EXPLANATION, INTERNALISM, AND REASONS FOR ACTION*

By David Sobel

I. Introduction

These days, just about every philosophical debate seems to generate a position labeled *internalism*. The debate I will be joining in this essay concerns reasons for action and their connection, or lack of connection, to motivation. The internalist position in this debate posits a certain essential connection between reasons and motivation, while the externalist position denies such a connection. This debate about internalism overlaps an older debate between Humeans and Kantians about the exclusive reason-giving power of desires. As we will see, however, while these debates overlap, the new debate is importantly different from the old debate.

Bernard Williams inaugurated the new debate about internalism. He argued that genuine reasons must be *internal*, that is, they must have a certain specified connection to the motivations of the agent whose reasons they are purported to be. Williams tells us that the most fundamental arguments for internalism stem from what I will call his *explanation condition*. Before we get bogged down in the attempt to formulate precisely what the explanation condition and internalism amount to, however, let me offer a quick road map of this essay. In this essay, I will try to reach a better understanding of (1) the thesis that Williams has labeled internalism, (2) the "interrelation of explanatory and normative reasons" that Williams claims exists, and (3) how Williams thinks (2) helps establish (1). I will argue that Williams's claim that reasons must be interrelated with explanation in a particular way, that is, his explanation condition, does not support internalism as he supposes. Furthermore, I will argue that Williams's explanation condition is false. Finally, I will argue that internalism is false.

The essay will have two major parts. In Section II, I try to understand the explanation condition and argue that it cannot be the key underpinning of internalism. I will argue that plausible interpretations of the explanation condition are either too weak to rule out externalism or so strong that they amount to the thesis of internalism itself. I also take issue

* I am grateful to David Copp, Janice Dowell, and Mike Weber for valuable comments on this essay. I am also grateful for helpful comments from the other contributors to this volume, Ellen Frankel Paul, and Carrie-Ann Biondi.

in this section with the way in which Williams argues to the conclusion that sound deliberation involves knowing the facts of the matter but not being motivated toward prudence or morality.

In Section III, I argue that the explanation condition and internalism are both false. The explanation condition and internalism posit a particular relationship between motivation and reasons for action. I offer arguments for resisting this particular understanding of the relationship. As I shall show, however, resisting this particular understanding of the relationship is compatible with maintaining that there is nonetheless a fundamental connection between motivation and reasons. The arguments I offer against the explanation condition and internalism do not tell generally against *subjectivism about reasons for action*—the view that it is the agent's subjective motivational set that makes it the case that an agent does or does not have a reason to ϕ. Rather, I argue that the best version of subjectivism must reject the explanation condition and internalism.

II. Williams and Internalism

A. The explanation condition as a motivation for internalism

Williams understands internalism to be the view that "A has a reason to ϕ only if he could reach the conclusion to ϕ by a sound deliberative route from the motivations he already has. The externalist view is that this is not a necessary condition, and that it can be true of A that he has a reason to ϕ even though A has no motivation in his motivational set that could, either directly or by some extension through sound deliberation, lead him to ϕ."[1] Williams argues in favor of internalism by trying to show us how only internal reasons can properly capture and respond to the force of the explanation condition.

Williams tells us that there are "two fundamental motivations for the internalist account" of reasons for action.[2] The first is what I have been calling the explanation condition. The second fundamental motivation turns out to be "another application of the same point" insisted upon by the explanation condition.[3] Thus, understanding and assessing the explanation condition is, Williams strongly insists, pivotal to understanding his case for internalism.

But how should we understand Williams's explanation condition? Here are the two most helpful passages in which Williams discusses it:

> [A fundamental motivation of the internalist account] is the interrelation of explanatory and normative reasons. It must be a mistake

[1] Bernard Williams, "Internal Reasons and the Obscurity of Blame," in Williams, *Making Sense of Humanity* (Cambridge: Cambridge University Press, 1995), 35.

[2] Ibid., 38.

[3] Ibid., 39.

> simply to separate explanatory and normative reasons. If it is true that A has a reason to ϕ, then it must be possible that he should ϕ for that reason; and if he does act for that reason, then that reason will be the explanation of his acting. So the claim that he has a reason to ϕ—that is, the normative statement 'He has a reason to ϕ'—introduces the possibility of that reason being an explanation. . . .[4]

> In considering what an external reason statement might mean, we have to remember . . . the dimension of possible explanation, a consideration which applies to any reason for action. If something can be a reason for action, then it could be someone's reason for acting on a particular occasion, and it would then figure in an explanation of that action. Now no external reason statement could by itself offer an explanation of anyone's action. Even if it were true (whatever that might turn out to mean) that there was a reason for Owen to join the army, that fact by itself would never explain anything that Owen did, not even his joining the army. For if it was true at all, it was true when Owen was not motivated to join the army. The whole point of external reason statements is that they can be true independently of the agent's motivations. But nothing can explain an agent's (intentional) actions except something that motivates him so to act.[5]

Williams takes it that there is at least a necessary condition on a consideration providing a normative reason for action—namely, that that consideration has a special kind of explanatory power. This claim is what I am calling the explanation condition. What exactly is the explanatory power that a consideration must have if it is to be able to generate a normative reason? One thing is obvious: the consideration need not be able to explain an actual action. To suppose otherwise is to suppose that a person could not fail to act as her normative reason instructed. Williams is clear that he rejects such a view. Rather, the consideration must be in some sense capable of explaining action. Capable, however, in what sense?

