Hypothetical Motivation

Donald C. Hubin

Noûs, Vol. 30, No. 1 (Mar., 1996), 31-54.

Stable URL:
http://links.jstor.org/sici?sici=0029-4624%28199603%2930%3A1%3C31%3AHM%3E2.0.CO%3B2-E

Noûs is currently published by Blackwell Publishing.

NOÛS 30:1 (1996) 31–54

Hypothetical Motivation

DONALD C. HUBIN
Ohio State University

The view that one has reason to do whatever one is moved to do is too crude.[1] Sometimes our motivation would not hold up to a moment's reflection: appreciation of the cause of our motivation, its nature, or the real effects of acting in accordance with it may annihilate the motivation. Better to say, some have thought, that we have reason to act as we would be moved to act if certain conditions were met: if we were vividly aware of all the relevant facts, capable of appreciating certain features of our action, thinking clearly about the situation, or whatever. This view, which I will dub 'the hypothetical motivation theory of reasons', is shared by a number of theorists having little else in common.[2]

Adherents of the hypothetical motivation theory disagree on the specific characterization of the idealized conditions and what must be true in these hypothetical circumstances in order for an agent to have a reason to act. There is at least near unanimity that false beliefs need to be purged—typically, supplanted with true ones.[3] Frequently, the idealization includes the assumption that the agent knows all "relevant facts."[4] This may not be enough for some people's tastes, though. While it may avoid the result that people have reasons based on ignorance and error, it doesn't preclude reasons based on insensitivity, selfishness or malice. Some would add the assumption that the idealized situation be one in which the agent is "vividly aware" of the relevant facts. This may help with insensitivity, but probably does nothing for selfishness and may exacerbate the problem with malice. (If one seeks the suffering of others, vivid awareness that an action produces this result excites, rather than extinguishes, desire.) The idea that one could, if constituted in a suitably malevolent way, have harm to others be a reason for acting seems all too close to the Humean view[5] to be attractive to some defenders of hypothetical motivation theories. So, some

proponents of these theories include in the idealization assumptions about the conative states or motivational dispositions of the agent.

There is another line along which idealizations might be demanded. One might think that an agent has reason to perform those actions he would be motivated to perform were his preferences coherent.[6] An individual with incoherent preferences, one might charge, cannot be given coherent advice, at least in certain hypothetical choice situations. Furthermore, since it will not be possible to define, for him, a utility function, it will sometimes not be possible to say what action maximizes his utility. If one holds that reasons and rationality are dependent on the notion of individual utility (either its maximization or something less), it might be plausible to conclude that an agent without preferences from which we can infer a utility function does not have reasons for acting.

Other conditions could be imposed. It could be held that one has reason to perform those actions one would desire to perform were one's set of desires non-arbitrary in certain specified ways. A defender of this sort of view might argue that while there may be no formal incoherence in a person's preference pattern, it may still be that his preferences depend on arbitrary and indefensible distinctions. One might, for example, hold that a preference for larger, rather than smaller, income is non-arbitrary, and that having income level be subject to diminishing marginal utility at any of a variety of rates is non-arbitrary as well. But suppose a person assigned utilities to income levels in much the normal way *except* that she assigned extremely low utility to all income levels that are equivalent to a prime number of 1990 U.S. dollars. And suppose that she did this for no further reason. One might hold her preference pattern to be arbitrary and, hence, not to have reason-giving force.[7] Does she really have a significantly diminished reason to accept a salary of n dollars when n is prime than she does to accept a salary of $n-1$ or $n+1$ dollars? Perhaps what she has reason to do depends on what she would be motivated to do were her preferences not arbitrary in the way that they are.

There is, then, considerable room for disagreement about the idealizing conditions to be employed in the hypothetical motivation theory of reasons for acting. There is room, too, for disagreement about what must be true in the hypothetical circumstances in order for there to be a reason. Many, though not all, hypothetical motivation theories of reasons tie the existence of reasons to hypothetical desires.[8] What one has reason to do is what one would, in some idealized situation, *desire* to do. But one might think that the appropriate motivating state is not desire at all. Perhaps it is the state of valuing or caring about something. Or, if one holds that cognitive states—perhaps certain sorts of beliefs—can motivate, one might hold a variant of the theory claiming that what one has reason to do is what one would, in some hypothetical situation, believe best.[9]

These differences, and others that could be identified, are, needless to say, of enormous importance; still, it is possible, and I think desirable, to consider the hypothetical motivation theory of reasons for acting at a level of abstraction that ignores these distinctions. Hypothetical motivation theories have the following in common: they hold that one has a reason to perform an act, *a*, in some actual situation, *s*, because one would, in some hypothetical circumstance, *h*, be motivated to do *a* in *s*. (We are using the notion of motivation to refer to whatever motivating psychological state[s] the theory in question takes to be relevant to the existence of reasons.) The specification of the idealized circumstances might be extremely difficult and, once they are specified, there might be overwhelming epistemic difficulties in applying the analysis. Still, one might think, we understand the view in broad outline and, at least at that remove, it is exactly right.

Not so. The hypothetical motivation theory is flawed, I will argue, not in its details but in its fundamental approach. The problem is not that it is impossible to construct some claim about the agent's motivation under some hypothetical circumstances in such a way that the claim will be true if, and only if, the agent has a reason to act as he would be motivated to act in the hypothetical circumstances. Specify the circumstances carefully enough, and this can be done.[10] Neither is the question whether the subjunctive conditionals proposed by the hypothetical motivation theory can be illuminating in deciding whether an agent has a reason to act. They certainly can be. The question concerns whether the subjunctive conditional constitutes a philosophical analysis of the notion of having a reason to act. Is the existence of a reason for acting *explained* by the truth of one of the subjunctive conditionals proposed by hypothetical motivation theories? Does the reason exist *because* of the truth of one of these subjunctive conditionals?

I will argue for a negative answer to these questions. Instead of the hypothetical motivation theory of reasons, I propose an "actual intrinsic motivation" account.[11] I do not give anything like a full exposition and defense of this position. Rather, I sketch it in sufficient detail (I hope) to render plausible the claim that this approach avoids the genuine problems with an actual motivation account without encountering those raised by the hypothetical motivation theory. Rather than fully developing the theory, I undertake the more modest task here of trying to show that it is a theory worth developing. First, though: the allure and inadequacy of the hypothetical motivation theory of reasons for acting.

The Attraction of the Hypothetical Motivation Theory

The hypothetical motivation theory of reasons is not without its attractions. Before criticizing the theory, we should appreciate these. The attractions

derive from both general methodological considerations and special features of the concept of rational advisability.

The question of what one has reason to do is an important substantive question that, if not directly connected with action (since we are all too capable of ignoring reasons), is at least connected with *rational* action. It is relevant to the assessment of our actions, and also of our characters insofar as our actions flow from our characters. As with most questions on which something of importance turns, there is great disagreement about just exactly when, and in virtue of what, one has reasons to perform an action.

Confronted with difficult controversy, it is always tempting to redescribe the problem in the hope that it will yield to an attack from a different direction. One popular sort of redescription involves the appeal to a hypothetical situation, process, or agent. So, we might say with Kant that what *is* moral duty for us, is simply what *would* be done naturally by morally ideal agents, or, with the hypothetical contractarians, that what *is* just, is what *would* be agreed to by us in some ideal choice situation.

