Normativity, Commitment, and Instrumental Reason

R. Jay Wallace

Philosophers' Imprint
<www.philosophersimprint.org/001003>
Volume 1, No. 3
December 2001

There are two tendencies in our thinking about instrumental rationality that do not seem to cohere very well. On the one hand, the instrumental principle—enjoining us to take the means that are necessary relative to our ends—does not seem to apply indifferently to any end that we might be motivated to pursue. There is, for instance, no genuine requirement to take the means that are necessary for realizing ends that one merely happens to desire. This encourages what we might call a moralizing tendency in reflection about instrumental reason: the supposition that instrumental requirements come on the scene only in relation to ends that have themselves been endorsed in some way by the agent, as ends that it would be good or desirable to achieve.

On the other hand, it seems undeniable that agents can display a kind of instrumental rationality in the pursuit of ends that they do not themselves endorse, when for instance they are in the grip of *akrasia.* People sometimes exhibit great intelligence and skill in executing plans that they view as dubious or questionable—think, for instance, of the extraordinary talent many of us display at procrastinating when it comes to tasks that we regard as worthy but difficult. It seems plausible to regard this kind of intelligence—cleverness, as we might call it[1]—as a form of instrumental rationality, relative to the ends that we are in fact pursuing. This verdict, however, conflicts with the moralizing tendency in our reflection about the instrumental principle, since the cases at issue are precisely ones in which

[1]I borrow this term (loosely) from English translations of Aristotle. In the Aristotelian context 'cleverness' means instrumental effectiveness with respect to ends that are in fact bad; I shall use it to refer to effectiveness relative to ends the agent believes to be bad.

R. Jay Wallace is Professor of Philosophy at the University of California, Berkeley.

people do not endorse the ends they are pursuing as good or worthwhile on the whole. There thus appears to be a latent tension in our thinking about instrumental rationality.

My ultimate aim in this paper will be to resolve this latent tension. The key to doing so, I shall argue, is to arrive at an improved understanding of the options in moral psychology that are available to us for conceptualizing both rational and irrational motivation. My thesis will be that we can account adequately for instrumental rationality only if we depart from a motivational psychology that makes do with the elements of belief and desire. In particular, we need to suppose that rational agents are equipped with a capacity for active self-determination that goes beyond the mere susceptibility to desires and beliefs.[2] On this volitionalist picture, as I shall call it, rational agency is made possible when we choose or decide what to do in ways that align with our own reflective verdicts about the reasons that bear on our deliberative situation.

Before we can see how this volitionalist account helps us with the problem of instrumental rationality, however, it will be necessary to take a detailed look at the nature of the volitional commitments involved in human agency. In particular, we must consider the extent to which such commitments are to be understood in normative terms. I shall argue

[2]There are other strategies for capturing the distinctive intelligence of human agency without departing from a belief-desire motivational psychology. For instance, some philosophers postulate higher-order desires with normative content—such as the desire to do what one ought—to explain how rational agency can reflect the agent's grasp of their reasons. I discuss this strategy (under the rubric of "meta-internalism"), and contrast it with the volitionalist motivational psychology that is more conventionally associated with the Kantian approach, in my paper "Three Conceptions of Rational Agency," *Ethical Theory and Moral Practice* 2 (1999), pp. 217-242. In the present paper I focus primarily on the more conventional, volitionalist version of the Kantian approach, though many of the points I make against normative interpretations of the will apply equally to meta-internalism.

against the interpretation of volitional commitment as an essentially normative stance. This issue is of substantial interest in its own right, with important implications in regard to the possibility of irrationality in action. But consideration of the issue will also point the way to an improved understanding of instrumental reason, enabling us eventually to resolve the latent tension in our thinking about the instrumental principle to which I called attention above.

The paper divides into four sections. In the first, I consider a number of arguments that have recently been advanced in favor of the normative interpretation of self-determination. These arguments purport to establish that genuine agency and self-determination presuppose the agent's commitment to normative principles. But I show that the arguments do not succeed: there is no general reason why agents should not be able to commit themselves to ends that they do not really endorse. Section 2 considers some differences between practical and theoretical reason. I contend that the proper counterpart of belief in the realm of action is not desire but choice or decision; but I suggest that the latter states do not involve the normative commitments characteristic of belief. In the third and fourth sections I return to the problem of instrumental rationality. Drawing on the version of volitionalism defended in sections 1 and 2, I develop a non-moralizing account of the normativity of instrumental reason. In particular, I show how we can explain the normative force of the instrumental principle without supposing that the ends to which the principle applies need be endorsed by the agent, as good or worthy of pursuit. Among the advantages of this strategy, it will emerge, is the attractive interpretation it makes possible of the phenomenon I referred to above as cleverness.

1. Choice and Normative Endorsement

Let us begin with some issues in motivational psychology. I suggested above that we should reject the belief-desire model of human motivation and postulate a distinctive capacity for self-determining choice, as a precondition of rational agency. It is not my intention to argue directly for this suggestion in the present paper. Instead I want to raise an interpretative issue about the volitionalist strategy: how are we to understand the choices that, on the volitionalist picture, are characteristic of rational agency?

Christine Korsgaard has recently offered a normative interpretation of volitionalism, as a framework for understanding the binding force of principles of practical reason.[3] She takes choice to be a matter of first-personal commitment to pursue an end, where this commitment is essentially normative. To choose to do *x* is, in effect, to accept a "law" or normative principle specifying, in general terms, which features of one's circumstances give one reason to do *x*. This stance commits one, in turn, to complying with a supreme unconditional principle of practical reason, the Kantian moral law, as well as with principles of instrumental reason instructing one to take necessary means to one's ends. This extremely ambitious approach can be understood as an attempt to extract from an interpretation of what we are doing when we act an account of the normative force of basic principles of practical reason.[4] Moral and instrumental principles are binding on us as agents, insofar as we necessarily commit ourselves to complying with them through the normative act of choice.

Notice that there are two kinds of normative commitment involved in human action, on this account of it. There are, first, specific normative commitments regarding the value of the actions we set out to perform; these commitments are enshrined in normative principles that specify our reasons for acting as we do, in the particular circumstances that confront us. But action on the basis of such principles involves, secondly, commitment to comply with more general principles of moral and instrumental rationality, principles whose normative force is explicated in terms of this second moment of commitment. I shall consider this second variety of normative commitment in sections 3 and 4, below, when I return to the question of instrumental rationality; for the moment, I want to focus on the specific kind of normative commitment involved in ordinary choices or intentions to act, on Korsgaard's interpretation of them.

To be clear, the issue is not whether choosing or intending to do something necessarily gives one reason to do it. A view of this sort, to the effect that intention and choice are what we could call objectively normative states, might seem to be the most promising basis for a moralizing approach to instrumental rationality. But this is not the view from which Korsgaard begins, and it is therefore not the view that I shall engage with here. Korsgaard's starting point is that intention and choice are subjectively normative states, involving our

[3]For the normative interpretation of choice, as involving acceptance of a "law," see Christine M. Korsgaard, *The Sources of Normativity* (Cambridge, England: Cambridge University Press, 1996), pp. 97-100 and 222-233; see chaps. 3 and 4 for the argument that this stance commits one to complying with the moral law; and for the argument that this commits one to complying with a principle of instrumental rationality, see Korsgaard, "The Normativity of Instrumental Reason," in Garrett Cullity and Berys Gaut, eds., *Ethics and Practical Reason* (Oxford: Clarendon Press, 1997), pp. 215-254, especially pp. 243-254.

[4]A similar strategy is adopted by J. David Velleman, who holds that our understanding of the nature of reflective agency can deliver a substantive criterion for normative reasons. See his *Practical Reflection* (Princeton: Princeton University Press, 1989), chap. 7, and "The Possibility of Practical Reason," as reprinted in his *The Possibility of Practical Reason* (Oxford: Clarendon Press, 2000), pp. 170-199, at pp. 188, 193, 198.

acceptance of a law that identifies, in general terms, our reason for acting in a given way. We must consider this thesis on its merits, including the very suggestive arguments that Korsgaard has advanced for thinking that choice and intention can be made sense of only on the assumption that they represent subjectively normative attitudes.

It is useful to think of the content of choices as specified by something like Kantian maxims. These may be treated as having the following schematic form: "I shall do x, (under circumstances c), in order to y/as a way of y-ing." Interpreted in this way, maxims articulate an agent's more or less provisional plan; choosing or deciding to do something can thus be thought of as committing oneself to a plan of action, the details of which can range from sketchy to quite complete.[5] Moreover, when one has reached a settled view about what one has reason to do, or which course of action it would be best to pursue, this view may be reflected in the content of one's choice. If, for instance, I believe it would be best to stop at a cafe after touring a new city for several hours, on account of my aching feet, my actual decision about what to do will ordinarily give expression to this normative belief. That is, I will commit myself to a maxim or plan of action with the following content: "I shall stop at a cafe, in order to rest my weary feet." When our intentions in acting rest on our conception of our reasons in this way, we should agree that the choices reflect normative commitments.

It is unclear, however, why choice should be thought of as necessarily normative, in the subjective sense distinguished above. In cases of *akrasia,* for example, we certainly

[5]I am indebted here—and indeed throughout my discussion of the volitionalist approach—to Michael Bratman's pioneering work on planning agency; see his *Intention, Plans, and Practical Reason* (Cambridge, Mass.: Harvard University Press, 1987).

appear to choose to act in ways that we ourselves do not regard as justified or best. What normative "law" is supposed to be implicit in choices of this kind? It cannot be the principle that we ought to be doing the action that we have chosen to perform, on pain of simply denying that *akrasia,* in the strong and philosophically interesting form, is so much as possible.[6] Perhaps, then, Korsgaard has in mind a normative "law" in a somewhat weaker sense—a principle, for instance, specifying that the action one has chosen to perform is at least *pro tanto* good.[7] But offhand, even this seems to go too far. There are cases in which we choose to do things without believing that there is anything genuinely good about them, in the actual circumstances at hand—the apparent value of the action we perform has the status of a *prima facie* good, not a *pro tanto* good. And there appear to be other, more alarming cases in which we choose to do things that we believe to be bad, precisely on account of their seeming badness.[8]

Here it is important to distinguish between the norma-

[6]In *The Sources of Normativity,* sec. 3.3.2, Korsgaard provides an explanation of how *akrasia* is possible, turning on the idea that we can "make an exception of the moment or the case" (p. 103). The question, however, is how this is to be reconciled with her claim about the normative "law" implicit in the stance of choice. Either "making an exception" is construed in normative terms, as endorsement of the principle that it is permissible to give in to temptation under the circumstances, in which case we are no longer dealing with a case of acting against one's better judgment. Or "making an exception" amounts to intentionally violating a normative principle we accept, in which case the claim about the normative "law" implicit in choice seems to go out the window.