Let us focus on Williams's claim that "If it is true that A has a reason to ϕ, then it must be possible that he should ϕ for that reason." We could, somewhat dimly, understand this merely to mean that it is a necessary condition of A having a reason to ϕ that there be a possible world in which A ϕs. I will call this thesis *Explanation I*. This is independently plausible and is ensured by the principle that "ought implies can," but it is not what Williams is after. The Explanation I formulation fails to distinguish, in the way Williams means to distinguish, considerations that can ground the truth of the claim that A has a reason to ϕ from considerations that cannot do so. Williams wants to be able to say that even if

[4] Ibid., 38–39.

[5] Bernard Williams, "Internal and External Reasons," in Williams, *Moral Luck* (Cambridge: Cambridge University Press, 1981), 107.

Owen does have some sort of reason to join the army, a consideration (e.g., of family honor) that does not appeal to anything in Owen's subjective motivational set cannot ground this reason.

What does Williams mean to add to Explanation I with the thought that for consideration C to support a reason for A to ϕ, A must not only be able to ϕ, but must be able to ϕ "for that reason"? It seems to be A's ϕ-ing "for that reason" that brings in the special explanatory element Williams is looking for. Claiming that A has a reason to ϕ does not yet give a potential ground of the normative justifiability of ϕ-ing. Thus, when Williams says that "If it is true that A has a reason to ϕ, then it must be possible that he should ϕ for that reason," it is not immediately clear what the "for that reason" is meant to refer to. The antecedent of the conditional does not appear to specify the right kind of thing such that it makes sense to say that one could do anything "for that reason."

I think we do best to understand Williams to be saying this: if a consideration C truly provides A with a normative reason (hereafter, reason) to ϕ, then it must be possible that A could ϕ and that at least part of the explanation for his doing so involves his contemplation of and subsequent motivation by C. If we so understand Williams, the possibly explanatory reason in the consequent of Williams's original claim does refer to the sort of thing that could potentially justify A's ϕ-ing. Williams's point seems to be that if ϕ has the status of being something that A has a reason to do, then it must be the case that A can ϕ for the same reasons that give ϕ-ing that status.[6]

Understanding Williams in this way would help us understand the other crucial passage in which he deploys the explanation condition. In the second extract above, Williams writes, "If something can be a reason for action, then it could be someone's reason for action on a particular occasion, and it would then figure in an explanation of that action." The consequent of this conditional invokes an "it" that is supposed both to refer us to the "something" in the antecedent and to be capable of serving as an agent's subjective ground for action. Thus, if the "it" is to be able to play the latter role, the "something" in the antecedent must be understood to be the sort of thing that could stand in the justifying relation. Furthermore, since the conditional has the form of "If it is true that . . . , it must be the case that it can seem to the agent that . . . ," we need to understand the "something" in the conditional's antecedent as something that is truly a reason for action. Therefore, we must again understand Williams as saying that if some consideration C can objectively ground a claim that A has a reason to ϕ, it must be the case that A could ϕ in response to the subjective ground provided by C.

[6] Notice that this formulation anticipates a discussion below to the effect that what Williams seems to really be after is not internalism but *subjectivism*. I conceive of the latter as an account of what makes it true that one has a reason to ϕ rather than merely an account of how to determine if one has a reason to ϕ.

Given this understanding of Williams, we might amend Explanation I so that it expresses the following claim: if consideration C gives A a reason to ϕ, it must be the case that A can ϕ *and* that in some possible world in which A does ϕ, his doing so is explained by his being motivated by C. Let us call this thesis *Explanation II*. This formulation avoids the problem that confronts Explanation I, because it tells us when a consideration lacks the power to provide A with a reason to ϕ. A consideration lacks this power when there is no possible world in which (1) A ϕ's, and (2) her ϕ-ing can be explained by the consideration.

Yet the Explanation II formulation cannot be exactly what Williams means, either, for this version of the explanation condition does not help support internalism. Recall that Williams's version of internalism claims that "A has a reason to ϕ only if he could reach the conclusion to ϕ by a sound deliberative route from the motivations he already has." Explanation II, however, makes no distinction between a consideration motivating via a sound deliberative route and motivating via some other means. For example, this version of the explanation condition would be satisfied if an agent who has no interest in counting blades of grass comes to have such an interest only in those possible worlds in which she undergoes radical brain surgery. According to Explanation II, the possibility of such surgery, and of the subsequent motivation to count blades of grass, means that the considerations in favor of counting blades of grass can provide reasons for A even if he does not actually undergo such surgery, and even though without the surgery he has no interest in his subjective motivational set that counting blades of grass would further. Because of this, Explanation II is a rather weak thesis (although I will find grounds for resisting it in Section III of this essay). Only considerations that could not, in any possible world (even including brain surgery scenarios or the like), motivate A to ϕ would be shown to not provide A with a reason to ϕ.

