all of the lotteries will rank the rest. It does not even tell us that a person will choose the highest ranked lottery of those that are available (only that he or she will prefer it). Decision theory simply places certain basic constraints on rational preference orderings. It certainly does not teach us to calculate expected utility as a method for making a decision.[3]

One reason it is easy to make the mistake of treating utility as a basis for choice is that it is easy to slip into thinking of utility as some kind of quantity measurable independently of prior preference. For instance, we might learn to calculate expected *monetary value* (for example the expected monetary value of buying an extended warranty on a television set), observe that monetary value is not always an accurate measure of value (since, for example, the action with the lower expected monetary value might involve less risk), and then think of utility as just a kind of adjustment to monetary value. Or we might think of utility as something like pleasure and suppose that in principle it could be measured by looking at what is happening in a person's brain. Or we might think of utility as happiness and ignore the fact that there is no evident means of placing all possible outcomes on a single happiness scale. This thought, that there might be a single scale of value and that we can make decisions by ranking our options on it, will be the subject of the next section.

Another source of the idea that decision theory provides a method for making decisions might be the mistaken thought that the rationality conditions can be used as a method for filling in the details of a preference ranking given some preferences to start with. One might reason as follows that the rationality conditions can be used to decide how to rank lotteries between which one could otherwise not decide. Suppose that my top-ranked outcome in some decision problem is B (best), and my bottom-ranked outcome is W (worst). Suppose moreover, that while I rank both C and D between B and W, I am unsure how to rank C and D relative to one another. But then I find that I would be indifferent between C and a lottery offering me a 70% chance

3 I have to admit that sometimes even authors of decision theory books get confused about this and describe the calculation of expected utility as a method of making decisions. See, for example, Resnik 1987, p. 99; Jeffrey 1983. pp. 1–8; Schick 1997, pp. 35–36. The mistake is further criticized by Luce and Raiffa, 1957, p. 22, pp. 31–32, Broome 1991, Pettit 1991, and Hampton 1994.

of getting B and a 30% chance of getting W. Likewise, I find that I would be indifferent between D and a lottery in which I have a 60% chance of getting B and a 40% chance of getting W. Well, the rationality conditions tell me that I should prefer a lottery in which I have a better chance of winning a better prize. And so I prefer the lottery that gives me a 70% chance of getting B over the lottery that gives me only a 60% chance of getting B. But I am indifferent between C and the first lottery and between D and the second lottery, and so I infer (again invoking the rationality conditions) that I should prefer C over D. But in fact, this reasoning is entirely specious. Yes, if I have all of the preferences described, then I should prefer C over D. But upon realizing that fact, nothing prevents me from changing my mind instead and deciding that I should not have been indifferent between C and the first lottery or should not have been indifferent between D and the second lottery. In general, the rationality of conforming to the rationality conditions gives one no reason whatsoever to rank any single pair of outcomes one way rather than another rather than changing one's rankings of other things.

The equivalence of the expected utility formulation and the rationality conditions formulation is bad news for proponents of the belief-desire law. The rationality conditions describe only static relations between preferences. They do not say anything at all about what a person will do given that he or she has a given preference ranking. But as I explained in section 2, the epistemological and ontological rationales for seeking a belief-desire law demand more than merely a characterization of static relations between states of mind. We need a generalization that somehow characterizes sequences of events. So the rationality conditions formulation and, consequently, the equivalent expected utility formulation cannot serve us as the example of a law of intentional psychology that we are looking for.

6 Imaginary Decision Theory

A proponent of the belief-desire law might insist that even if *actual* decision theory does not provide the desired content for the belief-desire law, nonetheless, a theory of decision could in principle be worked out that would serve our purposes. After all, people do make decisions based on their beliefs and desires, and there is a distinction to be drawn between the right way to do it and the wrong way. So in principle it should be possible for us to formulate the right way to make a decision based on one's beliefs and desires. Once we have the decision theory that truly tells us how we ought to decide, we will be able to formulate

the belief-desire law as follows: *Ceteris paribus*, people make their decisions in *that* way, whatever it is.

My contention is that this is wrong. Yes, decision theory can formulate methods that are applicable under certain conditions. For example, decision theory tells us how to calculate expected monetary value; so in a situation where all that matters is monetary value, we can make our decision by applying a method that decision theory teaches. Yes, decision theory identifies a unique solution for a two-person, zero-sum game (consisting of a pair of mixed strategies in equilibrium). So if by some prior decision we have deemed it adequate to represent some situation as a two-person, zero-sum game, then we can use decision theory to make a choice. But these concessions do not concede that decision theory sometimes recommends a choice of action *all things considered*.