[7]On the importance of *pro tanto* reasons and values in accounting for cases of *akrasia,* see S. L. Hurley, *Natural Reasons: Personality and Polity* (New York: Oxford University Press, 1989), part 2.

[8]For insightful discussion of cases of this kind, see Michael Stocker, "Desiring the Bad," *The Journal of Philosophy* 76 (1979), pp. 738-753, and J. David Velleman, "The Guise of the Good," as reprinted in his *The Possibility of Practical Reason,* pp. 99-122.

tive judgments that an agent genuinely accepts, and the normatively structured thoughts that may be prompted by the agent's desiderative and emotional states. It is states of these latter kinds that generally incite us to act at variance with our settled views about what there is most reason to do. Moreover, I believe that such desiderative and emotional states typically involve normative cognitions of one kind or another. Thus, on the first warm day of the summer term one may find that one wants to head off to the beach, and this desire will show itself in the thought that it would be good (because, say, pleasant) to spend the day amidst the surf and sand. I believe further that the connection between desiderative and emotional states and such normative cognitions helps to explain the fact that it occurs to us at all to perform actions that we do not really believe to be good. Thus, if I am angry or embittered, the fact that a prospective course of action would be bad may appear to render it attractive, and I will be tempted to opt for the course of action on account of its badness. But normatively structured thoughts of this kind are not to be confused with normative judgments or beliefs. Our intellectual capacities include the ability to entertain thoughts that we do not genuinely accept as true, and the gap between normative thought and normative judgment makes possible akratic action in the absence of a belief in the (*pro tanto*) goodness of what one is doing. One may act on one's desire to go to the beach, for instance, without really accepting that the pleasures thus made available provide a reason to skip the class one is scheduled to teach.[9]

[9]I elaborate on these suggestions about the cognitive structure of desire and emotion in my paper "Addiction as Defect of the Will: Some Philosophical Reflections," *Law and Philosophy* 18 (1999), pp. 621-654. Note that if my remarks here are on the right lines, there will be a different sense in which all choice might be said to be "subjectively" normative, insofar as choice or commitment always presupposes at least the *apparent* value of the ends chosen. I take it, however, that Korsgaard wishes to affirm the subjective normativity of choice in the different and stronger sense discussed in the text.

Of course, in what is doubtless the more common variety of akratic action, the agent accepts that there is something that is *pro tanto* good about the action that is performed. Furthermore, it is the fact that the action is believed to be genuinely good in some way that renders it an eligible candidate for choice, from the agent's point of view. Even when this is the case, however, we must be careful to distinguish between the agent's normative beliefs and the act of choice itself. After all, in many of the situations in which we judge that there is something *pro tanto* good about the action we are performing, we also judge that that action is not the one that it would be *best* to perform, on the whole. This is the general normative judgment that is authoritative for our practical reflection about what to do, and yet we choose to do something else instead. If we are to leave open the possibility of this kind of akratic action, we cannot understand choice exclusively in normative terms. Choice may often reflect or be based on normative commitments that the agent accepts, but it cannot be *identified* with such commitments without foreclosing genuine possibilities in the theory of action. There has to be something in the act of choice that distinctively goes beyond normative commitment if we are to leave room for *akrasia* and the other forms of irrationality to which action is characteristically subject.

Korsgaard offers two main arguments against this line of thought. The first of these appeals to the important idea that, as agents, we are not merely determined to act by the states of desire to which we are subject.[10] We are, as other

[10]What follows is an interpretation of Korsgaard's remarks in the "Reply" chapter of *The Sources of Normativity*, pp. 222-233. The task of interpretation is complicated by the fact that, when Korsgaard introduces the notion of a "law" in her argument for what she calls the categorical imperative, she does not explain

proponents of volitionalism should agree, active with respect to our motives, and Korsgaard contends that this makes sense only on the assumption that when we act, we endorse a universal normative principle. Thus she writes: "the special relation between agent and action, the necessitation that makes that relation different from an event's merely taking place in the agent's body, cannot be established in the absence of at least a claim to law or universality."[11] This claim is to be understood as "a claim that the reasons for which I act now will be valid on other occasions, or on occasions of this type—*including this one, conceived in a general way*."[12] Without a claim of this kind, Korsgaard suggests, agency effectively dissolves, insofar as we lose the conceptual resources for distinguishing between the choices of the agent and the results of psychological forces to which the agent is subject. Choice is intelligible only on the assumption that it is at least possible to fail to follow through on one's choice. This in turn supports the identification of choice with the act of commitment to a general principle, a universal law by reference to which some possible performances can be interpreted as failures.[13]

or even give an example of what she has in mind (see *The Sources of Normativity*, secs. 3.2.3-3.2.4). But in the "Reply" it seems clear that she takes the universal laws to which we commit ourselves in acting to be normative principles, specifying our conception of our reasons. She opens her discussion there by raising the question of "why the dictates of the free will must be universal in any sense at all" (p. 222). A little later, however, she characterizes "the point we were supposed to be establishing" as the thesis that "*reasons* are general" (pp. 224-225, my emphasis); and she begins talking about "the *normative* principles of the will" (p. 229, my emphasis). This strongly supports my assumption that "law" is to be understood throughout in the sense of a normative principle of action, specifying in general terms the agent's conception of their reason for acting.

[11]*The Sources of Normativity*, p. 228.

[12]*The Sources of Normativity*, p. 232.

[13]*The Sources of Normativity*, pp. 228-233.

This argument seems to me correct in the following respect: when our choices to act are based on our conception of what we have reason to do, they will entail that we accept some general normative principle. The reason for this is that the conclusions of normative reflection are best understood as implicitly general judgments.[14] If I conclude that my being in the mood for an action movie gives me reason to see the latest John Woo film, I commit myself thereby to a normative principle that is general, insofar as it could apply to other situations besides the present one: for example, that—other things being equal—one has reason to go to the kind of movie one is in the mood for, when it is a question of what would make for an entertaining evening.[15]

For present purposes, however, the relevant question is why one should identify the act of choice with the acceptance of a normative judgment of this kind. Korsgaard is surely correct to insist that our choices are recognizable as expressions of our agency only on the condition that it is in principle possible for us to fail to follow through on them. This means, perhaps, that their content—the plan of action given expression in a maxim—must be specifiable in terms that are to some extent general.[16] But there is no reason to

[14]Compare T. M. Scanlon, *What We Owe to Each Other* (Cambridge, Mass.: Harvard University Press, 1998), pp. 73-74.

[15]As this example illustrates, the commitment to generality that is at issue is a fairly modest one. Note in particular that the general judgment I have formulated incorporates an "other things equal" clause that the agent probably would not be able to unpack in non-trivial terms at the time of action. It is unclear whether this is at odds with Korsgaard's intentions, but it does make the talk about "universal laws" seem somewhat overblown. For discussion of this issue, see Michael Bratman, "Review of Korsgaard's *The Sources of Normativity*," as reprinted in his *Faces of Intention* (Cambridge, England: Cambridge University Press, 1999), pp. 265-278, sec. 4.

[16]Again, the commitment to generality at issue is fairly minimal; there is nothing to rule out such indexically formulated intentions as the intention to catch *that man* in order to return his hat to him.

suppose that the general specification must amount to a normative principle. Thus, in a spiteful and nasty mood I might resolve to burn all my roommate's books, without really supposing that what I am doing is best, on the whole; indeed, I might not really believe that it is good or justified in any way at all. In this case, the information supplied about the content of my resolution is enough to undergird the attribution of the resolution to me, as agent; it specifies a goal that I might in principle fail to reach—by, say, neglecting to burn the roommate's cookbooks in the kitchen. But identification of me as the agent of the choice does not require that I accept a normative principle justifying the action chosen, and in the case under consideration it would seem implausible to construe the choice as a commitment to a principle of this kind. Even if I accept that what I am doing is *pro tanto* good (insofar as it causes my roommate distress, say), my choice cannot be identified with acceptance of such a principle without rendering mysterious the phenomenon of akratic choice in the face of normative judgments about what it would be best to do on the whole. If an agent really accepts that a given action would be best, the identification of choice with normative commitment should entail that that is the action that is chosen, in fact—an apparent denial of the very possibility of clear-eyed *akrasia.*

On Korsgaard's behalf, it might be suggested that we can leave open this possibility by distinguishing between normative judgments and normative commitments. Akratic agents judge that some action *y* would be best, but commit themselves to an alternative action *x*, which they thereby affirm to be good along some dimension. But this just concedes the point I have been at pains to make in this section. The distinction between normative judgment and normative commitment can be drawn only if there is something in the act of commitment that distinctively goes beyond acceptance

of a normative principle or judgment, and this assumption calls into question the identification of volitional commitment with the acceptance of a normative principle or judgment. In any case, the example above makes clear that we do not need to identify choice with normative commitment in order to make sense of our authorship of our actions. So long as choices are interpreted in terms of an implicitly general plan of action, we have resources enough to render intelligible the attribution of them to the agent.

