Justification and Alienation

—draft—

Pamela Hieronymi
hieronym@ucla.edu
November 28, 2004

The past twenty to thirty years have seen an ongoing discussion of the psychological demands of morality: what demands does morality put upon us, psychologically, and can it legitimately do so? A handful of philosophers—most notably Bernard Williams—have claimed that morality demands too much. The worry, however, is not simply that morality is too stringent—morality's defenders may well accept its stringency. Nor, even, is the worry that morality may require remaking a person's character into morality's image. Morality may well require an overhaul of the morally reprobate. Rather, the worry on which I will focus is that the demands of morality do not allow one to be psychologically coherent—its demands do not address a single agent. To appropriate a phrase of Williams, the demands of morality can seem to be, "in the most literal sense, an attack on [the agent's] integrity."[1]

Williams, famously, made his argument by appealing to the particular psychology of each agent—to the agent's "ground projects," to what, for her, makes life worth living, or what constitutes her "subjective motivational set." According to Williams, the reasons on which an agent acts must find footing in these personal concerns and projects. These concerns and projects provide the motivational source of action. By addressing the agent from an impersonal or impartial point of view—a point of view other than that provided by our particular cares and personal projects—morality misses this motivational source. The difficulty is most clearly displayed in cases in which the impartial demands of

[1] Bernard Williams, "A Critique of Utilitarianism," in J. J. C. Smart and Bernard Williams, *Utilitarianism: For and Against* (Cambridge: Cambridge University Press, 1973): 117.

morality conflict with our more personal cares.[2] In such cases of conflict, on Williams' view, morality asks us to do the psychologically absurd: to alienate ourselves from the things that we think make life worth living and to act instead in accord with its impersonal, yet authoritative, demand. These episodic conflicts reveal the deeper and more systematic difficulty with moral demands—that they are, quite generally, wrongly related to the motivational bases of our actions. Sometimes the issue is put as a problem about the kind of reasons on which morality must rely. Morality makes its appeal by use of what are called "external" reasons, reasons which bear no necessary connection to our own concerns and projects. Such reasons, it is argued, will never explain anyone's action, and so are not ever really anyone's reasons for acting, and thus are not rightly considered reasons at all. By relying on them, morality forfeits its claim to be action guiding.[3]

In this paper, I hope to present a quite different way of understanding the worry that a moral theory may, in a literal sense, attack an agent's integrity—that moral theory may

[2] Williams considers a number of examples of conflict. In perhaps the most famous, a person is faced with many people in need of rescue, and chooses to rescue his spouse. Williams believes that concern for the requirements of impartial morality, in such a case, is out of place. It provides what he calls "one thought too many." (He says, of this case, "it might have been hoped by some (for instance, by his wife) that his motivating thought, fully spelled out, would be the thought that it was his wife, not that it was his wife and that in situations of this kind it is permissible to save one's wife." Bernard Williams, "Persons, Character, and Morality," *Moral Luck* (Cambridge: Cambridge University Press, 1981): 18.)

[3] The foundational papers in this discussion belong to Bernard Williams: "Morality and the Emotions," *Problems of the Self* (Cambridge: Cambridge University Press, 1973): 207–229, originally the Inaugural Lecture at Bedford College, London, 1965; "A Critique of Utilitarianism," in J. J. C. Smart and Bernard Williams, *Utilitarianism: For and Against* (Cambridge: Cambridge University Press, 1973): 75–150; "Persons, Character, and Morality," *Moral Luck* (Cambridge: Cambridge University Press, 1981): 1–19, first published in *The Identities of Persons,* ed. A. O. Rorty (Berkeley: University of California Press, 1976); "Internal and External Reasons," *Moral Luck* (Cambridge: Cambridge University Press, 1981): 101–113, first published in *Rational Action,* ed. Ross Harrison (Cambridge: Cambridge University Press, 1980); *Ethics and the Limits of Philosophy* (Cambridge: Harvard University Press, 1985). See also Michael Stocker, "The Schizophrenia of Modern Ethical Theories," *Journal of Philosophy* 73 (1976): 453–66; Susan Wolf, "Moral Saints," *Journal of Philosophy* 79 (1982): 419–259; "Morality and Partiality" *Philosophical Perspectives* 6 (1992): 243–259; "Meaning and Morality" *Proceedings of the Aristotelian*

ask an agent to act from some alienated point of view, or may fail to address a single moral agent. The form of alienation I hope to highlight does, in fact, appear when one acts on a certain class of reasons, which I will call "extrinsic" reasons. Unlike Williams' external reasons, extrinsic reasons are marked out, not by their relation to the agent's particular cares and concerns—her psychology or motivational set—but, rather, by their relation to the action they recommend. We will see that, although these extrinsic reasons count in favor of performing certain actions, they are not reasons for which you can directly perform those actions. Even if you find these reasons fully persuasive, you cannot, by responding to them, perform the action of which they count in favor. Rather, these are reasons for which you can, at most, act upon yourself to bring it about that you perform the action in question. Acting upon yourself in this way is the familiar (admittedly quite mild) form of alienation I want to highlight. Further, once we have isolated extrinsic reasons and examined how they generate this mild form of alienation, we will be able to see that relying on extrinsic reasons in the kind of justification moral theory hopes to provide will create a more extreme form of alienation, and so generate a problem about the integrity or coherence of the agent. If moral theory is to address a single agent, it must take care to avoid relying on extrinsic reasons. Unfortunately, determining exactly which considerations are extrinsic to the actions morality requires is no easy task.

Society 97 (1997): 229–315; and Peter Railton, "Alienation, Consequentialism, and the Demands of Morality," *Philosophy and Public Affairs* 13 (1984): 134–171.

KINDNESS AND ITS REASONS

In this section and the next, I hope simply to display the justificational structure that creates the mild form of alienation I have in mind—to explain how certain reasons can genuinely count in favor of an action while not being reasons for which one can directly perform that action, but rather only reasons for which one can act upon oneself, to bring it about that one performs the action of which they count in favor. I will illustrate this justificational structure and its limitations by appeal to an example—the example of a kind helping action. (Agreement with the details of this example is not particularly important, as the example is meant simply to illustrate a possible justificational structure and the difficulties that structure presents.)

The justificational structure in question is possible for actions which are identified by the reasons for which they were performed.[4] I take it that kind actions are so identified. While an action may qualify as, e.g., a *helping* action, independently of the reasons for which it was performed—simply in virtue of easing another's burdens—whether a helping action is also *kind* depends on the reasons for which one decided to help. If you help simply in order to assure the job is done competently, then that helping might be conscientious, but is not necessarily kind. Helping may be an act of kindness, or conscientiousness, or fair-mindedness, or prudence, or self-promotion, or spitefulness, depending on the reasons one took to count in favor of helping. A helping action is kind just in case it was performed for certain reasons, and not others.

[4] I think the structure appears, in general, for any activity which is identified by the reasons for which it is done or the reasons which it makes one answerable for. Cf. my "Controlling Attitudes," *Pacific Philosophical Quarterly* (forthcoming)..

Turning, then, to the reasons for helping kindly: As we have seen, some reasons (such as assuring the job is done competently) count in favor of helping, without touching on kindness. Notice, though, that the reasons which count in favor of performing a specifically *kind* helping action are of two quite distinct varieties. Certain reasons count in favor of helping and are such that, if you decide to help for those reasons, your helping will thereby qualify as kind. Relieving your colleague's exhaustion counts in favor of helping, and, presumably, if you decide to help in order to relieve your colleague's exhaustion, your helping will thereby qualify as kind. I call such reasons the *reasons constitutive of* a kind helping action.

Importantly, though, there are reasons which count in favor of performing a (specifically) *kind* helping action which are not among those that would qualify a helping action as kind. That helping kindly would improve your reputation or ease your sense of guilt could count in favor of helping kindly. Yet these are not reasons which would qualify your helping as kind. I call these "extrinsic reasons" for the kind action. Extrinsic reasons count in favor of performing a (specifically) kind action simply by showing something good about performing such an action, *without* being among the reasons that would qualify the action as kind.

Extrinsic reasons for kindness appear because we are reflective creatures, capable of thinking about the quality of our own actions, and it often matters to us whether we perform kind ones. One's career aims, or one's guilty conscience, or the criticism of a perceptive family member can all bear on whether it would be good to act kindly, without being reasons which would qualify an action as kind. In fact, it seems that for any action identified by the reasons for which it is performed, it will be possible, with some

imagination (and perhaps science-fiction), to construct some case in which one has other reasons for performing such an action.