One reason for the attractiveness of redescription in these cases is simply, as I have suggested, the hope that turning the problem sideways in order to attack it from the flank might produce insights that have been missed in the frontal assault. But there is another source of the attraction: redescriptions that involve hypothetical situations, processes or agents, seem to operationalize a concept in the sense of describing a procedure that could, in principle, be carried out to verify the applicability of the concept.[12] With certain (dangerous) assumptions, the operationalization may even appear to be a practical one.[13] With respect to hypothetical motivation theories of reasons, we may, for example, believe that to the degree that we approach the idealization, the actual motivation of the agent is a good indication of what reasons the agent has.[14]

These are general considerations. There are also attractions of the hypothetical motivation theory of reasons that are more closely associated with the specific problem the theory addresses. The crude view that one has reason to do what one is motivated to do attracts us, to the extent that it does, because it connects reasons with an agent's motivational structure. Reasons of the sort we are discussing—those relevant to rational advisability—ought not to be "shruggable." When it is pointed out that such a reason exists, a response of "I don't care" should manifest confusion or some noncontroversial sort of irrationality, perhaps incoherence. Unshruggability is the attraction of connecting reasons to an agent's motivational structure. But the crude theory makes the wrong connection. The crude theory must say that the agent has no reason to perform an action that would produce a result she is most strongly motivated to bring about if the agent, through ignorance or error, has no motivation to perform the action. Worse yet, one may be

motivated to bring about a result, believe that performing a given action will produce this result but simply fail to be motivated to perform the action. The crude theory will say, in this case, that the agent has no reason to act. We will not be able to rationally criticize behavior that seems manifestly irrational or to recommend behavior contrary to actual motivation.

Part of the attraction of the hypothetical motivation theory is based on these inadequacies of the crude view, and the apparent ability of the appeal to hypothetical motivation under idealized circumstances to remedy them. The problem might be conceptualized as follows: our motivation is sometimes inappropriate because we are not in an ideal situation for judging what we ought to do. Perhaps we are ignorant of, or insensitive to, some important consideration. We might, then, seek counsel from a better judge—and who, one might think, could be better than ourselves, freed of that which encumbers our judgment.[15] This remedy for the problems of the actual motivation theory does not appear *ad hoc*. The most obvious failings of the crude view—those in which the actual motivation is dependent on false beliefs—are addressed by the hypothetical motivation theory in an intuitively plausible and satisfying way. Indeed, the entirely natural way in which I introduced the problem with an "actual motivation" theory at the beginning of this essay seemed to incline toward the hypothetical motivation approach: I pointed out that some of our actual motivation *would not* hold up under reflection. In light of this, it is hardly plausible to deny that we find the appeal to hypothetical motivation a useful heuristic in talking about what reasons a person has to act.

But the theory we are challenging holds that hypothetical motivation is more than this. It offers hypothetical motivation as a philosophical analysis of the concept of a reason for acting. Some concepts, dispositions for example, seem to be essentially hypothetical in nature. Hypothetical motivation theories hold that the concept of a reason for acting, like the that of a disposition, is such a concept. Indeed, we can think of hypothetical motivation theories of practical reasons as holding that reasons for acting are based on certain motivational dispositions of agents. (I shall argue later that reasons are, in fact, based on *actual* motivation—though of a certain sort.)

Even if, as I argue, reasons for acting are not to be analyzed in terms of hypothetical motivation, we should not conclude that hypothetical motivation theories are without value. If we can establish the biconditional relation between the existence of reasons and some sort of hypothetical motivation, this might be illuminating even if not directly explanatory. The hopes of making progress on the problem of practical reasons by "turning the problem sideways" may be warranted; looking at the problem as the theory suggests may bear philosophical fruits. But I think we must be wary: there is

always the danger of satisfying ourselves that we are making progress when we are just going 'round—simply substituting one set of obscurities for another or, worse yet, presenting the old obscurities in new guise.

Evaluating Hypothetical Motivation Theories

Hypothetical motivation theories of reasons for action have it that the existence of reasons depends on the truth of certain counterfactuals. There is significant philosophical controversy about how counterfactuals are to be understood. But I do not mean to make anything of the difficulties arising from the analysis of counterfactuals, themselves—difficulties that will infect all counterfactual analyses. In particular, I will assume that there is some correct analysis of counterfactuals which entails that those counterfactuals which seem obviously true are, in fact, true.

The idea behind hypothetical motivation accounts of reasons is that reasons are based on some sort of disposition: a disposition to be motivated to act (or choose) under certain conditions. It is this disposition that the subjunctive conditional defining the hypothetical motivation theory is supposed to capture. We need to inquire into the subject of the disposition and how the relevant subjunctive conditional is to be evaluated.

Perhaps the subject is the human organism and subjunctives are not significantly different from those a psychiatrist might consider in deciding on a course of treatment for a patient. The psychiatrist will ask herself how the patient would react to various possible treatments. The answers she seeks will have the form of subjunctive conditionals appropriate for deciding between courses of action. ("If I were to use Freudian therapy alone, the patient would have an x% chance of significant improvement in his depression, . . .")

Suppose, now, that the patient has an unusual brain chemistry such that the use of Freudian therapy would, in fact, increase the severity of the patient's depression. This may be simply because the sorts of insights gained under Freudian analysis happen to consist, in this agent, of brain states that would produce a type of chemical reaction in the brain that results in increased depression. The causal connection may not be psychological in nature; it might be purely physiological. The doctor will, nonetheless, want to assess her actions in the awareness of this increased danger, for she is concerned with how *this* flesh-and-blood patient would respond to the treatment. The fact that other patients or even this patient (if we could block the unfortunate physiological side-effect of the Freudian-induced awareness) would, as a result of increased understanding, become less depressed, is not relevant.[16]

Is the defender of a hypothetical motivation theory interested in dispositions of the organism? I think not. Let us suppose that the idealizing condi-

tions the theory proposes are these: removal of false beliefs, introduction of all relevant true beliefs, and vivid appreciation of all relevant factors. Now we are to ask what a real flesh-and-blood agent *would* be motivated to do were he to be free of false beliefs and vividly aware of the relevant facts, *etc.* In possible worlds parlance, we are to ask what this human being (or his counterpart) is motivated to do in the closest possible world in which he is free of false beliefs, vividly aware, . . . *etc.*[17] But what a given human being would be motivated to do under non-actual circumstances (as well as under actual circumstances) is dependent in part on brute physiological features of the agent. This particular agent, even when idealized in the ways we are imagining, may become motivated to put pebbles in his navel when he becomes vividly aware of some complex of facts about number theory. The awareness of the set of facts may produce physiological effects on the agent—effects that have nothing to do with the content of the facts—and these effects may produce other psychological states: desires, or some other psychological states taken to be motivating.

Indeed, the sort of problem we have just raised is not confined to *idiosyncratic* physiologies. The same problem would arise were it a general physiological truth that awareness of certain facts set up a chemical response in the brain that produced some motivating state *via* a brute physiological connection. The defender of the hypothetical motivation theory of reasons is not interested in the "flesh-and-blood" dispositions of the human being; she is not concerned with the dispositions of the *organism*. She is, rather, concerned only with mental dispositions that have a mental etiology.[18]

We have focused on hypothetical circumstances that involve removal of false beliefs and vivid awareness of relevant truths. This was for expository convenience only. The argument that hypothetical motivation theories are not attempting to base reasons for acting on some disposition of the organism doesn't depend on this. Suppose that the theory holds that one has a reason to act if one would be motivated to act were one's preferences to be made coherent. The process one employs to render one's preference pattern coherent (and perhaps the mere possession of the coherent preference pattern one would have as a result of this process) may have purely physiological effects on one's motivation. And the same point can be made of other versions of the hypothetical motivation theory.