Korsgaard's second argument seems to take for granted that choices might be intelligible as such even if their agent does not accept an antecedent normative principle that justifies them. Thus, she imagines a "heroic existentialist" who chooses to pursue an end without supposing that there is anything independently good about the end to be pursued.[17] This sounds like the kind of case I have just described, except that Korsgaard goes on to add that the existentialist must at least view their own act of will as normative, as creating a reason to act where there was none before. In effect, it seems, the heroic existentialist endorses a general normative principle whose content is that one has reason to pursue those ends one has chosen to pursue; it is only that this principle does not and cannot provide an independent justification for the initial act of choice.

But why must one accept a normative principle of this kind in support of the action one has chosen to perform? As I cart the books out to the back yard and fling them onto the pyre, must I really suppose that the bare fact of my having

[17]"The Normativity of Instrumental Reason," pp. 250-253. It is not clear that Korsgaard herself understands this case as opening up a new line of argument, in part because she seems in some doubt as to whether the attitude of the heroic existentialist is a real possibility (see especially p. 251). But if it is conceded to be a possibility, the case calls in question the strong conclusion of her first argument, namely that choice is intelligible only if it is justified by a universal normative judgment the agent accepts.

resolved to do such a silly thing makes it a good thing for me to be doing? Korsgaard answers such questions as follows: "If I am to will an end, to be and to remain committed to it even in the face of desires that would distract and weaknesses that would dissuade me, it looks as if I must have something to *say to myself* about why I am doing that—something better, moreover, than the fact that this is what I wanted yesterday."[18] Well, in the case we have been imagining I do have something better to say to myself than that burning the books is what I wanted yesterday, namely that it's what I have resolved to do. Unless one is a heroic existentialist, this doesn't by itself count as a justification for the action; but why should it be thought psychologically impossible to carry out one's intentions if one doesn't have a way of justifying them to oneself?

The reasonable point to which Korsgaard is calling attention is that intentions that diverge from one's normative judgments will not form a reliable basis for long-term planning about the future. If I really believe that it would be best to go into the office next Saturday to work on admissions files, it would be peculiar for me nevertheless to say that I intend to stay home on Saturday and watch the game; my normative judgment about what I ought to do would lead me to hope that I do not come to act on the intention, and this would undermine the role of the intention in my planning for the future. Since future-directed intentions are plausibly understood in part in terms of the role they play in such planning agency,[19] it may be doubted whether we would really want to speak of intention in a case such as

[18] "The Normativity of Instrumental Reason," p. 250.

[19] Here I am drawing on Bratman, *Intention, Plans, and Practical Reason*. Also relevant are some of the considerations raised by Gregory Kavka in "The Toxin Puzzle," *Analysis* 43 (1983), p. 33-36; for a sophisticated recent discussion of these issues, see Michael Bratman, "Toxin, Temptation, and the Stability of Intention," as reprinted in his *Faces of Intention,* pp. 58-90.

this.

But this good point does not rule out the possibility of short-term intentions to act—still less intentions in acting—that diverge from our normative commitments. Granted, an agent who encounters large obstacles on the way to executing an akratic intention of this kind will find it hard to follow through on their intention, and will probably give up. But not necessarily: thinking that I really shouldn't do so, I might nevertheless choose to go out and buy a bottle of rum—and persist, despite discovering that the first shop I drive to is closed, and the second out of stock. In any case, there are plenty of situations in which we don't encounter any unusual additional obstacles on the way to carrying out our short-term akratic intentions. To suppose that the execution of such intentions must be impossible is, it seems to me, to neglect a large and interesting spectrum of cases of freely chosen human action, encompassing such phenomena as sheer willfulness, stubbornness, lethargy, habit, blind self-assertion, thoughtlessness, and various actions expressive of emotional states.[20]

To be sure, in not all cases of this kind is it equally plausible to suppose that agents really choose or commit themselves to acting in a way they do not endorse. People quite often decide on a course of action that they take to be sup-

[20] Compare Velleman, "The Guise of the Good." Velleman himself supposes, however, that agents who have chosen to do something precisely because it is bad must take the badness of the chosen action to be a reason for performing it, something that contributes to the intelligibility of what they are doing (see "The Guise of the Good," pp. 121-122). This further claim seems strained to me, given the rest of what Velleman says about the cases he discusses: a person completely unconcerned about the goodness of what they are doing would presumably be equally unconcerned about whether what they are doing is fully intelligible, or otherwise makes sense. The claim is connected with Velleman's account of full-blooded human action as motivated by a basal desire of the agent's that provides a criterion for something's being a normative reason (see, e.g., his "The Possibility of Practical Reason").

ported by good reasons, and continue on that course in the face of subsequent doubts or new reasons to reconsider their original normative assessment of what they are up to. There is a kind of "volitional inertia" that enables us to carry on as before, despite our having revised our judgment about whether what we are doing is well-advised. (Thus we may continue to flip through the channels on the television long after it has become clear to us that there is nothing on, and that it would be better to go back to work.) In these cases we may speak of practical irrationality in the absence of a deliberate choice that goes against normative principles we ourselves accept.[21] My point, however, is that there is also room for a more extreme kind of irrationality in action, in which we deliberately *choose* to act in a way that is at variance with our own normative beliefs about what it would be best to do. (We sometimes choose to turn on the television while knowing in advance that there is nothing on, and that it would be better to get back to work.) Furthermore, the same phenomenon of volitional inertia implicated in cases of irrationality without choice can help us to execute short-term intentions that exhibit this kind of extreme irrationality. Thus, if we can carry out even fairly complicated activities in the face of a revised normative assessment of their value, we should equally be able to carry out such activities when the choice to engage in them was not one that we initially endorsed—provided, perhaps, that we do not encounter too much resistance along the way.[22] One would need a power-

[21]See Hilary Bok, "Acting Without Choosing," *Noûs* 30 (1996), pp. 174-196, for an illuminating discussion of cases of this kind.

[22]It is interesting to note here that cases of carrying on in the face of "external" obstacles (the closed liquor shop) are more easily conceivable than cases of carrying on in the face of such "internal" obstacles as laziness, boredom, and distraction. If I believe sincerely that there is nothing to be said for burning my roommate's books, then—barring a particularly strong case of "volitional inertia"—I will almost certainly stop burning them when I become interested in a

ful philosophical argument to establish that appearances must be deceptive in this area, that we can follow through on our immediate intentions in the face of potential resistance only if we initially viewed them as justified.

Korsgaard's first argument aims to establish a conclusion of this kind, but as we have seen that argument does not succeed. Beyond that, I can merely venture a diagnosis of her position. Korsgaard seems to assume that there are only two options in the theory of the will: either we assume that the ends we pursue are fixed by our desires, or we grant that persons are capable of choosing for themselves what they shall do, where choice in turn is a matter of normative endorsement.[23] Since I agree with her that the first option is unattractive, I concede that we would have grounds for preferring the second, if it is the only alternative. But it should by now be apparent that I do not believe this to be the only alternative. Human agents have the capacity for a sophisticated kind of rational agency, insofar as they can reach in-

football game on television (thanks to Ruth Chang for this example). The reason for this, I would suggest, is that akratic action presupposes an emotional or desiderative state—involving, as I explained above, normative thoughts about the action one is performing—that is contingently incompatible with such states as boredom and distraction. Thus, the kind of intense anger or spite that might lead me to want to burn my roommate's books leaves little psychological space for (say) interest in a football game on TV; by the time I become interested in the game, my anger will have abated to the point where I am no longer even tempted by the prospect of burning the books.

[23]See, for instance, "The Normativity of Instrumental Reason," p. 251, note 74, where Korsgaard writes: "The heroic existentialist's ends are not merely the objects of his desires, but rather of his will, so he is not merely given them by nature: he has endorsed them, and to that extent he does see them as things he has reason to pursue." What I am questioning is why an object of one's will must necessarily be something one endorses, as reason-giving in any sense. For a different challenge to Korsgaard's conception of the options in the theory of action, see Bratman, "Review of Korsgaard's *The Sources of Normativity*," pp. 276-277.

dependent normative conclusions about what they have reason to do, and then choose in accordance with such normative conclusions. This capacity presupposes that we are equipped with the power to choose independently of the desires to which we are subject.[24] Once we have this power, however, it can be put to use in ways that are at odds with our own practical judgments about what we have reason to do. That is, we can treat our disposition to do what we ought as a further desire from which we set ourselves apart, choosing to act in a way that is at variance with our reflective better judgment. This may be regarded as a hazardous by-product of the capacity for self-determination that makes rational agency possible in the first place.

2. Intention and Belief

Reflecting on this side of agency, we see that there are important dissimilarities between our capacities in the practical and the theoretical spheres. In much writing about practical reason it is customary to press very hard an analogy between reasons for belief and reasons for action.[25] Korsgaard's treatment of the will provides an example of this trend. Thus, she rejects the assumption that the proper counterpart of belief in the practical sphere is desire; pointing out that believing is an essentially normative act, she suggests that the right analogue in the practical sphere must involve a similarly normative commitment, arguing that volition or choice is suited to play this role.[26] I would agree with Korsgaard that it is a mistake to attempt to reconstruct agency in terms of the concept of desire, but disagree that agency is like believing in being essentially normative.

The inherent normativity of believing is reflected in the fact, to which G. E. Moore famously called attention, that first-person assertions of the following forms are paradoxical: "*P* is true, but I don't believe it" and "I believe that *p*, but *p* isn't true." Moore's paradox brings out that to believe a proposition is to be committed to its truth,[27] and this normative aspect of believing, as we might call it, is connected with the further fact that there are clear limits, of a conceptual nature, on the possibility of believing something at will.[28]

Against this suggestion, it might be objected that normativity has to do with reasons, not with truth.[29] To say that belief is an inherently normative stance is to say that it is specially connected to conceptions of what one has reason to believe, and nothing would seem to follow from the considerations relevant to Moore's paradox about whether this is or

[24]I understand by desire here an occurrent state of being attracted or drawn to a course of action, which can become an object of reflective self-awareness. I discuss the volitional capacity to choose independently of desire in this sense in the following papers: "Three Conceptions of Rational Agency," sec. 3; "Addiction as Defect of the Will;" and "Moral Responsibility and the Practical Point of View," in Ton van den Beld, ed., *Moral Responsibility and Ontology* (Dordrecht: Kluwer Academic Publishers, 2000), pp. 25-47.