Without specifying which reasons, in particular, would qualify an action as kind (relying instead on our rough, commonsense intuitions about which considerations, in usual circumstances, would be constitutive of kindness), we can consider some general features of the distinction between constitutive and extrinsic reasons.

Notice, first, that the particular consideration in question—the content of the reason, so to speak—will not by itself determine whether a reason is constitutive of kindness. Relieving need and protecting vulnerability seem to be reasons constitutive of kindness. But an action is not kind just because it takes the needs and vulnerabilities of others as reason-giving. After all, malicious and cruel actions also take the needs and vulnerabilities of others as reason-giving. One might relieve a need out of spite, knowing that relieving this need will humiliate the needy. So, whether an action qualifies as kind will depend not only on *which* considerations one took to count in favor of helping—need or vulnerability or reputation—but also on *why* or *how* one so took them; it will depend on the reasons and background assumptions which explain one's doing so. Another person's need, taken as reason-giving, will be constitutive of kindness only if they are so taken in a certain justificational context. I will call this justificational context, together with the considerations whose being-taken-as-reasons it explains, the person's "operative justification" for the action. Assessing whether a particular action is kind will require assessing its operative justification. Certain operative justifications are constitutive of

kindness, others are not. Reasons are constitutive of kindness only if they are appear in operative justifications constitutive of kindness.[5]

We can now define extrinsic reasons as those which count in favor of performing an action from a certain operative justification, without themselves being a part of that justification. Thus, extrinsic reasons cannot be taken to lend support or justification to the constitutive reasons: if a reason is extrinsic to kindness, it cannot be taken to show that the reasons constitutive of kindness are reason enough to act. If, e.g., you take the fact that performing a kind helping action would improve your reputation to show that someone's exhaustion calls for help, your action will thereby fail to kind. You would be taking her exhaustion to be a reason to help because of concerns about your reputation, and so you action would be self-serving, not kind. Once an extrinsic consideration becomes the reason for which one takes an otherwise constitutive reasons to be reasons to act, the otherwise constitutive reasons are no longer constitutive of kindness. This is because any reason which is taken to show that another reason is reason enough to act will be part of the operative justification of the action performed, but extrinsic reasons are (by definition) not part of the operative justification of a kind action. By taking the

[5] I call these "operative justifications," as an extension of T. M. Scanlon's term "operative reason." Scanlon distinguishes "operative" reasons from reasons "in the standard normative sense" in his *What We Owe to Each Other* (Cambridge MA: Belknap Press, 1998): 18–20. A reason, in the standard normative sense, is, roughly, a good reason, a consideration that genuinely counts in favor of something. An "operative" reason is a consideration someone took to be a reason, and which thereby explains his action or attitude. Derek Parfit, in his *Rediscovering Reasons* (unpublished manuscript) makes the same distinction, but uses the term "motivating" where Scanlon uses "operative."

An operative justification gives a more or less complete account of the person's reasons for adopting a particular attitude or performing a certain action. (An operative justification may well be a "maxim," or subjective principle of action.) I hope it is clear, here, that the operative justification underlying a person's action or attitude can be implicit, inarticulate, even unconscious. A person can certainly be self-deceived about her own operative justifications. It may be indeterminate what a person's operative justification was. Finally, I don't think the force of the point depends on whether we can find psychological structures or processes that correspond to the "implicit" justification. It may well be that features of the justification are determined as much by features of the person's context as by features of the person's psychology. (Cases

extrinsic consideration to support the otherwise constitutive reason, one changes the operative justification of the action, and so spoils its kindness. The extrinsic consideration provides one reason too many.

Extrinsic Reasons and Alienation

Extrinsic reasons, then, cannot be taken to show that reasons constitutive of kindness bear sufficiently on whether to help.[6] If one takes extrinsic considerations to show that otherwise constitutive reasons are reason to help, then those otherwise constitutive reasons are no longer constitutive of kindness. I will now argue that extrinsic reasons therefore are not reasons for which you can directly perform the action of which they count in favor. They are rather reasons for acting upon yourself to bring it about that you perform that action for other reasons. Acting upon yourself in this way—to bring it about that you act for other reasons—involves a kind of mild alienation or split in one's point of view.

Consider first a case in which you do not already find the constitutive reasons convincing, but you are convinced by extrinsic reasons. Although you are unmoved by your colleague's exhaustion, you would like to be more kind in your interactions with your colleagues, because you think that doing so will advance your career aims. Your interest in kindness is extrinsic; your concern is strategic and prudential, not kind. Suppose you now see an opportunity to help one of your exhausted colleagues and

---

of ignoring or failing to notice might be examples of this.)

[6] This formulation must be careful: I have said that extrinsic reasons, as a matter of definition, "cannot be taken *to show* that the constitutive reasons bear . . ." rather than that they cannot be *reasons for taking* constitutive reasons to bear . . .". The verb "taking" would generate ambiguity that would falsify the claim—they might be *extrinsic* reasons for taking them to be reasons.

recognize that helping kindly would advance your career aims. You decide to help. Of course, if you decide to help your colleague simply in order to advance your career aims, your action will not be genuinely kind. You will have performed a strategic helping action, instead, merely acting as if you are kind.

It should be obvious enough that you can't perform the kind action for only extrinsic reasons. But the case should puzzle us more than it might, at first. At first it seems the problem is simply that, when you act on your extrinsic reason, your action doesn't qualify as kind. But, upon closer examination, it is actually not clear that, in such a case, you act on an *extrinsic* reason, at all.

A reason, as I understand it, is not specified simply by its content (need or vulnerability or reputation). Rather, a reason is a consideration (such as need or vulnerability) which bears on a question (such as whether to help).[7] Reasons for helping are considerations which bear positively on whether to help. All the reasons contained in the operative justification constitutive of a kind helping action must bear, eventually, on the question of whether to help—because it is by answering *that* question that one forms the intention to help. The intention to help qualifies as a kind one insofar as one answered the question of whether to help for the reasons constitutive of kindness.[8]

---

[7] I discuss the nature of reasons, at length, in my "Reasons, Actions, and Attitudes" (in progress). I say in the text above that a reason is a consideration that bears on a question. This covers over a complication. A person may *take* a consideration to bear on the question on which does not actually so bear. In that case, the person takes a consideration to be a reason, even though it is not (as we may say) a *good* reason, or not *really* a reason. Still, it is, really, *her* reason. So, it may be better to say that a reason is a consideration that *is taken* to bear on a question, and a *good* reason is one that is rightly so taken. (This means that "reason" will not be the fundamental normative notion.)

[8] Nothing I have said rules out the possibility that acting kindly "for its own sake" is among the reasons constitutive of kindness—that kindness is done, as it is sometimes put, "under the description 'kind'." Perhaps the fact that an action would be kind, if taken to be a reason for acting, would itself qualify the action as kind. I am quite doubtful of this, but I do not believe that anything in this paper rules it out.

An *extrinsic* reason for a kind helping action, in contrast, bears on a quite different question—not on whether to help, but rather on whether it is good, in some way, to perform a kind helping action (or to form a kind intention to help). Thus one will not, simply by finding these reasons convincing, settle the question of whether to help and therein form an intention to help (whether kind or no). Rather, by finding convincing the reasons which bear on whether it is good, in some way, to help kindly, you will answer the question of whether it is good, in some way, to help kindly. By answering this question, you will therein form the *belief* that it is good to help kindly. You might also form a desire to help kindly. But one will not, simply by finding these reasons convincing, form an intention to help, kind or no. These reasons bear on the wrong question.

(Of course, once you believe it is good to help kindly, and perhaps even desire to help kindly, it is tempting to think you could then simply decide to help kindly—that you could then consider the question of whether to help kindly, answer it affirmatively, and so intend to help kindly. But this thought is mistaken. You cannot decide to help kindly just because you have some reason that shows it useful to help kindly. Your reason might be extrinsic. Rather, one decides to help kindly by deciding to help, and doing so for reasons constitutive of kindness. So a belief that helping kindly would be good, for some reason, together with a desire to help kindly, will not, by themselves, enable one to help kindly.)

So , extrinsic reasons for helping kindly bear on the question of whether it would be good to help kindly. By finding them convincing, one will believe it good to help kindly.