The general point can be illustrated by considering an analogy—one that will be useful in clarifying some points yet to be discussed. Imagine that Impiety Products, Inc., a computer chip manufacturer, has brought out an extremely powerful math processing chip, dubbed "The Beast." It functions without mathematical error so long as it never hits a number that would be represented in base 10 as having a sequence of three '6's in a row. When confronted with a calculation that involves such a number, though, the chip literally short-circuits and puts out a random number as the an-

swer. In this case, we clearly do not want to identify the correct answer with the answer that the actual chip would give were it to calculate the problem.

There is one very misleading aspect of the analogy. In the case of the Beast, there is an objective standard of correctness for the calculations; we can determine that the Beast is flawed merely by looking at the output. If we have recourse to a similar standard for a theory of reasons, then the hypothetical motivation theory offers us nothing but a heuristic.[19] So let us suppose that we do not. Does denying such an independent outcome-related standard for reasons for acting force us to say that the only sense that can be made of the subjunctive conditional that defines that theory is the one capturing the dispositions of the human organism? I think not.

Consider, again, the Beast. It is true that it has the disposition to output random numbers when making certain calculations. But this is not a disposition of the algorithm that is being run. Understanding the logical design of the chip but knowing nothing about mathematical truths, we could determine that the chip is malfunctioning when it deals with numbers that cause the quirky behavior. ("The algorithm is willing but the silicone is weak.") Then, even without access to an *independent* standard of mathematical correctness, we might identify the correct answer to a problem with the one that *would* be produced by the correct application of the algorithm.[20]

Inspired by this analogy, we might view the person as the analog of the algorithm and speak of what *the person* would be motivated to do under the idealized conditions. We would then exclude as deviant and, thus, irrelevant, those counterfactuals based on purely physiological causal links between these conditions and the motivation of a human being. Suppose, then, that we understand the hypothetical circumstances as being ones in which the vivid awareness (or what-have-you) has no effects on motivation other than those it has *via* psychological connections based on the contents of mental states. Quite a hypothesis, to be sure, but it would avoid the problems for the analysis that are raised by "brute physiological processes."[21] It doesn't help with other problems, though.

An individual, Horatio—or Ratio, for short—might be extremely disturbed by the knowledge that the relation between the diameter and the circumference of a circle is not a rational number. Vivid appreciation of this fact, no matter what *else* Ratio is vividly aware of, may leave him without any motivation to eat. This is not due to any brute physiological phenomenon; it is the result of a psychological connection in Ratio between his appreciation of a proposition having a certain content and his motivation.

The hypothetical motivation theory would appear to have it that, even in the absence of any vivid awareness of the nature of *pi*, there is no reason for Ratio to eat. And this, despite the fact that he intrinsically desires to eat and intrinsically desires many things to which eating is a means.

Ratio is, of course, psychologically idiosyncratic. Perhaps problems with

the hypothetical motivation theory could be eliminated by requiring, as an idealizing assumption in the hypothetical situation, that the agent be psychologically normal.[22] This is a controversial correction, but even setting controversy aside, it is inadequate. Psychologically normal people might be such that vivid awareness of the Nazi atrocities in the concentration camps of World War II would kill their desire for food. Would the defender of a "vivid awareness" version of the hypothetical motivation theory want to say that the people of whom this is true have no reason to eat?

In order to avoid saying that the desire the agent currently has (in the absence of vivid awareness of the horrors of concentration camps) fails to provide him with a reason to eat the appealing food, we might declare the Nazi atrocities irrelevant to the act of eating. But, of course, part of what the hypothetical motivation theory seeks to do is to tell us what facts are relevant to our reasons for acting by asking us to consider how awareness of these facts would affect the motivations of the agent under idealized circumstances. If we have an independent test of relevance, it seems reasonable to ask why we lack such a test for the *way* in which the information is relevant. And if we have an answer to this latter question, we should suspect that, once again, the hypothetical motivation theory points to a mere heuristic.

Perhaps we have drawn the wrong lesson from the story of the Beast. Suppose that the Beast, when confronted with a computation involving a troublesome number, doesn't merely put out a number that has *no* mathematical relation to the correct answer. Instead, it follows an algorithm that is different from the one normally followed. This algorithm produces an output that, while not correct, *is* mathematically related to the correct answer. Even without appeal to an independent standard of correctness, we might distinguish these two algorithms and hold the Beast to be functioning (in some sense) correctly when it follows one and incorrectly when it follows the other.

To return to the case that interests us, we might, instead of requiring merely that the motivational change not be produced only through brute physiological processes, require that it be produced by a *nondeviant* psychological process. I shall not stalk this strategy here. As it stands, it is a promise, and promises are notoriously difficult to refute. A successful defense of the approach would require a number of developments. First, it would demand a typology of psychological processes that would allow us to justify the claim that there are really two different sorts of processes going on. Though we described the Beast as carrying out two different algorithms, we could just as well, it seems, have described it as carrying out one algorithm that leads to one subroutine when no problematic numbers arise and another when a problematic number arises. The difficulty of justifying the employment of a specific level of generality in one's typology will be uncomfortably apparent to those who have worried about either of the

following problems: the theory of act description for a rule-based theory of morality, and the theory of process description for a reliable-process account of justified belief.

More importantly, cashing the promissory note would require us to justify the claim that one of the algorithms was deviant *without* appealing to any independent standard for the correctness of the outcome. It is not clear to me how to do this. In the absence of a concrete proposal or any refutation of the abstract possibility of defending such a proposal, it seems best to leave this possibility for the hypothetical motivation theorist to pursue if she sees a way to do so.

Many versions of hypothetical motivation theories face the following additional, I think devastating, problem: they cannot handle correctly a case in which an agent is consistently "means/ends irrational."[23] Such an agent has ends that he values intrinsically, recognizes that certain actions are necessary means to these ends but never feels any motivation to perform the acts in question. Those theories that merely correct cognitive deficiencies (by, for example, requiring vivid awareness of all relevant facts) will do nothing to rectify means/ends irrationality. As a result, they will be forced to say that the agent has no reason to perform any actions that are merely means to desired ends (as opposed to ends in themselves). This is clearly the wrong answer. A possible solution seems obvious, but it is a mirage. The apparent solution is to add to the idealizing conditions the requirement that the agent be means/ends rational. Now, we are to ask what this agent would be motivated to do were he vividly aware of all the relevant facts (for instance) *and* means/ends rational. This apparent solution is a mirage because it offers no guarantee that the means/ends rationality will be brought about in the "right" way. Means/ends rationality requires a kind of coherence between one's means and ends, but it gives no priority to either. So, the requirement that agents be means/ends rational does not guarantee that they would be motivated to perform those actions which are means to the ends they actually desire. If the imposition of the other idealizing conditions leave Ratio and the sensitive observer of Nazi atrocities without the motivation to eat, the imposition of a requirement of means/ends rationality does not guarantee that we will acquire the motivation to eat. Coherence can be achieved by adjusting the ends to means as well.