[25]For examples of this tendency, see Philip Pettit and Michael Smith, "Freedom in Belief and Desire," *The Journal of Philosophy* 93 (1996), pp. 429-449; Peter Railton, "On the Hypothetical and Non-Hypothetical in Reasoning about Belief and Action," in Cullity and Gaut, eds., *Ethics and Practical Reason,* pp. 53-79; Scanlon, *What We Owe to Each Other,* chap. 1; and Velleman, "The Possibility of Practical Reason."

[26]"The Normativity of Instrumental Reason," pp. 248-9, especially footnote 69.

[27] For a much fuller presentation of the thought that this is one of the lessons of Moore's paradox, see Richard Moran, "Self-Knowledge: Discovery, Resolution, and Undoing," *European Journal of Philosophy* 5 (1997), pp. 141-161. I have formulated the paradoxical Moore-sentences in terms of the truth predicate to emphasize what Moran refers to on p. 157 of this paper as "the internal relation between belief and truth."

[28]See Bernard Williams, "Deciding to Believe," in his *Problems of the Self* (Cambridge, England: Cambridge University Press, 1973), pp. 136-151. For a nuanced discussion of the precise nature of the commitment to truth implicit in belief, see Velleman, "The Guise of the Good," pp. 110-115, and "The Possibility of Practical Reason," pp. 182-186.

[29]I am indebted to John Broome and Hannah Ginsborg for pressing me to be clearer about the sense in which belief is an inherently normative attitude.

is not the case. I agree that it is natural to understand normativity as a matter of reasons, and that the considerations just adduced do nothing to suggest that belief is intrinsically connected to reasons in a way that intention is not. In fact, as I shall suggest below, neither belief nor intention is an intrinsically normative stance, if normativity is construed in terms of the idea of intrinsic and necessary responsiveness to (judgments about) reasons. But belief is intrinsically and necessarily responsive to (judgments about) truth, and this yields a different sense in which belief—by contrast to choice or intention—can be claimed to be an essentially normative stance.

We may think of theoretical reason as a capacity to modify our beliefs directly, through reflection on the question of what we should believe.[30] Considerations pertaining to the truth of propositions are normative for theoretical reason, in this sense, insofar as they are immediately and constitutively relevant to theoretical reflection about what we should believe. Presumably we do not have reason to believe every proposition whose truth is epistemically accessible to us—many truths are too trivial or tangentially related to our interests to be the sorts of things we should bear in mind, or keep track of consciously, in the ways characteristic of belief. But when we do have reason to believe that *p*, the considerations that provide us with reason to do so will be considerations that speak in favor of the truth of *p*, and their status

[30]Some may object to this characterization that "oughts" and reasons have a comparatively modest role to play in theoretical reasoning, insofar as changes in our beliefs are often effected without reflection on the question of what we ought to believe. Nothing in the argument to follow hangs on my characterization of theoretical reason. For the record, however, it seems to me important to bear in mind that many of our beliefs are not arrived at through reasoning. Once we are clear about this, it becomes quite plausible to suppose that the cases of belief-revision that do involve reasoning are precisely cases in which there is at least implicit consideration of reasons.

as reasons for belief will be connected crucially to the fact that they point in this way toward the truth.[31] Reasons for action, on the other hand, are considerations that bear on the question of the goodness or value of action. Truth and goodness are thus the normative aims of belief and action, respectively: those aims by reference to which we make sense of considerations as reasons for belief and for action. But if believing that *p* is (*inter alia*) a commitment to the truth of *p*, it follows that there is a sense in which belief can be considered an inherently normative stance. It involves our acceptance that the aim by which theoretical reflection is properly governed has been achieved, and this in turn places constraints on our capacity to believe things at will. By contrast, we do not ordinarily suppose that our capacity to *intend* or *choose* is similarly constrained by the aim that is normatively regulative of practical deliberation, regarding the value of the alternatives open to us. The question of what action we are going to perform is not necessarily answered by our having determined to our own satisfaction what it would be best to do.[32]

This intuitive disanalogy between the practical and theoretical cases is reflected in the fact that first-person utterances of the following statements do not appear to be paradoxical at all: "I really ought to do *x*, but I'm going to do *y* instead," "*X* would be the best thing to do under the circumstances, but I intend to do *y*," "I've chosen to do *y*, though it's

[31]Of course, other kinds of considerations are sometimes brought to bear in assessment of our beliefs, as when it is said that we would be happier if we believed that our colleagues loved us. This kind of consideration is not however itself a reason for *believing* that our colleagues love us—though it might be a reason for *acting* in some way or other (e.g. for undertaking to induce the relevant belief by hypnosis or other indirect means).

[32]Compare Rogers Albritton, "Freedom of Will and Freedom of Action," *Proceedings of the American Philosophical Association* 59 (1985-86), pp. 239-251, at pp. 246-248.

not in fact the best alternative open to me," etc. Of course, if one has a thesis to maintain about the essentially normative character of choice, it would be possible to interpret the normative vocabulary deployed in such statements in an "inverted commas" sense. But no thesis of this kind can be motivated by reflection on those statements alone. Our sense that they are not paradoxical does not rest on our recognition of the possibility that the speaker might be deploying evaluative vocabulary insincerely. It rests, rather, on the conviction that our capacities for agency and choice can be exercised in a way that does not automatically align with our normative convictions. We do not think of choice as an essentially normative stance, and this is connected with our feeling that our active powers of self-determination in the practical domain present us with a set of alternatives for action that is wider than the set of actions we ourselves approve of.

The disanalogy between the theoretical and the practical cases is further reflected in the fact that, whereas *akrasia* in the practical sphere is an intelligible and even familiar phenomenon, strong *akrasia* of belief is rather harder to imagine.[33] The latter would be a case in which one judges that a given conclusion—say, *p*—is true, and yet one consciously and without self-deception believes that not-*p*.[34] But how can one believe that not-*p* in this way, if at the same time one consciously judges that *p* is true? T. M. Scanlon, who presses in a different way the analogy between theoretical and practical reason, has proposed an answer to this question. Drawing on the plausible assumption that belief involves an interconnected network of dispositions over time (including

[33]On this point, I agree with Hurley, *Natural Reasons,* pp. 130-135, 159-170.

[34]We might contrast strong *akrasia* of belief in this sense from weak *akrasia* of belief, in which one merely fails to accept *p* in the face of the judgment that *p* is (very likely to be) true.

dispositions to recall the proposition one believes and to feel convinced about it, to use it as a premise in further reasoning, etc.), Scanlon suggests that *akrasia* is no less conceivable in connection with belief than in the practical realm: "I may know, for example, that despite Jones' pretensions to be a loyal friend, he is in fact merely an artful deceiver. Yet when I am with him I may find the appearance of warmth and friendship so affecting that I find myself thinking, although I know better, that he can be relied on after all."[35]

There is no doubt that cases of this kind are possible, perhaps even common. The question is whether they qualify as cases of *akrasia* of belief in the strong sense defined above. I am not convinced that they do; here it is necessary to recall the distinction introduced above between normatively structured cognitions and normative judgments. Certainly Jones' appearance of warmth and conviviality can prompt in me the thought that he is a decent friend. But do I really *believe* that this is the case, if at the same time I know that his appearance of friendship is nothing more than an artful pretence? This seems highly implausible. Of course, the thought that Jones is reliable might turn into a belief, if the force of his warm appearance prompts me to reconsider my judgment that he is merely a deceiver. But if I remain committed to that judgment—if, as Scanlon puts it, I *know* that Jones is merely an artful deceiver—then the thought that he is reliable cannot be considered a proper belief. The reason, again, is that belief is an inherently normative stance, in the special sense distinguished above.

Granted, belief is probably best understood as involving a network of dispositions that extends over time. It follows, perhaps, that the normative judgment that *p* is true does not entail that one actually forms the sustained belief that *p*, in

[35]Scanlon, *What We Owe to Each Other,* p. 35.

the full sense of the word.[36] Much of the trivial but presumably reliable information one reads about in the daily *Times,* for instance, is not retained in memory, deployed in future episodes of reasoning, associated in one's thought with a feeling of conviction, and so on. Normative commitment to the truth of *p* may thus not be sufficient to ensure that one actually comes to believe that *p*. Nevertheless, normative commitment of this kind does seem necessary to the stance of believing that *p*. That is, when one believes that *p*, one is thereby committed to the truth of *p*. This is what rules out strong *akrasia* of belief, in which one consciously believes that not-*p* while also judging that *p* is true.

There is a different phenomenon that is often discussed under the heading of *akrasia* of belief.[37] This is the phenomenon whereby one believes that not-*p*, while also accepting that the available evidence speaks overwhelmingly in favor of *p*. To take a clear if somewhat hackneyed example: parents may find themselves clinging to the belief that their daughter is still alive, while granting that all indications point toward the conclusion that she went down with the other passengers in the shipwreck. What makes this phenomenon possible, despite the kind of normativity I have argued to be inherent in belief, is the logical gap between theoretical reasons and the truth of the propositions for which those reasons speak. Even when the available evidence points overwhelmingly to the truth of *p*, it is still possible that *p* is false, and the person who hangs onto the belief

[36]Some cases of this kind will be cases of weak *akrasia* of belief. But not all cases: there are many contexts, such as the one I go on to describe in the text, in which considerations such as intellectual clutter-avoidance make it perfectly rational not to believe (in Scanlon's dispositional sense) all of the propositions whose truth one is prepared to grant. Compare Gilbert Harman, *Change in View* (Cambridge, Mass.: MIT Press, 1986), p. 56.