Could extrinsic reasons also bear on the question of whether to help, and so be reasons for a helping action that might also be kind?

If you know that you do not find the reasons constitutive of kindness convincing, and you understand what kindness requires, then you know in advance that, even if you were to decide to help, your helping won't be kind. So it is hard to see how your extrinsic reasons, the reasons which bear on whether it is good to help kindly, can, in such a case, answer the question of whether to help. In such a case, it is unclear that they count in favor of helping at all.[9]

Of course, you might be confused about kindness. You might not understand, or might not bear in mind, that helping because your career aims show it good to act kindly won't be a kind helping. You might decide to help because career advancement shows a kind intention good to have, thinking that, by helping, you will advance your career aims. But, since you are mistaken about what kindness requires, your action misfires. You perform a helping action which isn't kind (an action which, in a sense, you had no reason to perform).

Alternatively, you might know what kindness requires, and be quite clear that you cannot act kindly, but take your career aims to give you reason to *act as if* you were kind. In any such case, you are not acting on extrinsic reasons for helping kindly, but rather on constitutive reasons for acting *as if* you were kind. Importantly, a single consideration, such as advancing one's career, can readily provide several different reasons, by bearing on several different questions. By bearing on whether it would be good to help kindly,

[9] Careful attention to one's situation seems to erode one's confidence that one can act on a particular reason. In this, it is reminiscent of Kavka's toxin puzzle (Gregory Kavka, "The Toxin Puzzle," *Analysis* 43 [1983]: 33–36). I find it noteworthy that there are such cases.

your career aims provide you with an extrinsic reason for a kind helping action; by bearing on whether to act as if you were kind, they can provide a constitutive reason for acting as if you were kind. Thus, you can readily take your career aims to be reason to help—by taking them to be a reason to act as if you were kind—without thereby acting on the extrinsic reason they also provide.

Finally, extrinsic reasons for acting kindly—reasons which are taken to bear on whether it is good to help kindly—can bear on whether to help, if (like Aristotle) you think that helping is somehow a way to become kind. In this case, your extrinsic reasons, which bear on whether it is good to act kindly, can bear on whether to help, because helping (kindly or not) is a means of bringing it about that you act kindly in the future. Still, when you act on your extrinsic reasons, you do not directly perform the action of which they count in favor. The action you perform, when acting on them, is not itself kind. Rather, you act so as to bring it about that you perform the action of which they count in favor.

Thus, extrinsic reasons bear on whether it is good to help kindly; they can also bear on whether to help only if helping is a way to help kindly. Thus extrinsic reasons for a kind helping action are not reasons for which you can directly perform the kind helping action of which they count in favor. They are rather reasons for which you can perform some action that will bring it about that you perform that kind action.

Notice what it would take to succeed in bringing yourself to act kindly, for extrinsic reasons. Your extrinsic reasons for acting kindly can motivate you to try to bring yourself act on the reasons constitutive of kindness. You might, e.g., look for the reasons to which the kind person responds. You might direct your attention and imagination in

certain ways, making an effort to attend more carefully to the needs and vulnerabilities of your colleagues. Of course, insofar as any such efforts are motivated by a concern for your career goals, they will be prudential or self-serving rather than kind. But suppose that, because you are on the look-out, you notice the exhaustion of one of your colleagues and realize that you could volunteer to do a certain task over the weekend, a task that would otherwise be assigned to her, spoiling her plans for a much-needed weekend get-away. Suppose you then decide to volunteer in order to allow her to take her trip. What would be required for you to have thus succeeded in bringing yourself to perform a kind action?

As we have seen, the kindness of an action depends not only on which considerations you take to be reason-giving, but also on the justificational context in which you so take them. In order for you to have succeeded in bringing yourself to perform a kind action, you must not take your colleague's exhaustion to bear on whether to help because finding it to be a reason to help would advance your career. Doing so would spoil its virtue. In order to succeed in bringing yourself to perform a kind action, you must come to be convinced of the adequacy of the reasons constitutive of kindness independently of the force of your extrinsic reasons. While your extrinsic reasons can be reasons to direct your attention and engage your imagination, the attending or imagining must bring you to act on other reasons, independently of the extrinsic ones. Your extrinsic reasons must, in a certain sense, be left behind. They could be part of the explanation of how you came to act on the constitutive reasons. They could also be reasons for being pleased that you

were able to act in this way.[10] But they cannot be taken to show that the constitutive reasons bear convincingly on whether to help.

Thus, in order for you to succeed in bringing yourself to perform a kind action, you must bring yourself to act on certain reasons, independently of your extrinsic reasons for doing so. Making it the case that you act on other reasons involves something like adopting another point of view.[11] Acting so as to make yourself act independently for other reasons involves the mild form of alienation I mean to highlight. When you act on your extrinsic reasons, you are not yet performing the action of which they count in favor, but rather are bringing it about that that you will perform an action for different reasons.

There is one final case to consider. Suppose you are already convinced by the reasons constitutive of kindness. You are thus able to perform a kind action. It is still true that performing a kind action might advance your career, and you may know this. If you know of yourself that take the constitutive reasons to be reason enough to help, then you know that, if you help, you will help kindly. So it seems your extrinsic reason could *now* be part of your reason for performing the kind helping action. After all, if you help, you will help kindly, thus advancing your career aims. Suppose, e.g., you are faced with a situation in which you could volunteer at work for a certain weekend task, relieving your colleague's exhaustion, or you could spend the same weekend cooking and cleaning for a neighbor who just brought home a new baby. Suppose that, in both cases, you would be

[10] Likewise, you might be pleased that some belief you hold allows you to avoid censure, or makes your life richer, without thinking that the fact it allows you to live, or makes your life richer, is a reason to believe.

[11] Often talk of point-of-view can be usefully understood in terms of operative justifications.

moved to relieve exhaustion, so that, in either case, your action would be kind. But, if you help your colleague, you will also, by performing a kind action, advance you career. It seems your extrinsic reason could then be a reason to help your colleague, rather than your neighbor, and, when you help, your helping will be kind.

I want to allow that, in such a case, your extrinsic reason for a kind helping action can be a reason for helping your colleague, and that such a helping can be kind. However, even in this case, there is a sense in which your extrinsic reasons are not reasons for directly performing a kind helping action; there is still a sense in which acting on them involves bringing yourself to act on other reasons, and so involves a mild form of alienation. In this case, the extrinsic reasons bear on whether to help only because you already know that, when you help, your helping will be kind. They are reasons to act because, when you act, you will be acting on other reasons. You are thus still doing something less than directly performing the action of which they count in favor.

We can clarify the sense in which this action is indirect by considering the complex structure of the operative justification(s) of this action and comparing it to a case of simple over-determination. Suppose you decide to help at work both to display your new computer skills and to relieve your colleague's exhaustion. In this case of simple over-determination, you act on two independent lines of reasoning, which happen to reach the same conclusion. The operative justification of your action thus contains two justificational structures, each independent of the other. In contrast, when you decide to help your colleague rather than your neighbor, you are acting on two lines of reasoning, one of which presupposes the other. It is useful, in this case, to talk of two different operative justifications for your action. The more basic operative justification bears on

whether to act, and contains the reasons constitutive of kindness. The parasitic operative justification also bears on whether to act, but it does so only *given that* you know that if you act, you will also act on the first, more basic, operative justification. Your action, then, involves a kind of self-reference, or split in point of view—part of your reasons for acting presupposes knowledge of your other reasons for acting. You wouldn't have the parasitic reasons if you didn't know that you already find the more basic reasons independently convincing (if you didn't have confidence in your own kindness). Your extrinsic reasons bear on whether to act only because you know that, by acting, you will bring it about that you act kindly. Extrinsic reasons are still not reasons for directly performing the kind helping action, but are again reasons for making it the case that you act for other, independent reasons.

I hope thus to have introduced the notion of an extrinsic reason and to have illustrated the way in which such reasons are not reasons for which you can directly perform the action of which they count in favor. They bear on whether it is good to perform some action—an action which is identified as an action performed for different reasons. By finding the extrinsic reasons convincing, you will, therein, *believe* the action good to perform. You might also desire or wish to perform the action, and may even decide to bring it about that you perform the action. However, if you understand what performing the action requires, you will not take these reasons to bear on the question of whether to act, unless you think that acting is a way to bring it about that you perform such an action. If you think that acting is a way to bring it about that you act on the constitutive reasons, then in acting on the extrinsic reasons (in taking them to be sufficient to decide to act), you are

deciding to bring it about that you act for other reasons. You do not, for extrinsic reasons, decide to act and thereby perform the action of which they count in favor.