Many problems stand in the way of a widely applicable, normatively correct rule of decision-making, but here is one important one: We often rank our options differently on different and incomparable scales of value. For example, suppose there are two paths I can follow in walking from my house to the university campus. I can take the straight route up Clifton Avenue, on a concrete sidewalk, along a monotonous row of parking meters and a steady flow of heavy traffic. Or I can walk through Burnet Woods along a curvy, narrow, paved road. The route up Clifton Avenue is shorter but unpleasant. The route through Burnet Woods takes more time but is more pleasant. There are other factors that may come into play as well. If I take the straight route on the sidewalk, I do not have to watch my step or pay attention to anything, and so I can think about other things. If I take the route through the park, I have to look out for the occasional passing car and be careful not to step in muddy patches. But let us focus on just the comparison with respect to time and pleasantness. On some occasions my choice may be clear. If going through the park will make me late for class, then I will take Clifton Avenue. If it's an especially beautiful day and I need to unwind and I have time to spare, I will take the route through the park. In other cases, there may be no clear choice. Suppose I could measure each route on a scale of pleasantness and time saved. Suppose also that I could combine these two measures into a scale of overall value. Then I could choose the route that had the highest level of overall value. But in fact I can do no such thing. I will make a choice, of course, and there will be some cause for my choice, but there need not be anything like a sufficient reason for my choice, grounded in beliefs and desires.

The sad fact is that there is no scale of overall value such that we might form our preferences by figuring out where the possible out-

comes of our actions lie on that scale. Sometimes people naïvely suppose that the end of all human action is pleasure. But when we consider the variety of things that people choose for themselves, it becomes very clear that if we want to call all of those things "pleasure," then we cannot rely on the ordinary connotations of the word. Alternatively, we might say that the end of every human action is happiness, but the term "happiness" is not one that can be used to discriminate between ends apart from the choices people do make so that an action might be chosen *because* it promotes happiness.

In response, it might be said that all that such cases show is that we may sometimes be indifferent between two options and there may be no decisive reason to prefer the one over the other. On the contrary, I need not be indifferent in such cases. I do choose, of course, but moreover, I need not choose by simply closing my eyes and allowing my passions to pull me. I may choose deliberately and I may definitely prefer what I choose. And afterward I can give reasons for my choice by citing the virtues of the option I have chosen. Despite the contrary ranking on the scale of time savings, I might definitely prefer the walk through the park, because it is more pleasant. What the example shows is that from the fact that a person is not indifferent between two options, one cannot immediately infer that there is some ultimate scale of value on which the chosen option ranks higher. Having a reason does not mean ranking the chosen option higher on some ultimate scale of value.[4]

It has been claimed that whenever we judge that one course of action is better than another *all things considered* and our judgment is subject to evaluation as correct or incorrect, then even if the alternatives have opposite ranks on relevant but incomparable scales of value, there must be some *more comprehensive* value that detemines which choice is correct (see Chang 1997, 2003). This might be true, given suitable definitions of the terms. Suppose we define a *judgment-all-things-considered* as a judgment that is made by ranking the options on a scale. And suppose that to deem such a judgment-all-things-considered *correct* is to regard the ranking on which it is based as correct. Then the claim at issue will be true. But then the fact will remain that we regularly make decisions that are not judgments all things considered in this sense and which are not governed by any more comprehensive value that allows us to rank the options on a single scale of value.

4 For an exposition of some of the deliberations that might lead one to a choice even in the face of incomparable evaluations, see Morton 1991.

By contemplating the choices I do make, I might learn something about the kind of person I am—whether, for instance, I am the kind of person who lets pleasure outweigh efficiency or not. By studying my past behavior in such situations I might even obtain some measure of how the two scales of value compare with one another in my decision-making. I might discover that when it is a choice between saving time and experiencing pleasure, if option *A* is *n* pleasure units more pleasurable than option *B*, then I will choose *B* only if *B* represents a time-savings of at least *t* minutes over *A*.[5] But even if I can draw this kind of conclusion, and even if I discover that, by this measure, the route up Clifton Avenue is preferable to the route through the park, it does not follow that I *should* take the route up Clifton Avenue. Today I might be in a different mood and require a higher level of time savings for every unit of pleasure I sacrifice, and there may be no reason why I should not require that today.