The most obvious way to deal with this problem is to hold fixed the agent's actual ends and force the agent to adjust the means to the ends on the hypothesis that the agent is means/ends rational. But this theory is a hypothetical motivation theory in name only. Since the agent is means/ends rational, fully informed, and so forth, and the ends are held fixed, the agent would be motivated to perform just those actions which are, in fact,

means to the actually desired ends. I think this is the right approach to the notion of rational advisability but it is merely a terminological variant of the actual intrinsic motivation theory I shall propose later. Hypothetical motivation plays no essential role.

One might, though, propose that the agent's actual ends not be held fixed—that the idealizing conditions be allowed to affect the ultimate ends the agent values. If such a theory is combined with an objective means/ends principle, the result is what I will call a "hypothetical ends theory."

Hypothetical Ends Theories

One of the problems we saw with the hypothetical motivation theory was that it seemed not to generate all of the reasons that in fact exist—there were "truant reasons." Ratio turned out to have no reason to eat because his numerical ponderings would destroy his motivation; a person's reason to eat may be undone by his sensitivity to human atrocities. There is a way around the truancy problem.[24] Both Ratio and I would, we may suppose, still have desires to which eating would be instrumental. Perhaps this fact ensures that Ratio and I have a reason to eat. Perhaps, while we would have no motivation to eat, we would have motivation to achieve some state to which eating is a means. We could, then, modify the hypothetical motivation theory of reasons to say the following: one has a reason to perform an act, a, in some actual situation, s, because one would, in some hypothetical circumstance, h, be motivated to bring about a state of affairs to which a is a means. The idea is that the ends that one would have in the idealized circumstances generate reasons to perform actions which are means to their achievement regardless of whether the motivation would be communicated to those means in the hypothetical circumstances.[25]

There is an initial attraction to this "hypothetical ends" view. While Ratio and I would, in the hypothetical circumstances, lack our motivation to eat, it is not because eating is not related to our motivational structure in a way that generates reasons for acting. It is because, given our psychological constitutions, we would not be motivated to perform those actions that we have reason to perform. Then, one might say, even in the actual circumstances Ratio and I have reason to eat. One might plausibly think, for example, that Ratio and I have reason to continue our lives based on the motivation we would have to do so in the idealized circumstances. Because reasons are communicated from ends to means, we have a reason to eat.

Even if this emendation were to work with the problem of truant reasons, it would not help with the opposite problem: reasons that are based on motivation that is inappropriately introduced by the hypothetical situation. We might call this the problem of "counterfeit reasons"—reasons the

theory improperly countenances. Motivation can, it seems, be inappropriately introduced as a result of either brute physiological processes or psychological ones.

It is not difficult to imagine a case in which some hypothetical process produces motivation as a result solely of a brute physiological chain of events. Depending on the physical entity in which psychological states are realized, it is, I suppose, possible for just about any psychological state to cause just about any other psychological state through brute physiological processes. Suppose that given my particular brain neurophysiology, becoming vividly aware of world hunger and human injustice would consist in a complex of neurophysiological states that would produce, through non-psychological connections, a strong motivation to beat my head against a brick wall. Surely, in this case I would still not have a reason to engage in such a destructive activity. Such a motivation would be a counterfeit, unjustifiably introduced by the hypothetical ends theory.

The problem is more difficult when we consider ends that one would have in the idealized circumstances as a result of psychological processes. The difficulty arises, I think, because the most common way in which a motivation might be provoked as a result of the sort of idealizing conditions typically imagined by hypothetical motivation or hypothetical ends theories is simply by the agent becoming aware that an action *is* a means to something he is intrinsically motivated to bring about. This can happen for one of at least two reasons: the agent may change his beliefs about the consequences of this or alternative actions; or, the agent may come to recognize what he genuinely cares about. But, if this is the only psychological process that produces motivation in the idealized conditions, there is no need to accept a hypothetical ends theory of reasons for acting. We could simply rely on an actual intrinsic motivation account. That is, we could hold that an agent has a reason to perform an action that is a means to some end he is actually motivated to pursue for its own sake.

The crucial question, then, is this: does an agent have a reason to perform some action because it is a means to some end that he would be intrinsically motivated to bring about under some idealized conditions where this motivation does not result from any intrinsic motivation the agent now has? Stephen Darwall offers a case that might lead one to think so (1983, pp. 39–42). As a result of watching a film portraying the plight of textile workers subject to poor working conditions, low pay, and environmental conditions that lead to "brown lung" disease, Roberta is motivated to political activity. While Darwall grants that such motivation *may* be dependent on some more general motivational feature of Roberta, say a general desire to combat injustice and alleviate suffering, he insists that it need not be. The motivation in question may be *created* by her vivid awareness of the suffering without her having any such pre-existing motivation,

and, one might think, this vivid awareness plausibly constitutes an appropriate idealizing condition for a hypothetical motivation theory.

We should agree with Darwall that there is some way in which we can, through psychological mechanisms, come to have motivation that is not based on any previously existing motivation.[26] The issue for us is whether we have, prior to acquiring the motivation, a reason to act to produce the end we would be motivated to pursue as a result of such a process. I think not. In the crucial case, remember, we may assume that Roberta is not one of those "who care about strangers, who care about evil and social injustice." The movie generates in her, absent any such general concerns, a motivation to alleviate the suffering of the textile workers.

Roberta has had a change of heart; she has become a different sort of person, a more caring, concerned person; she has undergone conversion, but it was not a rationally required conversion. I see no reason for us, or for Roberta, to think that prior to the conversion, she had a reason, in the relevant sense, to alleviate the suffering of the textile workers. Of course she has a *moral* reason to alleviate the suffering of the textile workers. And, furthermore, this reason does not depend on her actual desires. Perhaps part of our tendency to think that Roberta does have a reason (in the sense connected to rational advisability) to help the textile workers even prior to her change of heart is based on the recognition of this moral reason. But just as this moral reason doesn't depend on her actual desires, neither does it depend on the desires she would have after watching the movie. Suppose that Roberta were such that viewing the movie would lead her to desire to exploit workers herself. Would we still be inclined to say that even absent such actual desires, Roberta has a reason to exploit workers? Moral reasons are not generally dependent on an agent's motivation,[27] but we are concerned here not with moral advisability but with rational advisability. And provided it is correct to say of Roberta prior to her conversion that it was not *rationally* advisable, *ceteris paribus*, for her to work to end the exploitation, the hypothetical ends account of reasons offers no advantage over the intrinsic motivation theory.

The search for a solution to the truancy problem for the hypothetical motivation theory might lead one to the hypothetical ends theory, but both theories encounter the problem of counterfeit reasons. I propose, as a solution for both problems, that we reject the attempt to analyze or explain reasons in terms of hypothetical motivation at all. Better, I think, to understand reasons in terms of the actual ends of the agent.

Intrinsic Motivation Theory

Motivation founded on ignorance and error seems not to give us reasons for acting in the sense relevant to the rational advisability. The hypotheti-

cal motivation theory seeks to account for this by appealing to the "staying power" of this motivation in the face of idealized conditions. But, as we have seen, motivation may withstand (or be extinguished by) the idealization for the wrong reasons; the idealizing conditions may even create inappropriate motivation. Rather than grounding reasons on what an agent *would* be motivated to do in some hypothetical circumstance, it seems better to ground them on what he *should* be motivated to do. But, of course, if this 'should' is one of rational advisability, then we make no progress by endorsing this slogan. What *should* an agent be motivated to do?