MORAL PHILOSOPHY

We can now return to the importance of this class of reasons for thinking and theorizing about morality. Extrinsic reasons for kind actions are possible because the kindness of an action is determined by the reasons for which it is done, and yet, because we can think about the kindness of our own actions, and because it can matter to us in various ways whether our actions are kind, we can have reasons for wanting to perform such actions that are not among the reasons that would qualify the action as kind. I have argued, additionally, that these extrinsic reasons are not reasons for which one can directly perform a kind action, but are rather reasons for which one could at most act upon oneself to bring it about that one performs such an action. Insofar as acting on extrinsic reasons involves bringing yourself to act for other reasons, acting on them involves a certain (mild) form of alienation.

The existence of this class of reasons is of considerable importance for moral philosophy. Moral philosophy typically concerns itself, at least in large part, with reflection upon or construction of moral justifications for our actions. If (as seems plausible) many of the actions a moral theory hopes to justify are, like kind actions, identified by the reasons for which they are done, there will be extrinsic reasons for them. And, given what we have learned about extrinsic reasons, if one relies on extrinsic reasons in justifying such actions, morally, it seems that one is committed to the awkward position that those actions cannot, even in principle, be directly performed for the reasons

which justify them, morally. For such reasons, one could at most act upon oneself to bring it about that one acts on other reasons.

Some may wonder how bad this result is, in the end. The alienation I have displayed in the case of helping kindly is, after all, quite mild, and one might think that subjecting oneself to it—bringing it about that one acts for certain reasons—is just the sort of ability a moral agent might be expected to exercise.

I don't think we can be so sanguine about the possibility of relying on extrinsic reasons in our moral justifications. What is a mild form of alienation, for the agent involved in self-improvement, becomes severe, and in fact threatens the coherence of the agent, when incorporated into the justification provided by the moral theory. I will argue for this last claim in due course. But first I hope to show how it is that moral theory might come to rely on reasons extrinsic to the actions it requires.

If, in providing a justification for those actions typically identified as morally good or right, we simply hoped to provide considerations that show those actions in some way good or useful or necessary or appropriate (even *morally* good or useful or necessary or appropriate), our task would be relatively unrestricted. A great variety of considerations might, in this way, count in favor—perhaps even decisively in favor—of performing such actions. Morally good or right actions may well uniquely contribute to the stability and well-functioning of society, or to the continuation of the species, or to overall well-being. Perhaps all and only such actions are in accord with the commands of God, or constitute the fulfillment of human nature or the health of the human soul. Perhaps moral actions are the only ones which can be agreed upon as permissible by a community of equals, or the only ones that could be consistently willed by autonomous agents. The range of

considerations that show something good, even overwhelmingly morally good, about morally required action is plausibly wide and varied. However, the forgoing reflection on kind action suggests that the mere fact that a consideration counts in favor of performing an action does not guarantee that it is a reason for which one can perform that action. The reason may yet be extrinsic to the action of which it counts in favor. It may simply show that it is in some way good to perform an action that must be performed for other reasons. In that case, the reason offered is extrinsic. It is, at best, a reason to bring it about that one performs the action, not a reason for which one can directly act.

Moreover, it seems plausible that the reasons appealed to in various moral theory are quite often not be those found in the operative justifications of the actions the theory hopes to justify. For example, I suspect that a wide range of virtue terms ("generous," "honest," "fair-minded," "conscientious"), when used to describe actions, function in the same way I have suggested that "kind" functions: when used to describe actions, these terms pick out a rough-and-ready class of operative justifications. If one thinks that moral theory is committed to providing a justification of the actions we ordinarily identify as virtuous, then moral theory will encounter the hazard of extrinsic reasons. There will be reasons—likely even morally important reasons—that count in favor of performing an action that is virtuous in this or that way, which are not reasons constitutive of the virtue. Moreover, it seems likely that at least some of the reasons appealed to in standard moral theories are among them.

Importantly, though, the hazard of relying on extrinsic reasons is not restricted to those theories concerned to justify the actions we ordinarily identify as virtuous. Utilitarianism, e.g., has no qualms about revisionism, and so could be quite content to

displace ordinary virtuous action in favor of those deemed right by utilitarian principles. Yet familiar criticisms of utilitarianism can be well understood as worries about utilitarianism's reliance on reasons that are extrinsic to the actions it hopes to justify.

Utilitarianism justifies the undertaking of action or the pursuit of a project insofar as that action or project contributes to a certain outcome—the greatest overall happiness. But, if one accepts the plausible psychological claim that people will not be made happy by taking as the ultimate or final aim of their actions the greatest overall happiness, then the actions people undertake and the projects they pursue will contribute to the utilitarian's outcome only if they are pursued from operative justifications in which their pursuit is not ultimately justified by their service to the utilitarian aim. In other words, accepting the plausible psychological claim about the operative justification of actions productive of happiness leads one to conclude that the reasons appealed to by utilitarianism are extrinsic to the operative justifications of the actions it requires: they show something good about acting from such justifications, but are not part of them.

Since the utilitarian justification is extrinsic to the operative justification of the actions it requires, the agent cannot directly act on the utilitarian reasons in acting as utilitarianism requires. Rather, he can at most act on himself and bring it about that he acts for other reasons.

As I read him, Williams offers a structurally parallel concern about Kantianism. In the same way that utilitarianism requires agent to undertake projects other than the utilitarian one, Williams thinks that Kantian theory must presuppose that each agent is engaged with particular projects that give her life meaning, that give her "reason to go on

at all."[12] It must presuppose this, he thinks, if the theory is to have an agent to address. However, Williams insists that those projects that gives life meaning (e.g., loving relationships), are what they are, and so give life meaning, only if they are pursued in certain ways. In particular, one must, Williams thinks, take one's pursuit of those projects to be justified (or to no longer stand in need of justification) in ways that make essential reference to, and give special weight to, particular people, and so in ways that are essentially partial. Yet Kantian morality demands that one pursue such projects subject to an impartial condition of permissibility. Williams seems to think that such considerations, even as simply a limiting condition on the permissibility of an action, is incompatible with the pursuit of such personal projects. Rather, he thinks our pursuit of such projects must outrun considerations of impartial justifiability. Thus, it seems, he thinks of the impartial requirements of morality as extrinsic to the operative justification of actions that morality cannot do without.[13] If Williams is right about this, then the Kantian, like the utilitarian, relies on extrinsic reasons to justify actions that the theory can neither do without nor leave outside its jurisdiction.

There are, then, a variety of ways in which a moral justification might encounter the hazard of extrinsic reasons. It may give a justification extrinsic to the virtuous action it hopes to justify, or extrinsic to actions that may or may not be virtuous, but which the theory somehow requires.

---

[12] "Persons, Character, Morality," 10.

[13] While it is a difficult, substantive question about such projects whether considerations of permissibility against an impartial standard can be, as a limiting condition, constitutive of them, it seems clear that Williams thinks they cannot be. Herman ("Integrity and Impartiality," 39) and Scanlon (*What We* Owe, 165) both suggest they can be.

Let's return, now, to the thought that the relatively mild form of alienation highlighted in the example of kindness is not one to be overly concerned about. One might think that Williams is being overly dramatic in claiming that moral theory demands a sacrifice in integrity or coherence of the agent. In the case we considered, acting on extrinsic reasons did not constitute an affront to one's character. It rather introduced a fairly mild form of alienation—one which, after all, any moral agent might be expected to exercise. A moral agent, one might think, is one who exercises self-governance over the reasons on which she will act and ensures that she acts only on those reasons which are morally justified.[14]

This way of putting the issue is misleading. A moral agent presumably needs the ability to consider whether the reasons on which she acts are good ones—that is, she needs to be able to consider whether need or vulnerability are, in fact, reason enough to help—whether they settle the question of whether to help. But to grant this is only to say that the moral agent needs to be able to reflect upon whether the reasons on which she is considering will, all things (including morality) considered, settle for her the question on which they bear. The ability to bring yourself to perform an action for reasons extrinsic to it is a quite different ability. It is the ability to consider whether a certain operative justification is a good one, for reasons extrinsic to it, and then bring yourself to act from that operative justification, while leaving the extrinsic reasons behind.