The version of internalism that Williams wants to argue for claims that consideration C only gives one a reason to ϕ if one could reach the conclusion to ϕ for the reason that C via a sound deliberative route. Yet as we have seen, Explanation II is insensitive to the distinction between A's being motivated by C to ϕ via a sound deliberative route and A's being so motivated in other ways (such as radical brain surgery). Because of this, it is possible to accept Explanation II but reject Williams's internalism. Indeed, this combination of accepting Explanation II and rejecting internalism seems to be John McDowell's view.[7]

Thus, it is unclear how Explanation II could provide the fundamental motivation for internalism and against externalism. One could accept Explanation II but reject internalism by holding Explanation II while denying that it is a necessary condition on consideration C providing A a

[7] John McDowell, "Might There Be External Reasons?" in J. E. J. Altham and Ross Harrison, eds., *World, Mind, and Ethics: Essays on the Ethical Philosophy of Bernard Williams* (Cambridge: Cambridge University Press, 1995).

reason to ϕ that A would be motivated by C to ϕ after sound deliberation. That is, one could hold that a consideration that provides a reason must be able to motivate, but need not necessarily do so after sound deliberation. Because of the availability of this position, it is unclear how Explanation II could be thought to be the key to a defense of internalism against externalism.

As a final option, we could understand the explanation condition to express the claim that a jointly necessary condition of consideration C providing A a reason to ϕ is that (1) A could ϕ; (2) in some possible world in which A ϕs, his ϕ-ing can be explained by means of his contemplation of, and subsequent motivation by, C; and (3) in some possible world in which (1) and (2) are the case, A is deliberating soundly from his actual subjective motivational set. Let us call this thesis *Explanation III*.

Before considering the merits of Explanation III, let us pause to wonder what intuitive basis there could be for resisting Explanation III in favor of Explanation II. The only reason for doing so would have to be that one thinks it is a constraint on good reasons that they motivate after bad deliberation. After all, Explanation II posits a connection between reasons and what can motivate, while Explanation III posits a connection between reasons and what can motivate after sound deliberation. Thus, those who embrace Explanation II but resist Explanation III must champion a connection between good reasons and bad deliberation. I see no intuitive support for such a connection.

Let us turn now to Explanation III. Understood as I define it above, Explanation III just *is* the thesis of internalism, and thus is not a possible motivation for embracing that thesis. (The reader may want to go back and compare Explanation III with Williams's definition of internalism, which I quote at the beginning of this section.) The addition that we saw we needed to add to Explanation II to make it incompatible with externalism (namely, the bit about sound deliberation) was the only difference between Explanation II and internalism. Thus, Williams's explanation condition, which he took to be the centerpiece of his case for internalism, turns out to be, depending on how one interprets it, either too weak to support internalism or to be the thesis of internalism. In either case, Williams's claim that the explanation condition provides crucial support for internalism is misguided.

B. Sound deliberation: why does it involve knowing the facts?

There is another way in which Williams's argument for internalism is misguided. Consider how he argues that "sound deliberation" necessarily involves knowing the facts, but not necessarily being motivated to comply with prudence or morality:

> [I]f we are licensed to vary the agent's reasoning and assumptions of fact, it will be asked why we should not vary (for instance, insert)

> prudential and moral considerations as well. . . . The internalist proposal sticks with its Humean origins to the extent of making correction of fact and reasoning part of the notion of 'a sound deliberative route to this act' but not, from outside, prudential and moral considerations. . . . The grounds for making this general point about fact and reasoning, as distinct from prudential and moral considerations, are quite simple: any rational deliberative agent has in his S [i.e., his subjective motivational set] a general interest in being factually and rationally correctly informed.[8]

There are two problems with Williams's argument here. First, Williams warns against those who claim that "every rational deliberator is committed to constraints of morality." He rightfully tells us that "there has to be an argument for this conclusion. Someone who claims the constraints of morality are themselves built into the notion of what it is to be a rational deliberator cannot get that conclusion for nothing."[9] But Williams offers no argument for the claim that each rational agent has in his S "a general interest in being factually and rationally correctly informed." Williams, then, seems to be trying to get this claim for nothing. Williams may have in mind here some form of the thought that belief necessarily aims at truth and that believers therefore necessarily have an interest in having true beliefs. Such a line could perhaps be made persuasive. But surely Williams expects a champion of the claim that prudential and moral concerns are requirements of rational deliberation to do more than vaguely gesture toward promising argumentative strategies for establishing such a conclusion.