The most satisfactory approach, I think, is to ground reasons for acting (in the sense of considerations in favor of the rational advisability of actions) on the *actual intrinsic* motivation of the agent. This view depends on conceiving of motivation as having a roughly hierarchical structure: some motivation is dependent, in familiar ways, on other motivation together with beliefs about the world.[28] Ultimately, though, we find motivation that is not grounded in other motivation in this way. This motivation is intrinsic motivation. If we conceive of the motivating state as desire, then the view is one that makes the familiar distinction between intrinsic and instrumental desire and holds that it is only the former that generates reasons (in the appropriate sense). These reasons are communicated from ends to means so that, in the sense relevant to rational advisability, we have reasons to perform those actions that promote states of affairs we intrinsically desire regardless of our actual or hypothetical instrumental desires.[29] (But the motivating state need not be desire, of course.)

Such an actual intrinsic motivation view avoids those implications of an actual motivation theory that are clearly unacceptable. Suppose Winslow is motivated to engage in weight training only because he believes that doing so will win him the love of the fair Winnifred, which he is motivated to achieve for no other end. But Winslow is quite wrong in his belief; Winnifred's love turns on no such thing. His actual motivation to engage in weight training does not give him any reason to do so—not even a *pro tanto* one.[30] It doesn't matter whether this motivation *would* withstand Winslow's vivid awareness that weight training is not a means to his end. On the other hand, it may turn out that Winslow is motivated to achieve other states of affairs, say the preservation of his health, which would be promoted by his weight training. If so, then he does have a *pro tanto* reason to weight train. (Absent countervailing considerations, it is rationally advisable for him to weight train.) Again, this does not depend on whether he *would* be motivated to weight train by the vivid awareness of the connection between doing so and his intrinsic motivation.

So much for cases of ignorance and mistaken belief. What of cases of insensitivity? They should, I propose, be handled analogously. Ignorance

and mistaken belief lead to inadvisable action when the agent is led, by those defects, to act in ways that frustrate his intrinsic values. Insensitivity can have the same effect. Suppose that what I value most is the well-being of my children. I may still, through insensitivity (not ignorance), fail to *appreciate* the effect that my actions might have on them and, because of some momentarily strong desire, I may act in ways that are detrimental to what I value most. In this way, inadvisability may be born of insensitivity. While it may be true that a more sensitive appreciation of the effects of my actions will (typically) improve my choices as instruments for the realization of my intrinsic values, we are not thereby led to adopt a hypothetical motivation theory of reasons for acting. We need simply to evaluate actions as means to our intrinsic values.

Perhaps, it might be said, our intrinsic values reflect insensitivity. In what sense could this be true? I can think of only this sense: were we to reflect on them in a sensitive way, we would alter them. But notice that this cannot be because of any more basic values upon which these are based. The sensitive appreciation of the facts connected with our intrinsic values might cause those values to change. But I do not see how this change could be judged a rational improvement in the absence of some rationally mandated independent standard of intrinsic value. And, if we have justified such a theory, we hardly need recourse to hypothetical motivation as the basis of a theory of rational advisability.

I believe that the same sort of response is appropriate for all of the "defects" that are supposed to be corrected by the idealizations demanded by hypothetical motivation theories. These can be understood as defects insofar as they are likely to interfere with the promotion of that which the agent is intrinsically motivated to bring about. But, if this is the only sense in which they are defects, the intrinsic motivation theory is not embarrassed by these defects, for it recommends actions in virtue of their connection with those states we are *intrinsically* motivated to bring about, not in virtue of the particular "derived" motivation we may have as a result of some defect. If, on the other hand, we have earned the right to consider actual, "derived" motivation defective in some *other* sense, what need do we have of the notion of hypothetical motivation? We can appeal directly to the normative theory we employed to declare the actual motivation defective.

Intrinsic Motivation: Counterfeits and Truants

Intrinsic motivation theories of reasons for acting are superior to hypothetical motivation theories, I have argued; they avoid the problems of actual motivation theories without incurring those of the hypothetical motivation theories. But we might be leery of this last claim. There are reasons for

thinking that actual intrinsic motivation accounts suffer from their own problems with truant and counterfeit reasons. There are two problems I intend to address here. The first concerns motivation that is too fleeting and disconnected from the agent's other motivation to serve as a plausible ground for rational advisability. The second seems potentially more troubling, for it seems to indicate that the notion of intrinsic motivation cannot even be understood without introducing the very problems that led us to reject hypothetical motivation accounts of reasons for acting.

Ephemeral Motivation: Does the actual intrinsic motivation theory have its own problem with "counterfeit reasons"? Does it entail the existence of reasons that we should deny exist? It certainly appears so. Suppose that in the heat of my racquetball game, my consciousness is filled with a momentary motivation to physically harm my opponent, who is also my dearest friend. This momentary motivation fades quickly and I return to my normal sportsman-like self. Do I, for that instant, have a reason to harm my opponent (and dearest friend)? Such a thought has never crossed my mind before, I care deeply about my friend and all my other actions and intentions show this concern. Still, for a fleeting moment, I am motivated to harm him. Call this the problem of "ephemeral motivation":[31] How does the actual intrinsic motivation account prevent the generation of reasons based on fleeting motivation, unconnected to any of the agent's other motivation?

Whether there is a reason based on such motivation depends, I think, on what psychological state the specific actual intrinsic motivation theory holds to be relevant. Because a solution to the problem of ephemeral motivation requires a defense of a specific version of the actual intrinsic motivation theory, I do not propose to offer a solution here.[32] Rather, I will merely suggest a version of the theory according to which it is not at all obvious that ephemeral motivation, of the relevant sort, exists. If I am correct, then it is not at all obvious that actual intrinsic motivation theories are saddled with this problem.

First, though, it is worth pointing out that it is not entirely clear that we should be too troubled by the problem of ephemeral motivation. Provided the ephemeral motivation does not conflict in any way with other motivation that we have, what is the problem with allowing it efficacy in producing reasons for acting? And, if it does conflict with other motivation, we can blunt the implications of putatively counterfeit reasons by meeting them with *other* reasons based on this other motivation. Suppose, for a moment, that we take desire to be the motivating state. Now, it may well be that I have had, on rare and fleeting moments, the desire to slap my daughter across her mouth. While there may be ample provocation for this desire, I may not desire slapping my daughter as a *means* to anything whatsoever. I may not believe that slapping her will prevent the sort of

behavior that provoked the desire, teach her a lesson, or have any other consequence I seek. I may know, even as I have the desire, that slapping her will be counter-productive. The desire may be, in our sense, an intrinsic desire. So, according to the version of the theory we are considering right now, I have a reason to slap my daughter. But if doing so conflicts with other strong desires I have—say, to avoid inflicting any form of corporal punishment on my children—then perhaps there is no problem admitting that I have *a* (*pro tanto*) reason to slap my daughter. It is, of course, overridden, perhaps even "swamped"—so resoundingly overridden that it is easily overlooked. If, on the other hand, the ephemeral desire is one to tickle the bottom of my daughter's foot when she is watching television and doing so would not conflict with any other desires I have, it is not clear that I don't have a reason to act on this desire—ephemeral as it is.

But there may be ephemeral desires that we don't want to say generate even overridden reasons for acting. They may be desires that are so out of character, so poorly integrated with the individual's ongoing concerns and interests, that we view them as assaulting the agent, rather than expressing his nature. One might, plausibly, think that these ephemeral desires should not only not *determine* rational advisability, they should not even influence it. I am sympathetic to this claim, but I think the solution is not to flee from actual intrinsic motivation theories of reasons but to reject intrinsic *desire* as the motivational basis of reasons for acting.