We have seen that ordinary moral agents do sometimes demonstrate this ability; it is an important capacity for the project of self-improvement. However, the role of this ability must always be limited, and limited in a way that it could not be, if extrinsic reasons provided the moral justification for an action.

[14] Cf. Peter Railton, "Alienation, Consequentialism, and the Demands of Morality," CITE.

Moral justifications determine whether an action is morally permissible (or perhaps required). If that action is itself identified by the reasons for which it is performed, then a moral justification is one which shows whether acting on those reasons is permissible—whether one is justified in acting on those reasons. The cases of conflict cited by Williams, between moral demands and the personal or partial projects, highlight this role of moral justification. In the case of conflict it becomes clear that the utilitarian justification, e.g., is meant not only to guide you in developing projects and to break ties between them, but also to govern whether one's other projects are genuinely reason-giving. Likewise for Kantianism. Williams famously considers the case of a man who rescues his drowning wife rather than a drowning stranger. For this person, impartial considerations of permissibility are supposed to govern whether or not the essentially partial reason, "it's her," is a sufficient reason to save his spouse. In any such case, moral reasons seem to play the role of determining whether other reasons bear sufficiently on the question of whether to act. But extrinsic reasons, we have seen, cannot play this role. They can be reasons for brining yourself to act on other reasons, but then they must be left behind. They cannot govern whether one should act on constitutive reasons. To perform the action recommended by the extrinsic reasons, the constitutive reasons must be themselves independently convincing. Here, then, we encounter the incoherence that concerns Williams. These moral theories would ask one to allow extrinsic reasons to govern whether one acts on constitutive reasons. But extrinsic reasons, by definition, cannot play this governing role.

Or, rather, extrinsic reasons cannot play this role so long as the "governance" in question is rational. An extrinsic reason could "govern" whether one acts on constitutive

reasons, so long as the extrinsic reasons did so without actually being taken to bear on whether the constitutive reasons are reason-giving. This is possible if the extrinsic reason "governs" whether one acts on a constitutive reason, not by being taken to bear on whether the constitutive reason bears sufficiently on whether to act, but rather by giving one reason to exercise some power or means and make it the case that one acts on the constitutive reason. We can easily imagine such a governing relation occurring between two minds. Suppose, e.g., it would benefit your career if I act kindly. If you have some effective means to make it the case that I consistently act on reasons constitutive of kindness (perhaps simply by presenting to me constitutive reasons that you know I will find independently compelling), then you could, for extrinsic reasons, effectively govern whether I act on reasons constitutive of kindness, through your diligent exercise of this effective means. You can make it the case that I act on constitutive reasons. In this way extrinsic reasons could govern whether constitutive reasons are taken to be reason-giving, without being taken to bear on whether the constitutive reasons are reason enough to act.

Again, one might be tempted to think that this just is the ability of the moral agent—she can step back from her reasons and decide which reasons she will act upon. But, again, deciding to act on constitutive reasons for extrinsic reasons requires more than the ability to step away from one's inclinations and consider whether one has worthy reasons for acting. It requires the ability to settle this question for extrinsic reasons *without* those reasons appearing in the operative justification of one's action (lest they change its nature). Doing this, it seems, requires a division between minds, or between parts of a mind. One must ensure that the extrinsic reasons, for which one decides to bring it about that one acts on the constitutive reasons, are not taken to bear on whether the constitutive

reasons are reason enough to act. To so do, one must in some way hold apart the (sub)agent who is, for extrinsic reasons, authoritatively governing whether one acts on constitutive reasons, from the (sub)agent who is acting on the constitutive reasons—because the (sub)agent who is doing the extrinsic governance must take the question of whether to act to be subject to constraints or justifications that the (sub)agent who is acting on the constitutive reasons cannot recognize. Such an agent is, in fact, lacking in integrity—in the most literal sense.[15]

MORAL THEORY WITHOUT EXTRINSIC REASONS

In the face of these problems, it seems to me we should avoid relying on extrinsic reasons in providing moral justifications for action.[16] However, avoiding extrinsic reasons makes the task of reflection upon or construction of the moral justifications considerably more difficult. If we were simply interested in providing reasons that show it good or right, in some way, to act in those ways recognized as morally good or right, then our task would be to establish that the class of moral actions corresponds to a class of actions that share some common good-making feature, or that have in common some good effect. There are, surely, a wide range of good features and possible good effects of such actions. Again, those actions widely recognized as morally good or right may contribute to a stable society, or to the continuation of the species, or to overall well-being; they may be in accord with the commands of God, or constitute the fulfillment of human nature, or be

[15] This agent displays what Donald Davidson takes to be characteristic of irrationality: she is caused to act on a reason that does not rationalize (?) her action. He argues that this requires a divided mind. See his "Paradoxes of Irrationality," in *Philosophical Essays on Freud*, ed. Wollheim and Hospers (Cambridge: Cambridge University Press, 1982): 289–305, and "Deception and Division," in *The Multiple Self*, ed. Jon Elster (Cambridge: Cambridge University Press, 1986): 79–92. (Both are reprinted in CITE.)

the only actions which can be agreed upon as permissible by a community of equals, or the only ones that could be consistently willed by autonomous agents. But determining which account provides the moral justification of action is not simply a matter of determining which feature is genuinely shared by all and only the morally good or right actions—or even of determining which of these features seem, themselves, most morally important. After all, one could easily grant that all moral actions have some good effect, or share some good-making feature, while doubting that that effect or feature is distinctive of the reasons constitutive of those actions, as a class. Any such feature or effect, while perhaps useful in *locating* morally correct actions, and while perhaps useful in fending off certain forms of skepticism, may be a merely extrinsic reason for performing it.[17] So, if we want to provide an account of the justification or ground of morally good action, we must ask whether these features or effects are themselves reasons constitutive of the actions that moral theory seeks to justify. If they are not, then to justify those actions by appeal to them would be to justify the actions by extrinsic reasons.

Thus, if one wants to avoid relying on extrinsic reasons in one's account of moral justification, the tasks of moral philosophy become considerably more difficult. We will have to carefully attend to the actions and attitudes we recognize as morally good, or as required for a good life, in order to understand the reasons constitutive of them. We will

---

[16] Alternatively, one could defend the division in the mental life of the agent. I won't examine this route here.

[17] The forms of skepticism extrinsic justifications may meet are those which suspect that morality is in some way bad for us, or unhealthy. These might be met by showing that morality satisfies an external constraint. On the other hand, the limitations of extrinsic reasons show another kind of skepticism unanswerable. If the skeptic is defined as the one who is unconvinced by any constitutive reason, and if answering her requires giving her a reason that she will find convincing, not just for wanting to be moral or

then have to consider whether there is any feature of their operative justifications that gives unity to this class.[18] We will have to do so without allowing our moral philosophy to simply provide a subtle but uncritical description of the psychological status quo. We will also have to understand the proper role, in our thinking, of the other effects and features morally good actions and attitudes happen to share—effects and features which may not be accidental, and which may well play a role in the genealogy of morals, and yet which are, nonetheless, extrinsic to their justification.[19] Though I think that avoiding reliance on extrinsic reasons will thus add to the task of moral philosophy, I expect heeding the difficulty to yield fruitful results.[20]

---

for making herself moral, but for which she can directly perform a moral action, then she cannot be answered.

[18] I suspect this will be a project with a Kantian spirit.

[19] I expect there to be a route from moral psychology to metaethics. Some of the features of moral psychology I have here highlighted will, I think, set constraints on what we can think about the grounds of our moral justifications.

[20] Earlier versions of this material have benefited from comments from and/or conversation with Barbara Herman, David Jensen, Mark Johnson, Sean Kelsey, Christine Korsgaard, Gavin Lawrence, Richard Moran, T. M. Scanlon, Seana Shiffrin, Julie Tannenbaum, and the members of the Southern California Law and Philosophy Group.

APPENDIX: WILLIAMS AND EXTRINSIC REASONS

I will here consider in more detail Williams' well-known complaints against moral theory, offering both a more immediate and what I will call a "deeper" interpretation of them. I will then show how the "deeper" interpretation can be well understood as a concern about what I have called extrinsic reasons.

Consider, first, Williams' argument against the utilitarian in his "Critique of Utilitarianism." Williams there maintains that utilitarianism fails as a moral theory because it cannot make adequate sense of integrity. It cannot make sense of integrity because it "cannot coherently describe the relations between a man's projects and his actions" (100).