Second, even if the above claim that all rational agents want to be correctly informed could be established, this would not support Williams's conclusion that sound deliberation involves knowing the facts. Note that my argument here is that Williams's premise seems irrelevant to his conclusion; I am not claiming that his conclusion is false. That is, I am not disputing that sound deliberation involves knowing the facts. Rather, I am taking issue with how Williams hopes to argue for that claim.

Williams has championed a connection between what is in our S and our reasons for action. Thus, if Williams could establish that in each agent's S there is necessarily a motivation to be correctly informed, this might help him reach the conclusion that each agent has a reason to become correctly informed. However, this does nothing to make compelling the thought that being correctly informed is a requirement of sound deliberation.

There is a difference between claiming that one's motivations determine one's reasons and claiming that one's motivations determine what

[8] Williams, "Internal Reasons and the Obscurity of Blame," 36–37.
[9] Ibid., 37.

counts as sound deliberation. Williams seems to confuse these things. At least we can say that such a confusion is suggested by Williams's treating the claim that any rational agent has in his S a motivation to be factually informed as if it helped establish the claim that being so informed counts as part of sound deliberation. Consider that even if it could be established that all rational deliberators have an interest in viewing great works of art rather than schlocky knockoffs, this would not show that sound deliberation is deliberation done while viewing great works of art. There is no general reason that Williams offers (or that I can think of) for supposing that if all rational agents want something, then having that thing is necessary for sound deliberation. Surely, then, we might be convinced that Williams's internalism is correct—that our reasons are constrained by what we could be motivated to pursue after sound deliberation—without supposing that the content of sound deliberation is also determined by the agent's motivations.

There is a more plausible variant of the kind of position that Williams seems to be advocating here. Connie Rosati has recently proposed a version of internalism (in her case, internalism about a person's good) that she calls *two-tier internalism*. The central thought behind two-tier internalism is that for ϕ-ing to be good for an agent, not only must the agent be able to care about ϕ in some set of counterfactual conditions, but those counterfactual conditions themselves must answer to her concerns. That is, the appropriate counterfactual conditions in which an agent's reactions determine her good themselves must be such that the agent finds that her reactions in those counterfactual conditions are authoritative. Put most simply, Rosati's proposal has it that an agent's concerns not only determine her good, but that they also determine the appropriate way for the agent to be idealized such that her reactions from that idealized vantage point determine her good. She writes:

> [C]ounterfactual conditions C are appropriate only if the fact that a person would come to care about something X for her actual self when under C is itself something that she would care about while under ordinary optimal conditions. [We achieve "ordinary optimal conditions," Rosati says, when we are in "whatever normally attainable conditions are optimal for reflecting on questions about what to care about. . . ."] A person need not care about X itself while under ordinary optimal conditions in order for X to be good for her. But if her good is not to be alienated, the fact that she would care about X for her actual self under a particular set of counterfactual conditions had better be something that would prompt her concern under ordinary optimal conditions.[10]

[10] Connie Rosati, "Internalism and the Good for a Person," *Ethics* 106, no. 2 (1996): 307.

Rosati explicitly attempts to argue that the same thoughts that led internalists to internalism should lead them to two-tier internalism, and thus that we should allow an agent's concerns to shape what counts as a sound deliberative route for her. I cannot adequately explore Rosati's fascinating proposal here, but notice that she does not allow just any aspect of an agent's concerns to shape what counts as sound deliberation for that agent. Rosati specifically assigns this role to the agent's concerns about what forms of deliberation she finds authoritative.

If Williams or internalists generally want to argue that we can look to an agent's subjective motivational set to help shape what counts as sound deliberation, I think they would do better to follow Rosati's proposal rather than Williams's tacit suggestion. My own hunch, however, is that the internalist is best advised to sever the connection between what counts as sound deliberation and an agent's subjective motivational set. In my opinion, subjectivists like myself should argue that sound deliberation necessarily involves correct factual premises but not prudence or morality; however, I think we should do so on other grounds. That is, we should be subjectivists about reasons for action, but not about the vantage point from which an agent's reactions determine her reasons for action.

The subjectivist, at least by my lights, needs a defense of the thought that sound deliberation involves knowing the facts, a defense that (1) does not depend on finding certain elements in an agent's subjective motivational set, (2) does not also justify counting prudential or moral motivation as a necessary part of sound deliberation, and (3) is continuous with the general subjectivist framework rather than being ad hoc or incorporating objectivist elements. I believe that such a defense can be given, but this is not the place to attempt to make good on this claim.

III. Against the Explanation Condition and Internalism

In the previous section, I claimed that Williams's argumentative strategy for vindicating internalism is flawed in two ways. Of course, this by itself does not show that the explanation condition or internalism is itself false. In this section, however, I will argue for the falsity of both of those theses. I will argue that they are inadequate not because of their connection to neo-Humean subjectivism, but rather as a result of not being compatible with the best versions of such. Seeing the grounds for rejecting the explanation condition and internalism will lead us toward, rather than away from, an adequate subjectivism about reasons for action.