In acknowledging that one might in principle discover a person's rate of exchange between pleasure and time savings, and, more generally, that one might discover a person's rate of exchange between any two "incomparable" scales of value that do have a place in a person's decision-making, I have acknowledged that human behavior might be to that extent predictable. By studying a person's past choices, we might get a handle on how his or her choices relate to the qualities he or she perceives in the options between which he or she has to choose. And on that basis we might be able to explain and predict some choices. But that in itself offers no hope for the belief-desire law. These generalizations may pertain only to a given individual. Different people will exchange time savings for pleasure at different rates. Moreover, a given individual's rate of exchange may change over time. There is no evidence here for an objectively correct method of converting the two scales into a single scale that we can apply to every rational agent at all times.

The nature of the challenge posed by incomparable scales of value may be best brought out by considering a case where my choice is clear. Suppose it is my duty to administer an exam to my 9 o'clock class. I have

5 We cannot assume that the rate of exchange between two value scales is constant across the entire range of values on both scales. Nonetheless, if indifference curves can be plotted in the two-dimensional space of points defined by the two scales (such that the agent is indifferent between any two points on any given indifference curve), then at each point a rate of conversion can be defined. See Keeney and Raiffa, 1976, p. 83. But if a third dimension of value is added to the first two, we may find that the rate of exchange between the first two varies with the value of the third scale.

announced the exam in advance, and all the students will be there expecting it. I have the copies of the exam on my desk in my office. The classroom is directly across from my office. I can get to school, get the exams out of my office and get to my classroom in 16 minutes if I walk straight up Clifton Avenue at a brisk pace. If I go through the park, then it will take me 22 minutes at least and I will be late. My choice is clear: I take the straight route. I desire to get to class on time. I believe that the best way to get to class on time is to take the straight route. And yet one cannot maintain that rationality dictates that I must take the straight route. Suppose we add to the description of the case, without taking away anything I have already said about it, that a walk through the park would be *in some respects* preferable to me. In that case, I might decide that it is best to take that walk and start the exam late, or postpone the exam, or cancel class altogether. That is not what *I* would do in fact, and if someone were to behave like that, I would blame him for failing to meet his obligations, but could I accuse him of irrationality?

What I mean to imply with this rhetorical question is that we can never just look at some segment of an agent's beliefs and desires and on that basis conclude that only a certain action or a narrow range of alternatives qualifies as rational. If we look a little beyond that segment we will inevitably find additional, incomparable scales of value on which the options are differently ranked. And as soon as we find that two options are ranked differently on incomparable scales, no universal rule of decision-making decides between them. If we consider only some segment of a person's beliefs and desires, rationality does not dictate any particular course of action, because we cannot legitimately exclude the beliefs and desires that fall outside of that segment. But if we consider the sum total of a person's beliefs and desires, then again no rule of reason dictates a course of action, because within that sum total we will inevitably find that different options are ranked differently on incomparable scales of value.

What then of those cases where it seems we can make our decision on the basis of a rule of decision-making such as a principle of maximizing expected *monetary value*? We can do that, in fact, but only if we have somehow first determined that monetary value is the only relevant measure of value. Similarly, if our problem is just to choose a video that will be entertaining, then we can simply go with the one that we expect to be the most entertaining. These are all cases in which, by some other means, we have decided that all other dimensions of value can be ignored. But decision theory does not provide any resources for paring down the relevant dimensions of value; and so even in these

cases decision theory does not offer any method for making a decision all things considered.

These conclusions do not endorse any kind of nihilism about reason. I am not saying that there is no distinction between rationality and irrationality. For instance, I have not questioned the correctness of the Von Neumann-Morgenstern rationality conditions. (Actually, some of the conditions can be questioned. The Ellsberg Paradox (Ellsberg 1988 [1961]) is particularly troubling. But that is another matter.) Nor am I even saying that rationality is merely a matter of consistency, in some broad sense, among beliefs and desires. Actions too can be criticized as irrational. What I am saying is that we cannot conceive of the rationality of action as conformity to some all-purpose rule. So we cannot expect there to be a rule that we can appeal to in formulating a *ceteris paribus* law relating beliefs and desires to actions. An action is rational if it survives the process of criticism. I am not able to give a good theoretical account of that process, but I think I have adequately shown that it is not a matter of laying down some principle that selects an action on the basis of a person's beliefs and desires and checking whether a person conforms to it.