Williams' argument is based on a number of examples, one of which concerns George. George has recently earned his Ph.D. in chemistry but cannot find work. This is especially bad, as George's family depends on his ability to earn a wage. A friend proposes a job in a laboratory working on chemical and biological weapons. George finds chemical and biological warfare appalling. The friend points out that George's refusal will not stop the work; the research will be done, with or without him. In fact, if George took the job, he would keep the zeal and dedication of the alternate candidate out of the lab, perhaps slowing progress on the weapons. There is a suggestion of sabotage.

Williams points out that, for the utilitarian, the answer to George's dilemma should be obvious: George should take the job. To us, this answer seems far from obvious, but the problem with the utilitarian view, Williams points out, is not with its answer or its obviousness, but with the way the view arrives at its answer. The view does not properly

understand the relation between George's action and his own projects and attitudes. Here is Williams' own statement of his charge against utilitarianism:

> The point is that [the agent] is identified with his *action*s as flowing from projects and attitudes which in some cases he takes seriously at the deepest level, as what his life is about. . . . It is absurd to demand of such a man, when the sums come in from the utility network . . . that he should just step aside from his own project and decision and acknowledge the decision which utilitarian calculation requires. It is to alienate him in a real sense from his actions and the source of his action in his own convictions. It is to make him into a channel between the input of everyone's projects, including his own, and an output of optimific decision; but this is to neglect the extent to which *his* actions and *his* decisions have to be seen as the actions and decisions which flow from the projects and attitudes with which he is most closely identified. It is thus, in the most literal sense, an attack on his integrity (116–117).

So somehow, by failing to understand the relation between the person, his projects, and his actions, the utilitarian attacks his integrity "in the most literal sense." I take this to mean that the utilitarian threatens the unity or integration of the agent. Before considering this charge more carefully, let's look at Williams' critique of Kantianism.

In a later article, "Persons, Character, and Morality," Williams makes essentially the same charge against the Kantian. First he reminds of us of his concern with the utilitarian:

> A man who has [a project with which he is most closely identified] may be required by Utilitarianism to give up what it requires in a given case just if that conflicts with what he is required to do as an impersonal utility-maximizer when all the causally relevant considerations are in. That is quite an absurd requirement (14).

It is, in fact, the absurd requirement we just considered. Williams goes on:

> But the Kantian, who can do rather better than that, still cannot do well enough. For impartial morality, if the conflict really does arise, must be required to win; and that cannot necessarily be a reasonable demand on the agent. There can come a point at which it is quite unreasonable for a man to give up, in the name of the impartial good ordering of the world of moral agents, something which is a condition of his having any interest in being around in that world at all (14).

The Kantian overlooks the unreasonableness of this demand, Williams explains, because the Kantian omits "what is involved in having a character" (14). So, the charges are parallel: the utilitarian cannot understand integrity, and so makes absurd demands; the Kantian omits character, and so makes unreasonable demands.

It is important to see that this charge of "absurdity" or "unreasonableness" runs deeper than a simple complaint that the demands of these theories are hard, or extreme, or even that they may require a sacrifice of integrity or an overhaul of character in the case of the morally reprobate. Morality may well be hard or extreme, and may well require a sacrifice or overhaul in such cases. The charge is rather that these theories go wrong as moral theories by missing something fundamental about ordinary, non-reprobate moral agents and their actions. To better understand the charge, then, we must ask how Williams himself understands "the relation between a man's projects and his actions" or "what is involved in having a character."[21]

For this we can look, first, to Williams' views about reasons.[22] Famously, Williams believes that the truth of the claim that "*A* has a reason to ϕ" depends on the relation between the statement and *A*'s particular psychology. In particular, if *A* couldn't come to be motivated by the claim through some "sound deliberative route" from what Williams calls his "subjective motivational set," or his "*S*," then this claim that he has a reason will never explain any action of his. But the capacity to explain action is a necessary feature of any reason. Thus, whether *A* really has a reason depends on his particular psychology. Williams calls the reasons which could possibility motivate a person, given that person's

[21] My interpretation of Williams' arguments will differ from that provided by most others I have seen. I hope it will seem a deeper reading. Susan Wolf, for example, says that Williams' criticisms are "often referred to as the objection that morality is too demanding" ("Meaning and Morality," 229). She seems to agree with me that this is, by itself, too shallow an objection, but she hopes to deepen it by highlighting the particular *way* in which morality is too demanding, "the specific nature of the sacrifice Williams contemplates being demanded of the agent, namely, a sacrifice of that which gives the agent his reason for living . . ." (300). As will become clear, I do not think that sacrifice, of any sort, is really what is at issue in Williams' criticisms, though, certainly, his text leaves that reading open.

[22] Found in "Internal and External Reasons" (cited in the text) and "Internal Reasons and the Obscurity of Blame," in *Making Sense of Humanity* (Cambridge: Cambridge University Press, 1997): 35–45, first published in *Logos: Philosophical Issues in Christian Perspective* 10 (1989): 1–11.

S, "internal" reasons, and the remaining, purported reasons, which are not rightly related to the agent's *S*, "external" reasons. He argues that there are no external reasons.

Williams' views about reasons help to display his views about the relation between a person's projects and his actions, and so shows how Williams understands the place or importance of character. According to Williams, the members of a person's *S* includes include things like "dispositions of evaluation, patterns of emotional reaction, personal loyalties, and various projects . . . embodying the commitments of the agent" (105). These members are constitutive of his character, and provide a person with what Williams sometimes calls his "point of view."[23] Taken together, the members of a person's *S*—his desires, concerns, and projects, also known as his character—will, according to Williams, control deliberation and determine what, for him, counts as a reason, and so will determine what he can reasonably do.

We might, then, understand Williams to be claiming that the demands of the utilitarian and the Kantian are absurd or unreasonable because, by misunderstanding the relation between his projects and his actions, they end up requiring the agent to do things he has no reason to do. Authoritative yet impersonal moral principles are not among the agent's personal concerns or character, and so make demands that an agent may have no reason to meet.

If we were to understand Williams' criticism in this way, it might be answered by in some way rejecting Williams' internalism about reasons. One might, for example, point out that the demands of morality are impartial, but needn't, for that, be impersonal. One

[23] In "Persons, Character, Morality," Williams says, "an individual person has a set of desires, concerns, and, as I shall often call them, projects, which help to constitute a *character*" (5).

might maintain that morality can legitimately demand that an agent's *S* be shaped or formed, in moral education, to include among his personal concerns the impartial moral concern that the Kantian or utilitarian finds central. Alternatively, one could reject the broadly Humean psychology that Williams espouses and insist, with some Kantians, that we not only *have* reasons that are not dependent on the contingencies of our particular psychological profile, but we are capable of being motivated by those reasons, regardless of that profile.[24] We are all, as rational agents, capable of being motivated by the impartial moral principles.

While I think this interpretation of Williams is correct, and these replies to it forceful, I also believe that Williams also hopes for a deeper charge, underlying this one, which does not depend on his internalism about reasons and so cannot be met in either of these ways. This deeper charge would claim that acting on the reasons morality recommends would involve the agent in a kind of incoherence and so would be "in the most literal sense" an attack on his integrity. I will first sketch Williams' deeper charge, before reinterpreting it, using my own notion of extrinsic reasons.

Reconsider, then, the reply that morality can legitimately demand that an agent's *S* be shaped or formed, in moral education, to include the impartial moral concern that the Kantian or utilitarian finds central. Williams' deeper charge would claim that this is simply not possible. There is something about the impartial yet authoritative Kantian or utilitarian demand that prevents it from being integrated into a single, coherent character.

[24] There is an extensive literature responding to Williams claims about external reasons. Highlights include Christine Korsgaard "Skepticism about Practical Reason," *The Journal of Philosophy* 83, no. 1 (1986): 5–25; Rachel Cohon, "Are External Reasons Impossible?" *Ethics* 96 (1986): 545–56; Barbara Herman "Integrity and Impartiality" in *The Practice of Moral Judgment* (Cambridge: Harvard University Press, 1993): 73–93; and John McDowell, "Might There Be External Reasons?" in *World, Mind, and Ethics*, ed. J. E. J. Altham and Ross Harrison (Cambridge: Cambridge University Press, 1995): 68–85.