Consider, instead of an account of reasons for acting based on intrinsic desire, one based on what the person intrinsically *values*. At least the most obvious problems with ephemeral motivation seem to disappear. The acts of harming my racquetball partner or slapping my daughter, both of which I fleetingly desire, are nothing that I intrinsically value nor need they produce anything that I intrinsically value.[33] The values we hold are, I think, *essentially* non-ephemeral and are necessarily expressions of our selves. While it may make sense to think of a desire as being something that "assaults" one, one cannot be "assaulted" by one's values. Values are too closely identified with the self.[34]

Understanding Intrinsic Motivation: Suppose, then, that some appropriate motivating state can be specified, one that does not raise a problem with ephemeral motivation. What of motivation that is temporarily *not* present? Does the actual intrinsic motivation theory have a problem of truant reasons because of its reliance on *actual* intrinsic motivation. After all, in the middle of a hard-fought racquetball game, I may not have a thought for the well-being of my children, but my reason to promote their happiness survives the heat of competition. A moment of empty-mindedness does not mark the absence of all practical reasons. And I may well have good reason to cease my day-dreaming and concentrate on my upcoming lecture even though the ends fostered by doing so couldn't be further from my mind.

Of course, taking the motivating state to be something like intrinsic valuing rather than intrinsic desire, seems to help with this problem as well as with the problem of ephemeral desires. Surely, even when I am struggling to return a "kill shot", it is still true of me that I value the well-being of my children. This seems true because 'valuing' naturally receives a reading that is consistent with the agent having nothing related to it "present to mind." But, if we are willing to admit the distinction between occurrent and non-occurrent motivational states other than valuing (as we should), then other versions of the actual intrinsic motivation theory could make a similar move. For example, an actual intrinsic *desire* theorist could hold that all through the racquetball game I desire my children's happiness, though this desire is non-occurrent. It would be a mistake, the actual intrinsic motivation theorist must say, to suppose that reasons are generated only by intrinsic motivation that is "present to mind." To formulate the intrinsic motivation theory so as to require occurrent motivation in this sense would create a very serious problem of truant desires. So, we must recognize non-occurrent motivation. But how are we to understand this notion?

Non-occurrent motivation is frequently explained as a certain disposition toward occurrent motivation.[35] Unfortunately, this understanding creates a problem for the actual intrinsic motivation theory of reasons for acting. The problem is this: dispositional qualities seem to be cashed out naturally by the very sort of subjunctive conditionals that caused problems for the hypothetical motivation theory of reasons for acting. If non-occurrent motivation (desires, values, *etc.*) is just that motivation one would have if one were to reflect on the matter (or be vividly aware of the relevant facts or what-have-you), then it seems the realization of the hypothetical conditions could cause occurrent motivation when there is *no* motivation actually present.[36] This would raise a problem of counterfeit reasons. Conversely, there seems to be no set of conditions such that it is impossible for someone to be non-occurrently, intrinsically motivated to bring something about but not be disposed to be occurrently motivated under those conditions to do so. Understanding non-occurrent intrinsic motivation in this simple dispositional way raises the problem of truant reasons as well as that of counterfeit reasons—and on just the grounds that these problems arose for hypothetical motivation theories.

The problem confronting us here may tempt some to backtrack—reconsidering the arguments presented earlier against the hypothetical motivation theory of reasons for acting because of their implications in the present case. I have no such temptation. Whatever we say about hypothetical motivation theories of reasons, we will, in any event, have to distinguish two sorts of cases: that in which a person has a disposition to acquire the occurrent motivating state under certain hypothetical conditions but does

not currently have the motivating state even non-occurrently, and that in which a person has the motivating state non-occurrently and has the disposition to have it occurrently. Let us suppose that the motivating state is that of *valuing*. It is quite possible for a person to occurrently intrinsically value some state of affairs at one time, later not to value that state of affairs occurrently *or otherwise*, and still later to occurrently intrinsically value the state of affairs once again. It may even be that such a person is disposed to occurrently intrinsically value the state of affairs under certain specifiable conditions—perhaps whenever he vividly imagines the state of affairs. But *this* disposition to occurrently intrinsically value a state of affairs needs to be distinguished from the case of the person who consistently values a state of affairs sometimes occurrently sometimes not. The simple dispositional account of intrinsic valuing cannot make this distinction adequately.

What *is* the difference between the two cases? What does it mean to say that a person has some intrinsic motivation when there is no occurrent indication of this: he is not conscious of it, he is not doing anything that counts as acting in accordance with the motivation, *etc.*? I think the solution lies in understanding non-occurrent intrinsic motivation as being a real psychological state of an agent. It is identified, as I think all psychological states must be, by its functional characteristics which include its causal connections with stimulus situations, behavioral episodes and other psychological states. These will be analyzed in terms of a complicated and indefinitely large set of subjunctive conditionals, no one of which will be determinative of the existence of non-occurrent intrinsic motivation. What we are after, really, is something like a "theory of the person" and, in particular, we are seeking the best hypothesis about the conative aspects of his psychology.

Return once again to the analogy of the computer chip. The Beast gives the wrong answer in certain cases. The original problem with the Beast was that it short-circuited and failed to follow the algorithm that it normally followed. But even correcting for this—for example, by determining the algorithm it was designed to follow and the answer that would result—was not enough. The problem could be in the algorithm as well as in the hardware that is designed to follow the algorithm. (This was the point of the story of Ratio.) What we want to do is to determine the problem the Beast is "trying" to solve and offer the correct solution. This is no simple matter of observing the answers the Beast gives, the answers it would give in hypothetical situations, or even the answers it would give under the special hypothesis that it followed the algorithm correctly—though, of course, these things will constitute evidence for or against theories about what problem the Beast is "trying" to solve.

Our actual intrinsic motivation, occurrent or not, indicates what we are "trying" to do (in the relevant sense). Unfortunately, I have nothing particu-

larly illuminating or novel to say about how we justify the hypothesis that a person values (for example) some state of affairs. Certainly, the fact that, were a person to reflect on the matter, she would say she values some state of affairs and, more importantly, would adjust her behavior and plans accordingly, is important confirming evidence that she values it. But there are many other pieces of evidence. That she always feels empty and unfulfilled when she achieves that state of affairs may lead us to question this conclusion. That this value is not at all integrated into her life, that it seems to lie behind none of her actions or reactions to situations may also lead us to reject the claim that she values the state of affairs. Her tendency to say, on reflection, that she values it and even her readjustment of her plans may well be a "self-presentational" effect. Then, we may be inclined to say that what she really values is seeing herself as the sort of person who values the state of affairs, though she does not, in fact, value it.

The simple dispositional account of intrinsic values makes the mistake of identifying the nature of intrinsic valuing with one important piece of evidence for its existence. A better understanding takes the notion of intrinsic valuing to be functionally (and, therefore, dispositionally) characterized. But it does not identify the state of intrinsic valuing with any one specific (relatively simple) disposition.[37] This "better" understanding has obvious drawbacks. For one, we may seem considerably less capable of saying when a person actually values a state of affairs. Our problem will not merely be not knowing what the agent would do in some unobserved situation; even if, *per impossibile*, we knew all there is to know about her behavioral dispositions, we might still find ourselves quite unable to determine whether she values a certain state of affairs. She may have some dispositions that favor the hypothesis that she values it, and others that weigh in against this hypothesis. While we could, of course, render these "conflicting" dispositions coherent by ascribing to her some very complicated set of intrinsic values, the hypothesis that she has such complicated intrinsic values may be itself too preposterous to believe. It appears to me that the claim that a person intrinsically values some state of affairs is irremediably vague and that the evidence for such a claim is always ambiguous and incomplete.