Before we examine these objections to the explanation condition and internalism, we will need to take a brief detour through some literature on well-being. In that literature, one finds an account of well-being that is importantly similar to Williams's internalism about reasons for action. I will call this the *full information account of well-being*. John Stuart Mill's competent-judges test offered an early model of the account, Henry Sidg-

wick offered perhaps its first explicit formulation, and Richard Brandt, R. M. Hare, John Rawls, David Gauthier, James Griffin, Stephen Darwall, David Lewis, Peter Railton, and John Harsanyi have each developed and/or endorsed the view.[11] Roughly, the picture is this: an agent's life goes best if she gets those things that she would want if she had full knowledge of the options available.[12]

Sidgwick's formulation of the account went like this:

> [A] man's future good on the whole is what he would now desire and seek on the whole if all the consequences of all the different lines of conduct open to him were accurately foreseen and adequately realized in imagination at the present point in time.[13]

This formulation quickly runs into difficulties. For example, consider that even though our fully informed self would never want more information for itself, we are firmly convinced that sometimes it can be intrinsically in our interest to gain information. The fact that the fully informed agent lacks a desire for information clearly does not threaten the thought that it would be good for a noninformed agent to get information. Furthermore, our fully informed selves no doubt have a refined palate, and may well highly value expensive complex wines that taste just like the cheaper stuff to us. Yet it is implausible that one wine is much better for

[11] John Stuart Mill, *Utilitarianism* (Indianapolis, IN: Hackett, 1979), chap. 2; Henry Sidgwick, *The Methods of Ethics*, 7th ed. (Indianapolis, IN: Hackett, 1981), 111–12; Richard Brandt, *A Theory of the Good and the Right* (Oxford: Clarendon Press, 1979), 10, 113, 329; R. M. Hare, *Moral Thinking* (Oxford: Clarendon Press, 1981), 101–5, 214–16; R. M. Hare, "Replies," in Douglas Seanor and N. Fotion, eds., *Hare and Critics: Essays on "Moral Thinking"* (Oxford: Clarendon Press, 1990), 217–18; James Griffin, *Well-Being* (Oxford: Oxford University Press, 1986), 11–17; John Rawls, *A Theory of Justice* (Cambridge, MA: Harvard University Press, 1971), 407–24; David Gauthier, *Morals by Agreement* (Oxford: Clarendon Press, 1986), chap. 2; Stephen Darwall, *Impartial Reason* (Ithaca, NY: Cornell University Press, 1983), pt. 2; Peter Railton, "Facts and Values," *Philosophical Topics* 14, no. 2 (1986): 5–31; David Lewis, "Dispositional Theories of Value," *Proceedings of the Aristotelian Society*, n.s., 63 (1989): 113–37; John Harsanyi, "Morality and the Theory of Rational Behavior," in Amartya Sen and Bernard Williams, eds., *Utilitarianism and Beyond* (Cambridge: Cambridge University Press, 1973), 55. Several important caveats apply to some of the above authors' commitments to subjectivism, and some would decline the label. Robert Shaver raises some of these caveats in the case of Sidgwick. See Robert Shaver, "Sidgwick's False Friends," *Ethics* 107, no. 2 (1997): 314–20; see also David Sobel, "Reply to Shaver," published in 1997 in the e-journal BEARS, available at http://www.brown.edu/Departments/Philosophy/bears/9707sobel.html [posted on July 7, 1997].

[12] In David Sobel, "Well-Being as the Object of Moral Consideration," *Economics and Philosophy* 14, no. 2 (1998): 249–83, I consider ways that such a theory could try to respond to the fact that some of our concerns are moral or quasi-moral and hence not perfectly correlated with our well-being. I conclude that any such method will reveal that well-being is not the appropriate object of moral concern. I defend instead the *autonomy principle*, which would allow agents to throw the weight they are granted in moral reflection where they informedly see fit. For a different take on similar issues, see Stephen Darwall's "Self-Interest and Self-Concern," *Social Philosophy and Policy* 14, no. 1 (1997): 158–78.

[13] Sidgwick, *The Methods of Ethics*, 111–12.

me than another when I cannot tell the difference (assuming that it is only the taste of the expensive wine that causes our idealized self to prefer it over the cheaper stuff).[14] Hence, the presence of this desire in the informed agent does not give us grounds to suppose that satisfying this desire would be good for the nonidealized agent. In both of these examples, we are presented with a major problem in Sidgwick's formulation: the idealization process he postulates turns us into such different creatures that it would be surprising if the well-being of one's informed self and one's ordinary self consisted in the same things.[15]

In response to problems such as these, Railton has revised the full information account, proposing that

> an individual's good consists in what he would want himself to want, or to pursue, were he to contemplate his present situation from a standpoint fully and vividly informed about himself and his circumstances, and entirely free of cognitive error or lapses of instrumental rationality.[16]

The adoption of a "wanting to want" framework neatly eschews the implausible identification of interests between informed and ordinary selves, while retaining the insight that "the advice of someone who has this fuller information, and also has the deepest sort of identification with one's fate, is bound to have some commending force."[17]

Railton's move here with respect to discussions of well-being has been duplicated to some extent by Michael Smith in the sphere of reasons for action. Smith claims:

[14] I take this example from Griffin, *Well-Being*, 11.