[37]See, for example, Alfred R. Mele, *Irrationality: An Essay on Akrasia, Self-Deception, and Self-Control* (New York: Oxford University Press, 1987), chap. 8.

that not-*p* in the face of massive evidence to the contrary is exploiting this possibility. For this reason, belief resembles intention in respect to the different sense of normativity distinguished above: neither attitude is intrinsically and necessarily responsive to (judgments about) reasons. But there is no similar logical gap to exploit in the case in which one accepts not merely that the evidence speaks in favor of *p*, but that *p* is true; this is what rules out the possibility of strong *akrasia* of belief such as I have described. Nor is there any need to appeal to a gap of this kind to account for the possibility of *akrasia* in the sphere of action.[38] The akratic agent may choose to do *x*, while believing not merely that the evidence speaks in favor of the conclusion that some alternative action *y* would be better, but that *y* would in fact be better.

This is the respect in which theoretical reason seems disanalogous to practical reason. I have contended that there is no paradox involved in choosing to pursue an end that one acknowledges to be bad, tracing this to the idea that volition differs from belief in not being an essentially normative commitment. Having said that, however, I should also like to reiterate that there are complex and important connections between choice and normative *concepts.* Thus, in cases in which we choose at variance with our better judgment there must be something that makes the action chosen seem attractive, an eligible candidate for performance from the agent's point of view, and this will typically be a function of our normative cognitions. We might believe, for instance, that what we are doing is *pro tanto* good, while judging that it is not really best on the whole. Alternatively, states of emotion or desire can make it *seem* to us as if our actions are valuable in some dimension, even if we are aware that they are not valuable in fact. Furthermore, citing

[38]Thus, I would dispute Mele's claim that "strict incontinent believing is possible for roughly the reason that strict incontinent action is," *Irrationality,* p. 119.

these kinds of evaluative thoughts and cognitions can help us to understand akratic actions retrospectively, making it at least partially intelligible why the worse act was freely chosen, what made it seem attractive to the agent at the time. In this sense, evaluative cognitions can illuminate the reason why akratic agents act as they do.[39]

In practice, of course, cases of non-normative choice are the exception rather than the rule. Our capacities for self-determining choice are what make possible deliberate practical rationality in the face of temptation, and in a vast range of cases we exercise them in ways that we believe will facilitate rather than thwart the realization of aims we endorse. Exemplary in this connection is the phenomenon of decision, which we ordinarily understand as the deliberative resolution of uncertainty on an agent's part about what is to be done. In ordinary decision-making, the agent arrives through reflection at the decision *that* (say) *x* would be the best thing to do, where this in turn involves a corresponding orientation of the will—a decision *to* do *x*. There is no phenomenological gap between the two sides of decision, the normative judgment, on the one hand, and the formation of a corresponding intention on the other, and this is reflected in the use of a single term to refer to both aspects. In cases of this kind, the content of the intention with which the agent acts, specifying what it is that the agent decides to do, will be a maxim of the subjectively normative sort discussed in the preceding section, one that reflects the agent's judgment about their grounds for doing what they have chosen to do.

Furthermore, even when we fail to do what, by our own

[39]It has become something of a truism in philosophical discussions of *akrasia* that the akratic agent acts "for a reason." This way of speaking is harmless enough, if it is only meant to call attention to the fact that akratic agents act intentionally, and that their choices can be made sense of retrospectively in light of their other psychological states. It should not however be taken to imply that akratic agents necessarily endorse what they are doing, as even *pro tanto* good.

lights, we ought to do, there is plenty of room for the kinds of errors and mistakes in reasoning familiar from theoretical contexts. We often neglect to focus on the relevant considerations as the time for action approaches; or we revise our normative judgment under pressure of the temptation to which we are subject, telling ourselves that it's really all right to have just one more drink, watch one more series of plays, smoke one more cigarette, and so on. My point in this section has merely been that irrationality in the practical domain is not necessarily traceable to errors and mistakes of these kinds. We are familiar with a kind of deliberate, self-conscious irrationality in action that has no direct analogue in cases of belief. This is connected with our understanding of our own powers of action as agents, our sense that what we do is in some ultimate way up to us, and it shows that the volitions on which we act are not essentially normative in their commitments.

3. Cleverness and Normative Requirements

It is time to bring these reflections about motivational psychology to bear on the question from which I began in this essay, that of the normativity of instrumental reason. The instrumental principle introduced at the start of this paper enjoins us to take the means that are necessary relative to our ends. A central problem that is posed by this principle is to define the class of ends to which it applies—what we might refer to as the class of ends that are *privileged* with respect to instrumental reason.

On the one hand, it will not do to interpret this class of privileged ends as the ends that are fixed by an agent's desires. There is simply nothing irrational about wanting a given end—even wanting it very strongly—and failing to take the means that are necessary relative to the end. On the

other hand, it equally will not do to interpret the class of privileged ends as fixed by the agent's normative beliefs. Thus, consider a case of *akrasia,* in which you believe that it would be best on the whole to do *x*, but you do *y* instead. In a case of this kind, you of course fail to take the means that are necessary to bring about the ends that you believe it would be best to pursue. Yet it would seem peculiar to characterize the problem in this scenario as a breakdown of instrumental rationality.[40] Akratic agents do not go wrong in failing to take the means that are necessary relative to their ends, but in failing to have the right ends in the first place.

Now, Korsgaard's interpretation of the instrumental principle promises an improved account of the class of privileged ends, building on the basic idea of volitional *commitment.* According to Korsgaard, the instrumental principle tells us to adopt those means that are necessary in regard to the ends we have actively accepted or endorsed, as normative. Her contention is that the force of the instrumental principle—its bindingness on us, as a principle of practical reason—derives from the active nature of the stance of commitment to an end we endorse.[41] To be committed actively in this way is, among other things, to endeavor to bring about the end that one endorses as good, and this orientation of the will requires that one adopt the means that are (believed) necessary with respect to that end. The principle of instrumental rationality is thus construed as a constitutive principle of the will, a principle we necessarily commit ourselves to complying with through the normative act of choice.[42]

Of course, this account of the instrumental principle is couched in terms of the normative interpretation of the will that was criticized extensively above. The present question, however, does not concern the general adequacy of Korsgaard's account of volition, but the conditions for the application of the instrumental principle. Even if I am right that there can be intentional human action in the absence of normative commitment, it might still be the case that means-end rationality gets a grip on us only when the condition of normative commitment to an end of action has been satisfied.

But I believe we should reject this moralizing assumption. The reason is that it neglects the phenomenon I referred to earlier as cleverness, or effectiveness in the pursuit of ends one does not endorse. In cases of *akrasia*, for instance, agents often exhibit great practical intelligence in the pursuit of ends that they do not themselves accept as good—tracking down the one shop in town, for instance,

[40]That is, if instrumental rationality is construed as a matter of adopting necessary means to one's ends. As I explain later, there are broader norms of means-end coherence that often are violated in cases of *akrasia*, and that can help to explain what is ill-advised, from the agent's own point of view, about the course of action that is chosen.

[41]See Korsgaard, "The Normativity of Instrumental Reason," esp. pp. 245-251. I interpret Korsgaard in these pages as arguing for the instrumental principle from the distinctive features of the subjective attitude of commitment to an end that you normatively endorse. John Broome has suggested that Korsgaard is better interpreted as arguing from the assumption that the attitude of intention is (objectively) a reason to do what you intend (see his "Normative Requirements," *Ratio* [new series] 12 [1999], pp. 398-419, sec. 11). But Korsgaard never endorses this objective normative relation, nor would it be particularly plausible for her to do so (since intending an end does not itself seem to give you reason to adopt it). Her "constructivist" argument is best understood as building on the distinctive features of the stance of commitment to realize an end, a stance that Korsgaard believes to involve an element of normative endorsement. My contention will be that this general strategy is on the right track, though Korsgaard's own development of the strategy is flawed: it is not the stance of normative endorsement of an end that introduces rational constraints on our other attitudes, but the distinct and independent stance of commitment to realize the end (a point I believe Broome would also accept, as I go on to explain below).

[42]This is what I referred to in sec. 1, above, as the second moment of normative commitment in Korsgaard's reconstruction of rational agency.

where it is possible to purchase at midnight the bottle of rum they have just decided, against their better judgment, to acquire. Furthermore, this kind of practical intelligence seems correctly characterized as a matter of rationality, relative to the akratic agents' ends. Given their determination to achieve the chosen ends, it seems a requirement and not merely an option that such agents should take the means that are necessary to bring their ends about, and those who fail to do this exhibit a characteristic breakdown of rationality. Indeed, they display a lapse precisely in regard to the instrumental principle. This strongly suggests the need for an account of the normative requirement expressed in the instrumental principle that will apply both to cases in which agents take their ends to be well-grounded, and to cases in which they do not do so.[43]

But the phenomenon of cleverness helps to bring into focus a different aspect of the instrumental principle that is equally significant. This is that the principle imposes rational constraints on the attitudes of agents without entailing either that they have reason to take the means necessary relative to their ends, or that they are rationally required to believe that they should adopt the necessary means.[44] Thus,

[43]Earlier (in note 9) I suggested that even akratic choice could be said to be subjectively normative, insofar as it is prompted by emotional and desiderative states that present options as attractive along some dimension or other. Stephen Darwall has suggested to me that one might appeal to this phenomenon to account for the grip of instrumental requirements even in contexts of *akrasia*, treating akratic actions as actions that are chosen "on the hypothesis" that the chosen end is valuable. But I doubt whether this strategy can render cleverness fully intelligible. If instrumental requirements operate relative to the hypothesis that the end chosen is good or valuable, it is obscure why they should retain their force in cases of *akrasia*, since in these cases the agent precisely rejects the hypothesis that is supposed to be their basis.

[44]This is related to the "bootstrapping" problem identified and discussed by Bratman in *Intention, Plans, and Practical Reason*, pp. 24-27, 86-87; see also his "Intention and Means-End Reasoning," *The Philosophical Review* 90 (1981), pp. 252-265.

in cases of this kind agents do not believe an objective normative relation to obtain between the overall goodness of the ends under pursuit and the necessary means to the attainment of those ends. It is precisely the hallmark of cases of *akrasia* that the agents involved in them do not believe their ends to be the best, and the pressure to adopt the necessary means therefore cannot be accounted for by assumptions about the transmission of normativity across the relation of ends and means. We should accordingly balk at saying either that akratic agents have a normative *reason* for taking the means that are necessary relative to their ends, or that they are rationally required to believe that they have such a reason. On the other hand—and this is the point emphasized above—the fact that an agent has chosen to pursue x introduces rational constraints on their other attitudes and intentions that go beyond the constraints imposed by the mere desire for x, and that are independent of the agent's normative commitments in regard to the desirability of x-ing. These constraints amount to a requirement of instrumental consistency, and what I have referred to as cleverness is a matter of responsiveness to this requirement.