The deeper charge shows that if one were to adopt the Kantian or utilitarian concern as one of one's "ground projects," one would be forced into a kind of incoherence, because the impartiality and/or authority demanded by these projects makes them a kind of higher-order project that cannot be integrated into one's character.

In his "Critique of Utilitarianism" Williams asks, "What projects does a utilitarian agent have?" He answers, "As a utilitarian, he has the general project of bringing about maximally desirable outcomes." But, as Williams notes, in order for there to be happiness in the world there must be projects other than the utilitarian one:

> The desirable outcomes, however, do not just consist of agents carrying out *that* project [of bringing about maximally desirable outcomes]; there must be other more basic or lower-order projects which he and other agents have, and the desirable outcomes are going to consist, in part, of the maximally harmonious realization of those projects. . . . Unless there were lower-order projects, the general utilitarian project would have nothing to work on, and would be vacuous (110).

These more basic or lower-order projects are the familiar concerns of life: desires for things for oneself, one's family, one's friends, one's career ambitions, artistic endeavors, religious commitments, political projects, etc. The higher-order, utilitarian project is concerned with bringing about a certain state of affairs, viz., that in which utility is maximized. Because it is a higher-order project with this particular aim, Williams thinks the utilitarian project is guaranteed to conflict with the lower-order projects in a way that compromises one's integrity. It must conflict because, from the point of view of this higher-order project, one sees the lower-order projects as not especially one's own and as justified only if they serve as a means to the end of bringing about this state of affairs. But, Williams suggests, this detached, purely instrumental view is not one which can be coherently housed together in a well-integrated character with the point of view of these lower-order projects.

Consider what Williams says in his later work, *Ethics and the Limits of Philosophy*, about the view which indirect versions of utilitarianism must take toward certain dispositions of character. He points out that, from the point of view of the indirect utilitarian, the value of these dispositions is instrumental. But this point of view must conflict with that of the dispositions themselves:

> The [agent's] dispositions are seen [by the utilitarian] as devices for generating certain actions, and those actions are the means by which certain states of affairs, yielding the most welfare, come about. This is what the dispositions look like when seen from the outside, from the point of view of utilitarian consciousness. But it is not what they seem from the inside. Indeed, the utilitarian argument implies they should *not* seem like that from the inside. The dispositions . . . will do the job the theory has given them only if the agent does not see his character purely instrumentally, but sees the world from the point of view of that character (108).

This complaint, about indirect utilitarianism's treatment of dispositions of character, is structurally similar to Williams' complaint about (any version of) utilitarianism's treatment of the projects which provide one with satisfaction in life. The value of these projects "from the view of utilitarian consciousness" is detached and merely instrumental—they are a means for realizing a certain state of affairs. And yet, if they are to be of any use in realizing that state of affairs, one cannot value the project in this detached, merely instrumentally way. So, again, the view of one's projects (and so one's character) from the "outside" cannot be the view one must have of them from the "inside." Utilitarianism requires this to be so in order to achieve its desired state of affairs. This creates a problem for the utilitarian in determining where "in the mind or in society" his theory can be "located" (107–8). It is not clear that it can coherently be included among the members of the agent's *S*, if his *S* is to be well-integrated.

Let's turn now to Williams' criticism of the Kantian, which shares the same structure. Williams will try to show that the Kantian is also committed to requiring the agent to take up a higher-order point of view which is incompatible with a necessary lower-order point

of view and so committed to attacking the agent's character. But unlike the utilitarian, the Kantian project does not require other, "lower-order" projects to play a merely instrumental role in the accomplishment of that end.[25] Rather, the Kantian project demands that, in living out your other projects, you set for yourself certain particular ends and not transgress certain limitations. So, in confronting the Kantian, Williams must argue that even a higher-order project of this sort could not be plausibly incorporated into the character of the agent.

This is just what Williams sets out to do. His arguments here, like his argument against the utilitarian, aim to show that the point of view required by Kantian morality is incompatible with the point of view of other projects which the Kantian project cannot do without. Certain projects, Williams argues, are themselves a condition for our having any reason to go on at all. The Kantian project cannot do without these other projects. Yet, Williams will argue, the Kantian project also requires one to take up a point of view of those projects which is incompatible with them.

Williams' argument depends, first, on the claim that, in order be in a position to take anything to be reason-giving at all, one must be engaged in one's own particular projects which constitute the non-moral and (crucially) partial point of view. This step of Williams' argument gains support from his views about personal identity. The idea seems to be, roughly, that it is only from the point of view of your own particular, personal projects that you can conceive of yourself as someone with a future at all; so only from the point of view of the projects which constitute your character, conceived of as peculiarly yours, can you think of yourself as having any reason to "go on at all" (10).

[25] I take it this is why the Kantian can do "rather better" than the utilitarian, in Williams' view.

We might say that, if one is to think of oneself as a particular self with a future, and so be able to think one has reason to go on at all, one must think from a point of view from which the consideration "it's me," not further explicable in general terms, has relevance. One must, that is, occupy an essentially partial point of view.[26]

Williams then argues that the project of Kantian morality is, due to its impartiality, incompatible with these projects from which we conceive of our future and which thereby enable us to have reason to go on. This point requires argument because while it is has been shown that, due to the difference in relevance assigned to "it's me," the two points of view are not the same, it has not yet been shown that they could not be combined as the point of view of a single, well-integrated agent.[27]

Williams seems to make the required argument by use of an example in which a man, who is able to save only one of two people, saves his wife. With this example, Williams suggests that the impartiality of the Kantian project will conflict with and alienate the agent from his personal point of view in something like the way the utilitarian demand conflicts with and alienates the agent from his lower-order projects. If the man in this example were a Kantian agent, Williams maintains, then his "motivating thought, fully spelled out," as he rescues his spouse, must include the additional, alienating thought that it is permissible to do so only because doing so does not transgress the limitations imposed by impartial morality. The Kantian's motivating thought must include *both* his

[26] When discussing Kantianism in *Limits* Williams provides a rather different account of impartiality (see 65–70), one which is more explicitly tied to the idea of detachment.

[27] As Barbara Herman points out in "Integrity and Impartiality," "[w]hile it is (psychologically) true that attachments to projects can be unconditional [that is, unconditioned by impartial requirements], it is not a requirement of the conditions of having a character that they be so" (39). Barbara Herman, "Integrity and Impartiality," *The Practice of Moral Judgment* (Cambridge MA: Harvard University Press, CITE).

personal sense of attachment to his wife *and* his judgment that, in this case, it is permissible to save his wife. However, as Williams points out,

> it might have been hoped by some (for instance, by his wife) that his motivating thought, fully spelled out, would be the thought that it was his wife, not that it was his wife and that in situations of this kind it is permissible to save one's wife (18).

Even the merely additional thought about justifiability is enough, Williams thinks, to make it the case that the agent is no longer occupying the point of view appropriate to the personal relationship. The impartial requirement forces a kind of detachment.

So, with this example, Williams suggests that certain projects (notably, those attachments which give us reason to "go on") require that we give relevance to the considerations "it's me" or "it's her" not further explicable in general terms. Kantian morality subjects these projects to justification against an impartial standard. But, Williams seems to suggest, the projects will not remain themselves when subject to that constraint. The thought about permissibility against an impartial standard provides, famously, "one thought too many" (18). Because the two points of view are incompatible, the Kantian project cannot be well-integrated into the agent's *S*. Morality forces a sort of detachment, and so produces a rift in the agent.

Williams' deeper argument against the Kantian shares its structure with his deeper argument against the utilitarian. In both cases the theory demands that the agent do not only what she *doesn't* have reason to do, but what she *couldn't* have reason to do. The agent couldn't have reason to comply with the demands of the theory, because these demands arise from a project which could not coherently be a project of any single, well-integrated agent. It could not be the project of any single, well-integrated agent because it is a higher-order project that requires that one have certain other projects and yet

requires that one take up a point of view *of* those projects which is incompatible with them. And so, the attempt to take up the detached project from which the moral demand issues results, Williams thinks, in a compromise of one's integrity or character.

I believe that this deeper worry can be well understood using the notion of extrinsic reasons I developed earlier. In considering whether a reason is extrinsic, one focuses, not on the relation between the agent's particular projects and the demands of morality—not, that is, on internalism about practical reasons—but rather on the relation between the reason and the action the reason recommends. Certain reasons, it turns out, will generate a familiar form of alienation from one's own "projects" or from "the source of [one's] actions in [one's] own decision," quite apart from questions of any particular agent's psychological set. Thus we can restate Williams' deeper worry in a way that illuminates and perhaps explains, rather than relies on, the spatial metaphors of detachment, "inside" and "outside," or "view of" and "view from," and so allows us more purchase on why the moral demands generate not just a certain alienation, but an incoherence in the agent—an attack on the agent's integrity, in the most literal sense.