[15] I have presented these reasons for moving from a Sidgwickian view (and to a Railtonian view—see below) in David Sobel, "Full Information Accounts of Well-Being," *Ethics* 104, no. 4 (1994): 784–810.

[16] Railton offers this account in "Facts and Values," 16. But see ibid., 25 and Peter Railton, "Moral Realism," *Philosophical Review* 96, no. 2 (1986): 175–76 n. 17, for the claim that this account merely "tracks" one's good, that is, while the account shows what an agent's good is, it is not the case that an agent's good is her good because it fulfills the account's criterion. (I discuss this distinction in more detail later in this section.) Notice that Railton's compelling claim that it would be "an intolerably alienated conception of someone's good to imagine that it might fail in any way to engage him" (Railton, "Facts and Values," 9), is compatible with the claim that the full information account merely tracks one's good. In his more recent work, Railton claims that the subjective reactions from the approved vantage point are indicators of the presence of a fit between an individual and an end. See Peter Railton, "Aesthetic Value, Moral Value, and the Ambitions of Naturalism," in Jerrold Levinson, ed., *Aesthetics and Ethics: Essays at the Intersection* (Cambridge: Cambridge University Press, 1998).

[17] Railton, "Facts and Values," 14. Consider, however, that our idealized self could want our ordinary self to want X because the idealized agent knows that our ordinary self's doing so will be instrumentally effective in bringing about, albeit unintentionally, Y, something that the idealized agent finds to be best for our ordinary self. If we say that what is good for our ordinary self is what our idealized self wants our ordinary self to want, we seem to misdescribe these cases of indirection. Perhaps it would be better to focus on the kind of life that the idealized agent wants the ordinary self to have.

> [W]hat it is desirable for us to do in certain circumstances—let's call these circumstances the 'evaluated possible world'—is what we, not as we actually are, but as we would be in a possible world in which we are fully rational—let's call this the 'evaluating possible world'—would want ourselves to do in those circumstances. That is, it tells us that facts about the desirability of acting in certain ways in the evaluated world are constituted by facts about the desires we have about the evaluated world in the evaluating world.[18]

Let us, following a convention Railton uses in his essay quoted above, call the actual person whose reasons we are investigating A, and the idealized version of A, who engages in ideally sound deliberation, A+. Railton and Smith argue that to determine A's good or reasons for action, we should consult A+'s advice for A rather than what A+ himself finds motivating. Railton and Smith have fairly definite ideas about what ideally sound deliberation looks like. However, we need not agree with them on these matters to take the point that the ideally sound deliberator is best viewed as an advisor rather than as someone who will himself be motivated toward that which A has a reason to get. Sidgwick and Brandt, notably among others, do conceive of the idealized agent as someone who would himself be motivated to ϕ if and only if it is good for A to ϕ, and this can be seen to be a mistake even if we disagree with Railton and Smith about the contours of ideally sound deliberation.

The process of becoming an ideally sound deliberator can turn an agent into someone whose reasons for action differ from those of the agent's nonidealized self.[19] That is, the process of changing A into A+ can alter the reasons for action that this person has. If we are to look to A+ to determine A's reasons for action, we must take care, lest A's reasons that are present because he is a nonidealized agent get lost or altered. We are not interested in what reasons for action A+ has. This is why it is best to think of the idealized agent as an advisor. Partially as a result of these considerations, I think the move to *ideal advisor accounts* is a clear improvement over views such as Sidgwick's, Brandt's, and Williams's, each of which grants normative status to what A+ himself is motivated to do.[20]

[18] Michael Smith, *The Moral Problem* (Oxford: Blackwell Publishers, 1994), 151.

[19] This highlights a rather general problem for conditional theories. See Robert K. Shope, "The Conditional Fallacy in Contemporary Philosophy," *Journal of Philosophy* 75, no. 8 (1978): 397–413; and Robert K. Shope, "Rawls, Brandt, and the Definition of Rational Desires," *Canadian Journal of Philosophy* 8, no. 2 (1978): 329–40. I am grateful to Steve Darwall for these references.