7 Conclusion

The conclusion I draw from this survey of alternatives is that there is presently no reason to believe that there is any kind of law relating beliefs and desires to action that might satisfy the requirements of the epistemological or ontological rationales for looking for such laws. The challenge I have posed to those who believe that there must be such laws is to stop hiding under the cover of the excuse that "it's an empirical question" and to show us what the empirical question is by putting forward some candidates that we cannot reject out of hand just by thinking about them.[6]

Christopher Gauker
Department of Philosophy
University of Cincinnati
P. O. Box 210374
Cincinnati, OH 45221-0374
USA
christopher.gauker@uc.edu

6 An earlier version of this paper was presented at the Society for Philosophy and Psychology, Edmonton, Canada, June, 2002. Thanks to Georges Rey for his commentary on that occasion.

References

Ayer, A.J., 1970: "Man as a Subject for Science," in his *Metaphysics and Common Sense*, Cooper, Freemand and Co., pp. 219–239.

Broome, John, 1991: "'Utility'," *Economics and Philosophy* 7, 1–12.

Chang, Ruth, 1997: "Introduction" in Ruth Chang, ed., *Incommensurability, Incomparability and Practical Reason*, Harvard University Press, pp. 1–34.

Chang, Ruth, 2004: "'All Things Considered'," in John Hawthorne, ed., *Philosophical Perspectives*, Vol. 18: *Ethics*, Blackwell: 1–22.

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

Dennett, Daniel, 1971: "Intentional Systems," *Journal of Philosophy* 87, 279–328.

Dennett, Daniel, 1987: *The Intentional Stance*, MIT Press.

Ellsberg, Daniel, 1988 [1961]: "Risk, Ambiguity and the Savage Axioms," in Peter Gärdenfors and Nils Sahlin, eds., *Decision, Probability and Utility*, Cambridge University Press, pp. 245–269.

Fodor, Jerry, 1975: *The Language of Thought*, Harvard University Press.

Fodor, Jerry, 1991: "You Can Fool Some of the People All of the Time, Other Things Being Equal: Hedged Laws and Psychological Explanation," *Mind* 100, 19–34.

Grice, Paul, 1989: *Studies in the Way of Words*, Harvard University Press.

Hampton, Jean, 1994: "The Failure of Expected Utility Theory as a Theory of Reason," *Economics and Philosophy* 10, 195–242.

Horgan, Terence, and James Woodward, 1985: "Folk Psychology is Here to Stay," *Philosophical Review* 94, 197–226.

Jeffrey, Richard C., 1983: *The Logic of Decision*, 2nd edition, University of Chicago Press.

Keeney, Ralph L. and Howard Raiffa, 1974: *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*, John Wiley and Sons.

Lewis, David, 1970: "How to Define Theoretical Terms," *Journal of Philosophy* 67, pp. 427–446.

Lewis, David, 1972: "Psychophysical and Theoretical Identifications," *The Australasian Journal of Philosophy* 50, 249–258.

Lewis, David, 1974: "Radical Interpretation," *Synthese* 27, 331–344.

Loar, Brian, 1981: *Mind and Meaning*, Cambridge University Press.

Luce, R. Duncan, and Howard Raiffa, 1957: *Games and Decisions*, John Wiley and Sons.

Morton, Adam, 1991: *Disasters And Dilemmas: Strategies For Real-Life Decision Making*, Blackwell.

Newell, Alan, 1990: *Unified Theories of Cognition*, Harvard University Press.

Pettit, Philip, 1991: "Decision Theory and Folk Psychology," in Michael Bacharach and Susan Hurley, eds., *Foundations of Decision Theory*, Basil Blackwell, pp. 147–175.

Pietroski, Paul, 2000: *Causing Actions*, Oxford University Press.

Pietroski, Paul and Georges Rey, 1995: "When Other Things are Not Equal: Saving *Ceteris Paribus* Laws from Vacuity," *British Journal for the Philosophy of Science* 46, 81–110.

Resnick, Michael. D., 1987: *Choices: An Introduction to Decision Theory*, University of Minnesota Press.

Rey, Georges, 1997: *Contemporary Philosophy of Mind*, Blackwell.

Schick, Frederic, 1997: *Making Choices: A Recasting of Decision Theory*, Cambridge University Press.