We thus arrive at the following position: an adequate account of the instrumental principle must explain its applicability even to cases of cleverness, but without delivering the questionable conclusion that the deliberate pursuit of an end always yields a reason to take the means that are necessary relative to the end. I now want to sketch the outlines of a response to this problem, proceeding in two steps. The first step concerns the kind of requirement represented by the instrumental principle. We are looking for an interpretation of this principle according to which it imposes constraints on the attitudes of agents, without giving them reason to take

the necessary means to their ends. A natural way to achieve this result is to construe the principle as governing combinations of attitudes—as a normative requirement, in the specialized sense recently distinguished by John Broome.[45] The distinctive features of a normative requirement, in this sense, can be illustrated by considering the relevance of logical principles, such as *modus ponens,* to theoretical reasoning. Suppose you believe p and you also believe $p \rightarrow q$. *Modus ponens* clearly applies to the contents of your beliefs in a situation of this sort, and it therefore imposes constraints on what you should believe. But it would be a mistake to interpret these constraints as licensing you to conclude that you ought to believe q, or even that you have reason to believe q. If for instance your belief that p is itself poorly grounded (by comparison with the considerations that speak against the supposition that q), it may be that the best way to comply with the relevant constraints on belief would be to give up your belief that p, rather than to form the new belief that q. The principle of theoretical reasoning based in *modus ponens,* in other words, functions as a constraint on combinations of your attitudes. We might put this by saying that its normativity is non-detachable, and of wide scope. The requirement could be expressed by saying that you ought to bring about the following: that you believe q, if you believe p and $p \rightarrow q$. It would be incorrect to understand the requirement as

[45]See especially Broome's "Normative Requirements." Gilbert Harman makes a related point about the bearing of logic on theoretical reasoning (*Change in View,* chap. 2). The basic idea that instrumental requirements constrain combinations of attitudes is also implicit in some of the literature on hypothetical imperatives: see, e.g., Stephen L. Darwall, *Impartial Reason* (Ithaca: Cornell University Press, 1983), pp. 14-16; P. S. Greenspan, "Conditional Oughts and Hypothetical Imperatives," *The Journal of Philosophy* 72 (1975), pp. 259-276, secs. 4 and 5; R. M. Hare, "Wanting: Some Pitfalls," as reprinted in his *Practical Inferences* (Berkeley: University of California Press, 1972), pp. 44-58, especially pp. 45-49; and Thomas E. Hill, Jr., "The Hypothetical Imperative," as reprinted in his *Dignity and Practical Reason in Kant's Moral Theory* (Ithaca: Cornell University Press, 1992), pp. 17-37, especially pp. 23-24.

saying, or entailing, that if you believe p, and you believe $p \rightarrow q$, then you ought to believe q.

I submit that we would we do well to interpret the instrumental principle along similar lines, as a constraint on combinations of attitudes that does not license detached normative judgments to the effect that we have reason to take the necessary means to our ends. Thus if you intend to do x, and believe that you can do x only if you do y, then the instrumental principle imposes a normative constraint on your attitudes. You can comply with this constraint either by giving up the intention to do x, or by forming the intention to do y. But it does not follow from the constraint, together with the fact that you intend to do x and believe that you can do x only if you do y, that you ought to intend to do y. In the cases of cleverness that we have been considering, for instance, this would seem to be precisely the wrong thing to say. If the instrumental principle is to apply in the right way to cases of this kind, we will need to understand it as a normative requirement in Broome's sense, imposing strict and non-detachable restrictions on sets of attitudes that include the intention to do x, and the belief that one can do x only if one does y.

This is the first step toward an adequate understanding of the instrumental principle that I announced above. The second step is to elucidate the normative force of the requirement that is embodied in the instrumental principle. Given the understanding of this principle as a strict and non-detachable constraint on combinations of attitudes, what is it that makes it a rational constraint? We can appreciate the task to be addressed here by recalling the comparison of desires with intentions. As I noted above, the fact that you desire that you do x, and believe that you can do x only if you do y, does not seem to have the same rational implications

for your further attitudes as the fact that you intend to do *x*. So there must be something about the attitude of intending to do *x* that goes beyond the attitude of desiring that one do *x*, in a way that brings a distinctively rational requirement into play.

To understand what this additional distinctive feature of intention might be, it may help to return to Korsgaard's elucidation of the instrumental principle. On her account, as we have seen, the ends to which the instrumental principle applies are those that agents have actively committed themselves to realizing, where the stance of active commitment is in turn taken to involve an attitude of normative endorsement. There are two distinct parts to this account. One is the idea that normative endorsement of an end by the agent is a condition for the applicability of the instrumental principle. The second is the volitionalist idea that the ends to which the principle applies are ones to whose realization the agent is actively committed. Korsgaard evidently supposes that these two ideas stand or fall together, contending that the stance of commitment to an end must be cashed out in terms of the different notion of endorsement of a normative principle. But I have argued that Korsgaard is wrong to link these ideas in the way she does: the volitionalist capacity for the distinctive stance of commitment to realize an end can be exercised independently from the kind of normative endorsement to which Korsgaard appeals. Once we are clear about this, perhaps we can define in terms of the notion of active commitment the privileged class of ends to which the instrumental principle applies.

This suggestion certainly has an air of plausibility about it. If you are actively committed to doing *x*, in the way characteristic of the stance of intention, then it seems that you must also be committed to taking the steps that you believe to be necessary to your doing *x*, on pain of irrationality. The

question, however, is this: *why* would you be irrational if you failed to intend the means that you believe to be necessary, relative to some end that you intend to realize? What in particular is it about the stance of commitment to realize the end that makes this combination of attitudes rationally required, even in the absence of normative endorsement? Until we have answered this question, we will not really have explained the force of the normative requirement embodied in the instrumental principle.

Now, Broome has proposed a straightforward response to this problem.[46] The normativity of the instrumental principle, he suggests, may be traced to the very same logical constraints that underlie the requirement to adjust one's beliefs in accordance with *modus ponens.* Suppose you intend to do *x*, and believe that *y* is a necessary means to bringing it about that you do *x*. In this scenario, the objects of your intention and your instrumental belief can be represented as propositions to which *modus ponens* applies, as follows:

(a) You will do *x*
(b) For you to do *x*, it is necessary that you do *y*.

Modus ponens tells us that these two propositions can be true only if a third proposition is also true, namely:

(c) You will do *y*.

The validity or truth-preserving character of this pattern of inference is what grounds the rational requirement that we should adjust our beliefs in accordance with it. Beliefs are essentially attitudes toward the truth of propositions, as we saw in section 2, above; the internal aim of the attitude of

[46]See John Broome, "Practical Reasoning," in José Bermùdez, ed., *Reason and Nature: Essays in the Theory of Rationality* (Oxford: Oxford University Press, forthcoming); also his "Normative Requirements," especially sec. 6.

belief, as we might put it, is to track the truth. It follows that we are not rational in our beliefs if we do not arrive at them in a truth-preserving way, and this explains the relevance of logical principles such as *modus ponens* to theoretical reasoning.

But Broome suggests that intentions are equally attitudes toward the truth of propositions. They are not, to be sure, attitudes whereby we take propositions to be true; but they can be understood as attitudes whereby we are set to *make* propositions true. Once we grasp this point, Broome contends that we will see the direct relevance of the very same logical principles to practical reflection that involves intentions and beliefs about necessary means to our ends. If you intend to bring it about that (a) is true, and you believe that (b) is true, then you must intend to bring it about that (c) is true on pain of irrationality. For to fail to do this is to thwart the internal aim that is constitutive of your initial state of intention—the aim, namely, of bringing it about that (a) is true.

This is an elegant response to the present problem. It builds on the plausible idea that intention involves a commitment to bringing something about, cashing this out in terms of the attitude of being set to make a proposition true, in a way that renders considerations of validity immediately relevant to the assessment of the rationality of intentions. On closer inspection, however, it may be wondered how much this proposal really explains. One difficulty with it is that it does not seem to distinguish in the right way between our attitudes toward the means that are believed necessary to realize our ends, and our attitudes toward necessary but unintended consequences or side-effects of our pursuit of our ends. In the schema above, (b) could characterize equally either of these two relations. For instance, given my end of getting to work in time for my 11:00 a.m. class, I may

believe both that it is necessary that I set off in the car before 10:30 a.m., and that I crush numerous acorns (which are littering the street at this time of year).[47] Following Broome, if I am set to make it true that I get to work in time for my 11:00 a.m. class, then these two beliefs of mine would seem to require that I be set to make it true both that I drive off in the car before 10:30 a.m., and that I crush numerous acorns. But it would be odd, to say the least, to suggest that an intention to do the latter is rationally required by the other attitudes that have just been ascribed to me.

Broome responds to this difficulty by distinguishing between things that are necessary as consequences of my *x*-ing and things that are necessary as means to my *x*-ing.[48] Items of both kinds fall within the purview of the attitude of being set to make true. In the example just sketched, for instance, given my beliefs and my initial intention, I must be set to make it true both that I set off in the car before 10:30 a.m., and that I crush numerous acorns. But Broome suggests that not all the things that I am set to make true are things that I intend. If I intend to *x*, then I must rationally be set to make true anything that I believe to be necessary to my *x*-ing; but I must intend only those things that are believed to be necessary as *means* to my *x*-ing. But this proposal raises two new and very challenging questions. First, how are we to distinguish necessary means from necessary consequences of our actions? And second, how are we to distinguish the distinctive attitude of intending that *p* from the more generic attitude of being set to make it true that *p*? Broome's initial appeal to the idea of being set to make something true looked to be a way of elucidating the distinctive stance of intention, showing how it goes beyond the

[47]It is an interesting question how the notions of possibility and necessity relevant to instrumental rationality are to be understood; I come back to this issue in section 4.