[20] I take the useful term "ideal advisor account" from Connie Rosati, "Persons, Perspectives, and Full Information Accounts of the Good," *Ethics* 105, no. 2 (1995): 296–325. Rosati goes on in that paper to critique such accounts. I critique such accounts in "Full Information Accounts of Well-Being." Although both of these papers are critical of such accounts, both agree that the move from the simpler accounts (we might call them *direct motivational accounts*) to ideal advisor accounts is a step in the right direction. Although both papers' critiques are offered against full information accounts of well-being, they are equally effective against full information accounts of reasons for action.

If we are persuaded by these sorts of considerations to look to A+ as an advisor to A, then Williams's explanation condition and his formulation of internalism are both threatened. Let us start with the explanation condition. As noted in Section II, Williams claims that "If it is true that A has a reason to ϕ, then it must be possible that he should ϕ for that reason," and that "If something can be a reason for action, then it could be someone's reason for acting on a particular occasion, and it would then figure in an explanation of that action." In Section II, we had some difficulty generating a precise formulation of these thoughts such that they could play the role that Williams wanted them to play. But now we are in a position to see that it is the central idea here, not just a particular formulation, that is false. The central idea of Williams's claims, I take it, is that if consideration C provides A a reason to ϕ, then it must be the case that A could ϕ because C motivated him to ϕ.

If I have a reason to ϕ, then I have a reason to ϕ in the actual world. Thus, Williams's explanation condition might be thought to express the claim that the consideration that makes it true that I have a reason to ϕ in the actual world must be able to explain my ϕ-ing in the actual world. But counterfactual sound deliberation by A+ in some other possible world, and A+'s subsequent motivation to recommend to A that he should ϕ, cannot explain A's ϕ-ing in the actual world. A might, for example, lack epistemic access to the information that A+ has, with the result being that A could not act for the considerations that the information makes available to A+.

It will rightly be objected that this by itself does not threaten the explanation condition, since that condition need not specify that it is A's actions in the actual world that must be able to be explained. But if these are not the actions that the condition requires be explicable, to what actions does the condition refer? Perhaps the thought is that if consideration C truly provides A with a reason to ϕ, then C must be able to explain A's action after A has deliberated soundly. However, the above analysis of ideal advisor views makes clear that A+—who *is* A after ideally sound deliberation—need not himself be motivated or take action toward that which A has a reason to do. The fact that for C to provide A with a reason to ϕ, A+ must in some sense recommend to A that he ϕ on the ground provided by C does not support the claim that C could explain A's or A+'s ϕ-ing. Thus, on ideal advisor views, it can be true that consideration C provides A with a reason to ϕ without it being the case that C could explain A's or A+'s ϕ-ing. Therefore, adherents to the most plausible versions of subjectivism about reasons for action must reject Williams's explanation condition.

Put in a different way, the problem with the explanation condition is that it cannot accommodate the existence of what I will call *fragile reasons*. One has a reason to ϕ, at least according to ideal advisor views, if one's ideally informed self would in some sense recommend ϕ-ing to one's

actual self. One's reason to ϕ is fragile if the process of becoming ideally informed results in the ideally informed agent lacking a reason to ϕ. I call such reasons fragile because the process of becoming an ideally sound deliberator destroys them. To put this in terms of A and A+, we can say that A's reason to ϕ is fragile if and only if A has it but A+ lacks it.[21] For example, suppose there is a distinctive taste that, once one has tasted it, one is glad to have done so but has no desire to do so again. After one has tasted it, one would recommend to versions of oneself that have not tasted it to try it, but considering the taste itself could never motivate one, whether informed or not, to try it.

There are likely to be fragile reasons when the considerations that ground the reasons for A to ϕ involve the fact that A is not ideally epistemically situated to determine his reasons. The fact that extreme alterations are needed to make a person an ideal advisor suggests that it will frequently be the case that the reasons of A will differ from the reasons of A+. Thus, I suspect that fragile reasons are common.

Reasons can be so fragile that the only vantage points from which one could appreciate the way in which ϕ-ing furthers something in the actual agent's subjective motivational set are vantage points in which one lacks a reason to ϕ. These are what I will call *superfragile reasons*. Superfragile reasons are reasons that one cannot have and be motivated by simultaneously. The case of the singular taste offered above might be an example. We should expect superfragile reasons when appreciating the considerations that make it true that ϕ-ing would further something in A's subjective motivational set itself makes it the case that the agent who so appreciates these considerations himself lacks a reason to ϕ. If there are superfragile reasons, then there are cases in which no vantage point that a person could take up would be such that from that vantage point a person would both have the reason to ϕ and be motivated by the consideration that gives rise to that reason.

Fragile reasons are the key to my rejection of Explanation III and internalism. Superfragile reasons are the key to my rejection of Explanation II. Because the existence of superfragile reasons is more contentious than the existence of fragile reasons, my case against Explanation II is weaker than my case against Explanation III and internalism. However, as I argued in the previous section, I see no intuitive basis for Explanation II except that which is better captured by Explanation III.