We have already considered how how the utilitarian reasons are extrinsic to the justification of the necessary lower-order projects: The utilitarian moral justification bears on whether it is good to pursue certain other projects. It justifies the pursuit of those projects insofar as they contribute to a certain outcome. But those projects will fulfill their role only if they are pursued from operative justifications in which their pursuit is not ultimately justified instrumentally by their service to utilitarianism. Thus the reasons appealed to by utilitarianism are extrinsic to the operative justifications of the

projects it requires: they show something good about acting from such justifications, but cannot be part of them.

Since the utilitarian justification is extrinsic to the operative justification of the required lower-order projects, the agent cannot directly act on the utilitarian reasons in pursuing those projects. Rather, he can at most act on himself and bring it about that he pursues the lower-order projects for the reasons constitutive of them. If the agent does not start with a concern for these projects that is independent of his utilitarian aims, the utilitarian reasons give him reason to develop one. If he has competing lower-order projects, both of which he finds worthwhile for reasons constitutive of them, his utilitarian reason can allow him to break a tie between them, by giving him reason to chose the one that creates the most happiness. But whenever the agent pursues the lower-order projects, the utilitarian reasons can be at most merely additional, parasitic reasons for acting, as in the case in which one's career aims give one an additional reason to volunteer at work rather than to help one's neighbor.

We have also interpreted the difficulty for the Kantian in terms of extrinsic reasons. Williams insists that those projects that gives life meaning and so give on reason to go on, e.g., loving relationships, are what they are, and so give life meaning, only if they are pursued in certain ways. One must, Williams thinks, take one's pursuit of those projects to be justified (or to no longer stand in need of justification) in ways that make essential reference to, and give special weight to, particular people, and so in ways that are essentially partial. Yet Kantian morality demands that one pursue such projects subject to an impartial condition of permissibility. Williams seems to think that such considerations of permissibility, even as simply a limiting condition on action, is

incompatible with the pursuit of such personal projects. Rather, he thinks our pursuit of such projects must outrun considerations of impartial justifiability (18). Thus he thinks that the impartial requirements are extrinsic to them. While it is a difficult, substantive question about such projects whether considerations of permissibility against an impartial standard can be, as a limiting condition, constitutive of them, it seems clear that Williams thinks they cannot be.[28] If Williams is right about this, then the Kantian, like the utilitarian, relies on extrinsic reasons to justify actions that are essential to life.

One might now think that, if the worry is about acting on extrinsic reasons, Williams is being overly dramatic in claiming that moral theory demands a sacrifice in integrity. The cases we have considered, of bringing yourself to act on the reasons constitutive of kindness, show that acting on extrinsic reasons does not necessarily constitute an affront to one's character. It rather introduces a fairly mild form of alienation—one which, after all, any moral agent might be expected to exercise.[29] One might think the ability to step back from one's own motives and determine whether they are good motives to act from is, in fact, the defining ability of a moral agent. A moral agent, one might think, is one who exercises self-governance over the reasons on which she will act.

This objection founders on the authority of morality. A moral justification must demand a kind of authority that extrinsic reasons cannot have. The cases of conflict Williams brings up make just this point. In the case of conflict it becomes clear that the utilitarian justification, e.g., as a moral justification, is meant not only to guide you in developing projects and to break ties between them, but also to govern whether one's

[28] Herman ("Integrity and Impartiality," 39) and Scanlon (*What We* Owe, 165) both suggest they can be.

[29] Cf. Railton, "Alienation, Consequentialism, and the Demands of Morality."

lower-order projects are genuinely reason-giving. Likewise for Kantianism; impartial considerations of permissibility are supposed to govern, for the rescuer, the whether or not the essentially partial reason, "it's her," is a sufficient reason to save his wife. Any such governing reasons seem to play the role of determining whether the supposedly independent constitutive reasons bear sufficiently on whether to act. But extrinsic reasons, we have seen, cannot play this role. They can be reasons for brining yourself to act on other reasons, but then they must be left behind. To perform the action recommended by the extrinsic reasons, the constitutive reasons must be themselves independently convincing. Here, then, we encounter the incoherence that concerns Williams. These moral theories would ask one to allow extrinsic reasons to govern whether one acts on constitutive reasons. But extrinsic reasons, by definition, cannot play this role.

Or, rather, extrinsic reasons cannot play this role so long as the "governance" in question is rational. An extrinsic reason could "govern" whether one acts on constitutive reasons, so long as the extrinsic reasons did so without actually being taken to bear on whether the constitutive reasons are reason-giving. This is possible if the extrinsic reason "governs" whether one acts on a constitutive reason, not by being taken to bear on whether the constitutive reason bears sufficiently on whether to act, but rather by giving one reason to exercise some power or means and make it the case that one acts on the constitutive reason. We can easily imagine such a governing relation occurring between two minds. Suppose, e.g., it would benefit your career if I act kindly. If you have some effective means to make it the case that I consistently act on reasons constitutive of kindness (perhaps simply by presenting to me constitutive reasons that you know I will

find independently compelling), then you could, for extrinsic reasons, effectively govern whether I act on reasons constitutive of kindness through your diligent exercise of this effective means. You can make it the case that I act on constitutive reasons. In this way extrinsic reasons could govern whether constitutive reasons are taken to be reason-giving, without being taken to bear on whether the constitutive reasons are reason enough to act. They would have a kind of authority, and yet remain extrinsic to the action performed.

One might be tempted to think that this just is the ability of the moral agent—she can step back from her reasons and decide which reasons she will act upon. But notice that deciding to act on constitutive reasons for extrinsic reasons requires more than the ability to step away from one's inclinations and consider whether one has worthy reasons for acting. It requires the ability to settle this question for extrinsic reasons without those reasons appearing in the operative justification of one's action. Doing this, it seems, requires a division between minds, or between parts of a mind. One must ensure that the extrinsic reasons, for which one decides to bring it about that one acts on the constitutive reasons, are not taken to bear on whether the constitutive reasons are reason enough to act. To so do, it seems one must hold apart the (sub)agent who is, for extrinsic reasons, authoritatively governing whether one acts on constitutive reasons, from the (sub)agent who is acting on the constitutive reasons—because the (sub)agent who is doing the extrinsic governance must take the question of whether one acts to be subject to constraints or justifications that the (sub)agent who is acting on the constitutive reasons cannot recognize. Such an agent is, in fact, lacking in integrity—in the most literal sense.[30]

[30] This agent displays what Donald Davidson takes to be characteristic of irrationality: she is caused to act on a reason that does not rationalize (?) her action. He argues that this requires a divided mind. See his

I suggest, then, that Williams' deeper criticism points to a problem about relying on extrinsic reasons. Williams thought that the impartial or instrumental concern of the utilitarian or the Kantian forced a kind of detachment that then led to an incoherence in the agent. The moral reasons, for being impartial or instrumental, could not be integrated into the point of view from which one performs the projects required for moral life. The underlying worry, I suggest, is that moral reasons, for being impartial or instrumental, are extrinsic to the actions they are supposed to govern. If a moral theory asks an agent to allow reasons extrinsic to some action or project to govern whether she performs that action or participates in that project, it will be asking the agent to divide her mental life.

Because this deeper criticism concerns extrinsic reasons rather than external ones, it will not be resolved by either simply reeducating the agent's motivational set or claiming that a rational agent is able to act on the requirements of impartial reasons, regardless of her particular, contingent desires. One would rather have to either defend this division of the mental life of the agent or else avoid reliance on extrinsic reasons. So long as it is agreed that morality governs actions or projects that are undertaken for certain characteristic, constitutive reasons, and yet governs them by reference to reasons that are extrinsic to them, we will face a problems about integrity.

"Paradoxes of Irrationality," in *Philosophical Essays on Freud*, ed. Wollheim and Hospers (Cambridge: Cambridge University Press, 1982): 289–305, and "Deception and Division," in *The Multiple Self*, ed. Jon Elster (Cambridge: Cambridge University Press, 1986): 79–92. (Both are reprinted in CITE.)