And the problems with determining a person's values "begin at home." I do not think that our own values are open to direct introspection. As we do with others, but with somewhat different evidence, we formulate hypotheses and attempt to "make sense" of our behavior, our emotional responses, our internal dialogue, and all the other parts of our lives. And we could always be wrong. Still, in attempting to determine what is rationally advisable, we are trying to find out how to bring about what we value, what we care about—not what we *would* care about in other circumstances.[38]

Notes

[1]Throughout, I shall be talking about reasons in the sense that is relevant to "rational advisability". That is, I am not concerned with what it is reasonable for the person to do given his epistemic situation, which may involve non-culpable ignorance and error; I am concerned with what advice should be given the agent if the advisor adopted the normative perspective of the agent. Allan Gibbard (1990, pp. 18–19) has clarified the distinction between what it is reasonable for an agent to do and what it is advisable for the agent to do. I qualify the notion of advisability because I believe that there are many different perspectives from which one might advise an agent. We are not concerned with advice from a political perspective or from the moral point of view, for example.

[2]Richard Brandt (1979, especially Chapters VI and VIII), Richard Fumerton (1990, p. 137ff), David Gauthier (1986, Chapter II), James Griffin (1986, Chapters I and II), Shelly Kagan, (1989, Chapter 8, especially pp. 300–307) to name just a few. For his purposes, Kagan doesn't need to defend a hypothetical motivation *account* of reasons. It will do to show that the fact that one would be motivated to act in a certain way is *evidence* that there is a reason so to act. Depending on how weak we take this evidence to be, and what sort of defeaters we allow, this claim may be unexceptionable.

[3]Were we concerned with reasons in a sense relevant to the rationality of agents, rather than in a sense relevant to rational advisability of actions, then eradicating *all* false beliefs and, *a fortiori*, instilling all true beliefs would be inappropriate. Clearly we would not want to impugn the rationality of an agent for not acting as one should with full and error-free information if it was not a rational failing of his that was responsible for his errors and ignorance. Furthermore, there is at least one sense of the rational appraisal of *action* in which what is relevant are the beliefs that it is epistemically rational for the agent to have. However, as noted above (note 1), our concern here is with the advisability of action. (See Gibbard, 1990, pp. 18–19 and Hubin 1991, p. 4.)

[4]A fact may be considered relevant if it would affect the motivation of the agent given the other idealizing assumptions. This means that the relevance criterion depends crucially on these other idealizing assumptions. We might also need to worry about "order of introduction" problems: either of two facts may affect an agent's motivation if she is made aware of it *before* being made aware of the other and not affect her motivation if she has already been made aware of the other.

[5]On the Humean view, reason may endorse absolutely any action, given the appropriate passions. See Hume (1968) Book II, Part 3, § 3.

[6]By 'coherence' here, I mean such formal requirements as the following: that weak preference (the strict-preference-or-indifference relation) be reflexive and transitive, that strict preference be asymmetric, irreflexive and transitive, *etc.* Nozick (1993, p.140) takes this to be only "[o]ne tiny step beyond Hume, not something he need resist."

[7]Derek Parfit (1984) pp. 120–126.

[8]Brandt, Gauthier and Griffin think that it is the fact that one would, under certain non-actual conditions, have certain *desires* that entails reasons for acting. Kagan suggests an abstract account of a view that does not *require* desires to enter the picture even hypothetically, and his preference seems to be for fleshing out the account in a way that leaves them out.

[9]This form of the theory needs to exercise some care in characterizing the idealization of the cognitive states. If the removal of false beliefs and the infusion of true ones includes the very belief that is supposed to have motivational efficacy, then the hypothetical motivation framework seems to idle. Why not just say that one has reason to do what would produce the best outcome (or whatever the content of the motivating belief is supposed to be)?

[10]And, of course, there is always the trivial case in which we hypothesize that the agent is motivated to do all and only those things he has reason to do.

[11]I use this label with some reservations. The psychological state I find most plausible as the source of reasons in the sense relevant to the rational advisability of action is "intrinsic valuing." As will become clear, I do not think that the fact that an agent intrinsically values a state of affairs *entails* that he has any motivation to bring it about. (See note 37.) I think,

though, that it is plausible to hold that there is what some have called a 'conceptual connection' between intrinsic valuing and motivation such that it is not incorrect to think of intrinsic valuing as a "motivational state." Hence, I consider the theory that bases reasons on the intrinsic values of the agent a species of intrinsic motivation theory. Perhaps it is worth saying in defense of this terminology that even having a desire to do something, which we have consistently taken to be paradigmatic of a having a motivating state, does not *entail* motivation. Perhaps some forms of *ennui* present good examples of situations in which states we understand as motivating, fail to motivate.

[12]Whether this operationalization constitutes any real progress depends, as always, on whether the concepts involved in the operationalized analysis are any better understood than the concept we seek to operationalize.

[13]It is a mark of a certain sort of robustness for a theory to be such that small deviations from the idealizing conditions do not occasion dramatic deviations from the predicted result. If the idealization involves, for example, the absence of friction, then a robust theory will make predictions that are close to what we observe in real situations in which slight friction is present.

Not all theories are robust in this way. I suspect that the results of game theory are not. Common game-theoretic solutions depend on unrealistic assumptions about the rationality of the agents and the extent of their knowledge of the situation and one another's preferences and rationality. It is not clear that diminishing slightly the common knowledge of mutual rationality, for example, will lead to results that approximate the idealized case.

Whether the operationalization offered by a hypothetical motivation account of reasons is practical in the sense indicated above depends on whether that theory is robust in this sense. The answer to this last question is not obvious.

[14]Kagan suggests that we can gain understanding of what would motivate us in the ideal case by noting what motivates us as we approach the ideal (1989, p. 298).

[15]The notion that in appealing to our hypothetical unencumbered selves we appeal to a better judge is suggested by Kagan (1989, pp 301–7).

[16]It is our hypothetical judgment understood in as a disposition of the human organism to which Gibbard (1990, pp. 177–9) argues we must grant fundamental authority.

[17]For simplicity of exposition, I use Stalnaker's semantics for the counterfactual (1968). Substituting the more adequate Lewis (1973) or Pollock (1976 and 1981) semantics leaves the arguments in this paper unaffected.

[18]Of course, it is possible to construct a specification of the hypothetical situation so that we can still be talking about dispositions of the organism and avoid the problems raised. However, the hypothetical situation will have to be constructed so that non-psychological effects of the (other) idealizing changes are ruled out. The resulting "dispositions of the organism" will not be the sort of dispositions we are used to considering. Pigs may have the dispositional property of flight in the sense that, were we to alter their physical characteristics sufficiently, they would fly. Similarly, a human being, prone to a certain sort of motivation as a purely physiological result of altered beliefs, may nonetheless be said not to be disposed to this motivation on the grounds that, were the beliefs altered *and* all purely physiological effects of that alteration blocked, he wouldn't be motivated in the specified way. This is, though, just a way of shifting focus from dispositions of the flesh-and-blood organism in favor of dispositions of some more abstract entity, perhaps a person.

[19]For a discussion of related points, see Arthur Ripstein (1993).