[48]See "Practical Reasoning," sec. 3.

attitude of merely desiring that p, in a way that renders transparent the relevance of logical principles to the assessment of the rationality of our intentions. But now we discover that being set to make something true is not sufficient for intending it: what then is the further feature of intentions that distinguishes them from cases in which we are set to make something true without intending it?

4. The Cognitive Conditions of Intention

Perhaps satisfactory answers to these questions can be devised. But I do not wish to pursue that issue here, because it seems to me that there is an alternative account of the instrumental principle available, one whose plausibility does not depend on such controversial matters as these. We may begin by returning to the comparison of intentions with desires. I have already suggested that to intend an end is to be committed to realizing it or bringing it about, in a way that goes beyond the attitude involved in merely desiring the end. One specific respect in which these attitudes differ is that the commitment to realize an end is constrained by one's beliefs about the possibility of realizing the end, whereas desires are not similarly constrained. One could want the process of global warming to stop immediately, without believing that this is so much as possible, given the rest of what one believes about the current state of the world. But intentions are different in this respect. Some philosophers have gone so far as to suggest that intentions presuppose (or may in part be identified with) the belief that what one intends will in fact come to pass;[49] but this thesis is

[49]See, for example, Gilbert Harman, "Practical Reasoning," as reprinted in Alfred R. Mele, ed., *The Philosophy of Action* (Oxford: Oxford University Press, 1997), pp. 149-177. The basic account of instrumental rationality that I am about to offer owes much to Harman's discussion of practical reasoning in this paper, and in his *Change in View*. But of course I work out the basic strategy in a way very different from Harman, starting (for instance) from much weaker, and hence more plausible, assumptions about the cognitive conditions of intention.

controversial, and anyway unnecessary to account for the principle of instrumental rationality. It will suffice to maintain what is at any rate more plausible, namely that the intention to do x requires at least the belief that it is *possible* that one do x.

It is worth pausing to compare this suggestion to the assumptions implicit in Broome's account of the instrumental principle. Broome suggests, in effect, that the stance of intending to do x is rationally answerable to considerations regarding the possibility of one's x-ing. To intend to do x is to be set to make true the proposition that one x's, and this internal aim of intention will be thwarted if the proposition in question cannot be true, given the rest of what one believes. There is an assumption here about the possibility of realizing the aim of intention; but the assumption is made by the theorist of instrumental rationality, not necessarily by the agent. Given the characterization of the internal aim of intention, the theorist assumes that this aim will be thwarted if the propositional object of intention cannot be true together with other propositions that the agent believes, about necessary means. By contrast, I have suggested that agents who intend to do something must *themselves* believe that it is possible for them to do what they intend. The difference between these proposals can be brought out by reflecting on the attitudes of agents who firmly believe that they cannot do something (say, x). For all Broome says, it might be the case that such agents intend to do x, though their intentions would then be irrational. By contrast, if I am right in suggesting that intention presupposes the belief that it is possible for one to do what one intends, then agents who believe that they cannot do x should not even be described as in-

tending to do x in the first place.

Once we are clear about this, it seems to me that intuition and reflection support the proposal I have put forward.[50] For it seems to me that we simply do not describe people as intending to do x in circumstances in which it is clear that they do not believe it to be possible for them to do x. Nothing the agent might do under these circumstances, we might say, would correctly be described as intending to do x.[51] If this is on the right lines, however, it suggests a different account of the normative force of the instrumental principle. For consider now a situation in which the following attitudes can be ascribed to you: you intend to do x, you believe that your doing y is necessary if x is to be brought about, and you believe that you will do y only if you intend to do y.[52] Given that the intention to do x brings with it the belief that it is possible for you to do x, your further beliefs about y-ing and its relation to your doing x entail that you will be subject to an incoherence in beliefs if you do not either abandon the original intention to do x, or adopt a new intention to do y. Failing to take either of these steps, you will be left in effect with the following incoherent set of beliefs (assuming you are minimally self-aware): the belief that it is possible that you do x, the belief that it is possible that you do x only if you also intend to do y, and the belief that you do not intend

[50]I do not mean that Broome would necessarily disagree—he does not take a stand on the issue I have raised, nor does his account of instrumental rationality require him to do so. But if he agrees with me about this point, then he should be open to the suggestion that intentions introduce rational constraints on our attitudes that operate via independent rational constraints on coherence of beliefs.

[51]This is one of the lessons of Albritton's "Freedom of Will and Freedom of Action."

[52]I do not mean, of course, that these beliefs must be explicitly and articulately present to the consciousness of the agent who is engaged in instrumental reasoning, only that they are implicit in the agent's understanding of their situation and their possibilities for action.

to do y. The incoherence of these beliefs is a straightforward function of the logical relations among their contents, suggesting that the normative force of the instrumental principle can be traced to independent rational constraints on your beliefs—in particular, to constraints on certain combinations of beliefs (a normative requirement, in the specialized sense discussed above).

To this it will be replied that theoretical constraints on rational belief formation cannot by themselves account for the requirement to intend to do y, even given the intention to x and the other beliefs about the relation between x-ing and intending to y specified above.[53] Granted those other attitudes, you will be subject to an incoherence in belief of you do not form the belief that you intend to do y. But from the rationality of this belief it does not follow that it would be rational for you to intend to do y. Theoretical constraints on rational belief can get you only as far as the *belief* that you intend to do y; to go beyond that, to a rational requirement that you form the intention to y, we need an additional principle of distinctively practical reason. Moreover, without an additional principle of this kind, we will not have succeeded in accounting for the force of the instrumental principle. This can be seen by considering the (slightly comical) situation in which you intend to do x, and believe that your intending to y is necessary to your doing x, but mistakenly or wishfully come to believe that you intend to y. In this scenario you seem to have satisfied fully the constraints on rational belief spelled out above. But we should presumably want to say that you have not yet complied with the instrumental principle, which remains in force and says that your

[53]I am grateful to John Broome and a second reviewer for *Philosophers' Imprint* for pressing me to think about the line of objection sketched in this paragraph. Compare Bratman, "Intention and Means-End Reasoning," pp. 255-256 (note 4).

attitudes are not in rational order so long as you do not in fact form the intention to do y. Conversely, a situation is imaginable in which you fully comply with the demands expressed in the instrumental principle while violating the requirements of theoretical reason appealed to above. Thus you might intend to x, and believe that your intending to y is necessary to your x-ing, and also intend to y; but if you mistakenly or carelessly believe that you do not intend to y, you will be subject to the kind of incoherence in belief characterized above.

This line of objection turns on the possibility of divergence between your intentions and your beliefs about what you intend. But what, in practice, does this possibility amount to? Ordinarily, future-directed intentions are fairly accessible to consciousness. This is connected to the distinctive role that intentions play in shaping our deliberations. Someone who is entirely unaware of their alleged intention to do x, or who has completely forgotten that this is what they intend to do, cannot really be described as having the intention to do x any longer. In what would the distinctive commitment to bringing it about that one x's consist, under the circumstances described? We can imagine an unconscious *desire* that is expressed in some course of behavior that a person performs, fixing the real point of the behavior (as opposed to the account of it that the agent would be inclined to provide when asked). But a fully unconscious desire cannot amount to the kind of commitment to realize a plan of action that is represented by intentions, for the simple reason that it is inaccessible to consciousness, and so incapable of attaining the functional features implicit in talk of *commitment* (as opposed to mere desire).[54] By the same token, it is equally difficult to imagine a scenario under which

[54]I have in mind here the functional features of intention stressed by Bratman, in *Intention, Plans, and Practical Reason*.

you sincerely believe that you intend to do x, while utterly failing to have this intention in fact. Your belief about what you are committed to—if it is sustained and sincerely held over time (as opposed to being a passing vivid impression or hallucination)—will itself start to shape your deliberations and reflections in the ways characteristic of intention.

In light of these considerations, I would suggest the following answer to the objection sketched above. First, assume, as I have been suggesting, that intentions are readily accessible to consciousness. In this situation it would indeed seem independently irrational for you to have false beliefs about the content of your intentions. You may of course be momentarily distracted or forgetful, but reflection at the level of minimal self-awareness can bring you to see or recall what you are really intending to do. It follows that you will be subject to rational criticism if you believe that you intend to do y without really so intending, or if you form such an intention without believing that that is the case. Indeed, the rational requirements at issue here are sufficiently stringent that they constrain our attribution of intentions to agents, in the ways sketched in the preceding paragraph. But this is enough to plug the gap in the argument to which I earlier called attention. The objection was that theoretical considerations of coherence of belief merely require you to believe that you intend to y, if you intend to x and believe that your intending to y is necessary to your x-ing; they cannot require you actually to form the intention of doing y. But if you can rationally believe that you intend to y only if you in fact intend to y, then rational requirements can indeed bring you to form this intention. Under the circumstances described—where you intend to do x, and believe that you can do x only if you intend to do y—the only rational way to bring your beliefs into coherence is to commit yourself to doing y, in the way we have seen to be characteristic of intention.

Assume, next, that there can be intentions that are not

readily accessible to consciousness. I have suggested that we would have trouble making sense of intentions—commitments to realize a plan of action, as opposed to mere desires or wishes—that operate at the level of the unconscious. But suppose for the sake of argument that I am wrong about this. In the situation where this is the case, it would not seem irrational for you to have false beliefs about the content of your intentions, and my argument for instrumental rationality would consequently fail to get a grip. But equally, it seems to me doubtful that intentions that are cut off in this way from conscious belief really do introduce rational constraints on our further attitudes, of the kind represented by the instrumental principle. Someone who has an unconscious intention to (say) avenge an imagined childhood slight would not seem irrational if they fail to take the means that they (unconsciously?) believe to be necessary to that end. The postulated inaccessibility of the intention to consciousness already itself prevents the attitude from behaving in the ways characteristic of intention, and this makes it seem odd or mistaken to suppose that it introduces further rational constraints on one's attitudes. In this respect, the imagined scenario involving unconscious intentions differs markedly from the case of akratic intentions that has been my primary concern.

These considerations support the account of the instrumental principle that I have been developing. That account does not represent the principle merely as an application of the considerations that determine coherence and consistency relations amongst beliefs. The account additionally makes assumptions about the relations between intentions and beliefs: first, that to intend to *x* is to believe that it is possible that one do *x*; second, that one can rationally believe one intends to *x* only if one really does intend to do *x*; and third,

that if one intends to *x*, it is rational for one to believe that one intends to *x*. But these further assumptions seem independently plausible, as I have endeavored to show above. Furthermore, when we try to imagine a scenario in which the further assumptions do not hold (such as the scenario involving unconscious intentions whose connections to self-awareness have been severed), we lose our grip on the basic idea that intentions introduce the constraint on our further attitudes that is expressed by the instrumental principle.

To be sure, this core requirement is not the whole of what we ordinarily think of under the rubric of instrumental rationality. The requirement corresponds roughly to the analytic principle of rational willing that Kant introduces in the second section of the *Groundwork* to explain the possibility of hypothetical imperatives: "Whoever wills the end, wills (so far as reason has decisive influence on his actions) also the means that are *indispensably necessary* to his actions and that lie in his power."[55] Beyond this core requirement, however, there are broader norms of means-end coherence that can be brought to bear in assessment of action.[56] Practical reason does not enjoin us merely to take the means that are absolutely necessary to realize our ends but also to take those optional means that would facilitate the realization of our

[55]This is from p. 417 in the pagination of the Prussian Academy edition (emphasis mine); the translation used is James Ellington, *Grounding for the Metaphysics of Morals* (Indianapolis: Hackett Publishing Co., 1981).

[56]I borrow the expression "means-end coherence" from Michael Bratman (see his *Intention, Plans, and Practical Reason,* pp. 31-33). I should emphasize that nothing in Bratman's discussion of this phenomenon supports the suggestion I go on to discuss in the text, namely that means-end coherence makes sense only against a background of normative endorsement of one's ends. If one could account for this broader phenomenon in non-moralizing terms, then a unified treatment of the whole of means-end rationality might be possible. But Bratman does little to indicate what a non-moralizing account of means-end coherence might look like, nor does such an account seem very promising to me.

overall system of plans and values. In order to produce a decent paper for the conference that is coming up in the summer, it may not strictly be necessary that one start working on it tomorrow, but that might still be a good idea, given one's overall set of aims and ambitions.

Means-end coherence of this kind, to the extent it goes beyond the core demand that one adopt those means that are necessary relative to one's ends, probably does presuppose that one endorse the ends that are in question. Thus akratic agents typically do not have an eye to the coherence, in this sense, between their immediate activities and their larger system of projects and values. Indeed in many typical cases what makes akratic activities ill-advised, from the agent's own point of view, is precisely the fact that they do not cohere well with the agent's larger system of projects and values. (One realizes that it would be best to get started on one's paper for the summer conference, and yet one goes out to a movie instead.) Practical requirements of means-end coherence may thus apply only to those ends that the agent has endorsed, as good or worth pursuing on the whole. To the extent this is the case, the theory of instrumental rationality will not be a monolithic subject, and the account I have sketched of the instrumental principle should not be interpreted as a complete account of what is customarily understood as instrumental or means-end rationality.

But we should not infer from this important insight that the instrumental principle, in the form I have discussed, is of merely minor or secondary significance. Granted, the principle applies only to cases involving means that are believed to be necessary for the attainment of our intended ends. This might seem to entail that it applies very rarely, since it is seldom the case that the means we adopt are strictly necessary to the realization of our ends—necessary, that is, in the sense of logical or physical necessity. But the principle

allows for natural extensions to cases that do not involve strict necessity of these kinds.

Consider the quotidian example introduced earlier, in which I intend to get to campus by 11:00, believe that I can do so only if I set out in the car by 10:30, and believe further that I can set out in the car by 10:30 only if I intend to do so. The beliefs about necessary means involved in this example are almost certainly false if interpreted in the sense of physical or logical necessity. Thus I might hold these beliefs while lucidly conceding that it is both logically and physically possible that I will arrive at work by 11:00, even if I do not set out in the car by 10:30—a helicopter might come by at 10:50, for instance, and deposit me at my office ten minutes later. The point, however, is that nothing in my overall set of beliefs gives me any reason at all to think that this will happen.

This suggests that the notions of necessity and possibility relevant to the instrumental principle are epistemic notions, determined by our background assumptions about developments in the world and the way those developments will shape our concrete options for action. In deliberating about what to do we take many parameters as fixed, such as the fact that helicopters will not descend and whisk us off to the office; even if they are not ruled out by logical principles or the laws of physics, such occurrences seem so unlikely that we do best to proceed on the assumption that they are simply off the table. The notions of necessity and possibility at work in practical deliberation are thus notions of what is necessary and possible in the strict sense, given a host of unspoken assumptions about what is going to happen in the world.

Background assumptions of this kind may in turn have a variety of distinct sources, extending beyond purely epistemic considerations relating to the independent probability of certain events. They might, for instance, be based in prior

plans or decisions made by the agent. If in addition to the commitment to get to the office by 11:00 a.m., I also have a standing intention to commute to campus with my car, then this will exclude from deliberative consideration methods of technically possible locomotion that do not involve my driving. To adopt such means would be incompatible with the belief that it is possible that I take my car to work, where that belief accompanies, in ways already canvassed, the intention to commute to campus with my car. Another source of constraints on the standpoint of deliberation are normative views that are held by the agent: thus, my assumption that I will get to work on time only if I drive my car takes it as fixed that I will not steal my neighbor's car. This is an option that is ruled out not by logic or the laws of physics but by morality, and its incompatibility with moral justification may in turn make it reasonable for me to conduct my deliberations on the assumption that the option is simply off the table. Sometimes the fundamental rationale for such assumptions will again be an epistemic one—for instance, the utter improbability of my stealing my neighbor's car. This event is of the kind that is under my intentional control, so what makes it improbable is my own view about the moral impermissibility of casual theft, together with the fact that I am appropriately responsive to considerations of this kind.[57] In other cases the rationale for the assumption that stealing the car is off the table may be more directly a normative one. In these cases, it is not the fact that I take my stealing the car to be practically improbable that renders it irrelevant to my deliberation, but the different fact that I take it to be normatively out of bounds.

These remarks only scratch the surface of a large and complex subject. But I hope that enough has been said to show how the instrumental principle, as I have interpreted

[57]Compare Bernard Williams, "Moral Incapacity," as reprinted in his *Making Sense of Humanity* (Cambridge, England: Cambridge University Press, 1995), pp. 46-55.

it, can be extended to cover a wide range of cases involving means that are not without qualification necessary for the attainment of our ends. The principle may not account for the whole of means-end rationality, but it expresses a core requirement that is both central to our deliberative experience and applicable, as I have shown, to situations in which agents are involved in the pursuit of ends that they do not themselves endorse. Furthermore, the derivation of this core requirement from theoretical principles governing belief formation helps to make intelligible its ready extension to a range of cases involving means that are not strictly necessary for the attainment of our ends, but that are necessary only relative to background epistemic and normative assumptions that shape our deliberative point of view. Our belief as agents that it is possible for us to do what we intend serves to mediate between these background assumptions and our beliefs about the strict necessity of various means, in ways that seem to accord with our deliberative experience. If by contrast we follow Broome in deriving the instrumental principle from considerations involving the internal aim of intention, it becomes harder to see how the principle could admit of natural extensions beyond the most narrow cases. The normativity of the core instrumental principle, on his account of it, stems from the idea that the pursuit of ends that it is impossible to realize would thwart the constitutive aim of intention. But the notion of possibility relevant to this account is the theorist's notion of possible truth, not the agent's understanding of what it is possible to achieve; it is accordingly obscure on Broome's account how the core instrumental principle could be extended to cases that do not involve strict logical or physical necessity, in the ways that come perfectly naturally to us when we deliberate about what we are to do.

Finally, the account I have presented seems to explain

how automatically the core requirement of instrumental rationality tends to be complied with as we execute our intentions and plans. As we saw in section 2, above, there is less scope for deliberate irrationality in the sphere of belief than there seems to be in the sphere of action, and this reflects itself in the difficulty we have imagining a willful violation of the core requirement of instrumental reason.[58] When people believe that an available means is necessary relative to one of their alleged ends, but fail to adopt that means, we tend to question whether they are really committed to the end after all. The core requirement of instrumental reason thus functions more like a constraint on interpretation than do other principles of practical reason. This is one of several features of the core requirement that point toward the account of it I have developed in this section, according to which the requirement derives from basic theoretical constraints on the coherence of beliefs.[59]

[58]This difficulty was brought home to me by some remarks of Brad Hooker's.

[59]Earlier versions of this paper were presented to the Jowett Society of Oxford University, to the Berliner Workshop zur praktischen Vernunft, to the Philosophy Department of the University of California at Berkeley, to the section on metaethics and methodology of the Netherlands School for Research in Practical Philosophy in Utrecht, and to a workshop on moral and social philosophy at the Australian National University; many thanks to the audiences on all these occasions for very helpful feedback. A list of the numerous people whose probing and constructive questions led to improvements would have to include at least the following: Karin Boxer, Jan Bransen, Ruth Chang, Stephen Darwall, Hannah Ginsborg, Brad Hooker, Joanna Perkins, Joseph Raz, Samuel Scheffler, and Theo van Willigenberg. Lengthy written comments by John Broome and a second reader for *Philosophers' Imprint* were especially helpful, and prompted extensive changes during the last round of revisions. My interest in the phenomenon of cleverness grew out of memorable discussions with Michael Smith held in Princeton in the 1980's.