[20]Again, we could present this in terms of a disposition of the chip by speaking of the output the Beast is disposed to produce if it correctly follows the algorithm, but then this is a universal disposition—*everything whatsoever* is disposed to yield the same output if it correctly follows the algorithm.

[21]It dims hopes for a practical operationalization of the concept of reasons. Now we not only need to observe how motivation changes when the subject gains knowledge (or whatever the idealizing assumptions require). We need to know whether these changes result from psychological processes or merely from brute physiological ones.

[22]Caution is advisable here. One doesn't want to characterize psychological normality in such a way as to rule out the effect of psychological characteristics that are essential to the agent. Otherwise, the dispositions we find will be dispositions *of the agent* only in a Pickwickian sense, at best.

[23]This example and its implications were pointed out to me by James Dreier.

[24]It is no help, though, with the corresponding problem of inappropriately generated motivational states. It may be, either as a result of brute physiological processes or psychological connections, that the idealizing conditions would produce motivation to perform some act that we think the agent has no reason to perform. One of the nicest examples of this, partly because it depends on a phenomenon with which most of us are familiar, is Allan Gibbard's. We may take vivid awareness of relevant facts to be a relatively noncontroversial idealizing assumption. But, the fact that vivid awareness of what we could do with a large bribe for performing some immoral action would leave us motivated to do it may merely be taken by us as a reason not to dwell on this aspect of the choice.

[25]If, to avoid an objective means/ends principle, we employ the other idealizing conditions to fix the agent's ends and then ask what, *given these ends*, the agent would be motivated to do, we encounter a dilemma. If means/ends rationality is not assumed, then the agent may not be motivated to perform actions that are means to the ends he would desire in the hypothetical circumstances. On the other hand if means/ends rationality *is* one of the idealizing conditions, we achieve only a terminological variant of the hypothetical ends view described in the text.

[26]It is plausible to think that what is going on is what Michael DePaul calls 'a formative experience' (1988, pp. 619–635).

[27]I unabashedly beg the question against internalism about *moral* reasons.

[28]The notion of *dependency* here—of some motivation being based on other, more basic motivation—is to be interpreted structurally, rather than causally. To cast the point in terms of 'desire' it is, of course, possible to come to desire something intrinsically because one sees that having such an intrinsic desire is instrumentally valuable—that is, satisfies some other desire. (See Hubin, 1991, p. 23.)

[29]It sometimes seems wrong to say that we have any reason at all to perform an action that satisfies one of our intrinsic desires if it also thwarts a much stronger intrinsic desire. I think it is reasonable to view this as a matter of pragmatics, not semantics. It is wrong (though not false) *to say* that one has a reason to kill one's children because one desires money and the insurance would pay off handsomely. The reason is "swamped"; there is no practical point to mentioning it. At least, I think, this is a plausible thing to say about these cases. If this is unacceptable, we can revise the claim to hold that we have reasons to perform those actions that promote states of affairs we intrinsically desire regardless of our actual or hypothetical instrumental desires *provided that these actions do not thwart stronger intrinsic desires*. For my present point, the addition makes no difference.

[30]I follow Shelly Kagan (1989, p. 17) in using the term 'pro tanto' to refer to an actual but overrideable reason—what is typically referred to as a *prima facie* reason. This avoids the ambiguity of the latter phrase between a *pro tanto* reason and a consideration that *appears* to be a reason but may, in fact, be no reason at all.

[31]This may be connected with the problem Fumerton refers to as "ethereal" values. "Although the issue is rather complicated, it seems to me that in the final analysis a value that would not be sustained through the realization that it has been satisfied is too ethereal to give us a reason to act" (1990, p. 140).

[32]I develop and defend more fully an "intrinsic valuing" version of the actual intrinsic motivation theory in "Values and Desires" (in progress).

[33]Of course, harming my friend *may* lead to my winning the game and slapping my daughter *may* lead to her being quiet, both of which ends I intrinsically value, let's suppose. If so, I don't object to seeing *these* values as generating reasons for acting, albeit swamped ones. The objection is that the ephemeral and psychologically disconnected desires, even if intrinsic desires, should not necessarily generate reasons for acting.

[34]This is, of course, mere assertion. While I believe it to be correct assertion, argumentation will have to wait for another occasion.

[35]I think this is a very common and tempting line. Richard Fumerton suggests it for just this problem (1990, pp. 138–141).

[36]Perhaps Darwall's case of Roberta is such a case.

[37]Geoff Sayre-McCord has pointed out to me that denying the identification of the state of intrinsically valuing with any simple disposition makes it implausible that any such disposition is *entailed* by the existence of intrinsic valuing. In particular, it probably is not a necessary

truth that one is disposed to be motivated to bring about what one intrinsically values. I believe that this is true and that this breaks a necessary connection between motivation (and even any interesting disposition to be motivated) and intrinsic valuing or the existence of practical reasons. Perhaps this renders the position I defend an externalist position. (I'm not too concerned about the definitional point.) On the other hand, intrinsic valuing is surely (as some might have said in an earlier time) *conceptually* connected to motivation, or at least to an interesting disposition to be motivated. This connection is evident in light of the fact that the connection with motivation is one of the most central causal connections by which we identify the psychological state of intrinsically valuing.

[38]I am indebted to James Dreier, Daniel Farrell, Geoff Sayre-McCord, Peter Vallentyne and two anonymous referees for this journal for helpful comments on earlier drafts of this paper.

References

Brandt, Richard. *A Theory of the Good and the Right,* (Oxford, England: Clarendon Press, 1979).

Darwall, Stephen. *Impartial Reason*, (Ithica: Cornell University Press, 1983).

DePaul, Michael. "Naivete and Corruption in Moral Inquiry", *Philosophy and Phenomenological Research* 48 (1988): 619–635.

Fumerton, Richard. *Reason and Morality,* (Ithaca: Cornell University Press, 1990).

Gauthier, David. *Morals by Agreement,* (Oxford: Clarendon Press, 1986).

Gibbard, Allan. *Wise Choices, Apt Feelings: A Theory of Normative Judgment,* (Oxford: Clarendon Press, 1990).

Griffin, James. *Well-Being: Its Meaning, Measurement and Moral Importance,* (Oxford: Clarendon Press, 1986).

Hume, David. *A Treatise of Human Nature*, (Oxford: Clarendon Press, 1968).

Hubin, Donald C., "Prudential Reasons", *Canadian Journal of Philosophy*, 10 (1980): 63–81.

———,"Irrational Desires", *Philosophical Studies*, 62 (1991): 23–44.

Kagan, Shelly. *The Limits of Morality*, (Oxford: Clarendon Press, 1989).

Lewis, David. *Counterfactuals,* (Cambridge, MA: Harvard University Press, 1973).

Nozick, Robert. *The Nature of Rationality*, (Princeton, NJ: Princeton University Press, 1993)

Parfit, Derek. *Reasons and Persons*, (Oxford: Clarendon Press, 1984).

Pollock, John. *Subjunctive Reasoning,* (Dordrecht: D. Reidel Publishing Co. , 1976)

———, "A Refined Theory of Counterfactuals", *Journal of Philosophical Logic* 10 (1981): 239–266

Ripstein, Arthur. "Preference" in *Value, Preference and Morality* edited by Christopher Morris and Rey Frey, (Cambridge, 1993)

Stalnaker, Robert. "A Theory of Conditionals", *American Philosophical Quarterly*, monograph series 2 (1968): 98–112.