The existence of fragile and superfragile reasons shows us that it is a mistake to insist that the same consideration that provides one with a reason must also be able to explain action in accord with that reason. This, it seems to me, strikes at the heart of the explanation condition and

[21] There will, of course, also be cases in which A lacks a reason to ϕ but A+ has one. However, the example of fragile reasons as I define them in the text is sufficient to make my case.

internalism, and shows us that both are just wrong. Yet giving up the claim that reasons and motivations are connected in this way does not force us to give up the thought that one's reasons are determined by one's subjective motivational set.

It might be that there cannot be fragile considerations that ground the fact that it would be rational for A to ϕ. Considerations that ground rationality claims cannot so radically exceed the ken of the agent as can the considerations that ground his reasons. Rationality is a matter of making good use of the information that one has or could reasonably be expected to get. Thus, considerations that it was reasonable for one to be unaware of cannot undermine the claim that one was rational to ϕ. Claims about rationality, then, should be relativized to take into account the agent's predicament and epistemic situation.

Williams's project does not engage in this sort of relativization. On the contrary, the deliberation that Williams claims can close the gap between our current motivations and our genuine reasons is deliberation that, in many cases, we are unable to carry out. Often, for example, the relevant facts that one would need in ideal deliberation have not yet been discovered. Additionally, the deliberation that Williams thinks can close the gap will in many cases be deliberation that it would be impractical for actual people to pursue.

Whatever the merits of his proposal, Williams is hoping to capture the sense of having a reason to ϕ in which one might retrospectively say of oneself, "I had a reason all along to ϕ and did not realize it or have any reason to suspect it until now." In everyday parlance, we do speak as if we could have had a reason to ϕ "all along" even if we had never had any information that would have made ϕ-ing a rational choice at the time. This shows that reason claims, unlike rationality claims, need not be relativized to the agent's epistemic predicament. Therefore, reason claims are significantly more likely to be fragile than are rationality claims. It is thus important to my case that Williams is offering an account of reasons, not an account of rationality.[22]

Let us turn now from considering how the move to ideal advisor views undermines the explanation condition to the issue of how it undermines internalism. As noted above, internalism is the claim that "A has a reason to ϕ only if he could reach the conclusion to ϕ by a sound deliberative route from the motivations he already has. The externalist view is that this is not a necessary condition, and that it can

[22] I make this case much more fully in David Sobel, "Subjective Accounts of Reasons for Action," *Ethics* 111, no. 3 (2001): 461–92. I also argue in that essay that attention to the distinction between an account of reasons and an account of rationality undermines Christine Korsgaard's case against the instrumentalism of Hume and Williams that she offers in Christine Korsgaard, "Skepticism About Practical Reason," in Korsgaard, *Creating the Kingdom of Ends* (Cambridge: Cambridge University Press, 1996); and Christine Korsgaard, "The Normativity of Instrumental Reason," in Garret Cullity and Berys Gaut, eds., *Ethics and Practical Reason* (Oxford: Oxford University Press, 1997).

be true of A that he has a reason to ϕ even though A has no motivation in his motivational set that could, either directly or by some extension through sound deliberation, lead him to ϕ." Thus, as an internalist, Williams supposes that if A has a reason to ϕ, it must be the case that, via sound deliberation, A could reach the conclusion that he himself ought to ϕ. The problem again is that there need not be, at least according to ideal advisor views, a single version of A who both (1) has a reason to ϕ, and (2) is himself motivated to ϕ or would conclude that he ought to ϕ after sound deliberation. According to ideal advisor views, the crucial motivation that we should fix on is what A+ recommends to A. The normatively special motivation is, on such views, not a motivation or conclusion for A himself to take action toward ϕ.

Thus, it will sometimes be true that even though A "has no motivation in his motivational set that could, either directly or by some extension through sound deliberation, lead him to ϕ," the best subjectivist accounts of reasons for action will nonetheless claim that A has a reason to ϕ. Fragile reasons work like this. Therefore, fragile reasons are, according to Williams, external reasons. As a result, the best subjectivist accounts of reasons for action must tolerate external reasons as Williams defines them.

Again, the arguments offered here to reject internalism emerge from an acceptance of subjectivism about reasons for action. The thought is that the most plausible version of subjectivism must follow the path of ideal advisor views. When we follow this path, we recognize that it is not a necessary condition on consideration C providing A a reason to ϕ that there be any particular version of A that can conclude via a sound deliberative route that he ought to ϕ.

Williams is arguing for internalism rather than subjectivism. Williams's internalism is compatible with either a "tracking" or a "truth-making" interpretation.[23] To understand this distinction, consider the American holiday of Groundhog Day. On February 2 of each year, groundhogs are observed as they emerge from their holes; if, rather than venturing outside, the groundhogs return to their holes (upon being scared by their shadow, as I hear it), it is said to mean that there will be six more weeks of winter. Now, it is reasonably clear that the groundhogs' behavior is not thought to make the winter linger. We cannot blame the cold on them, for this would be to blame the messenger. Rather, the groundhogs' behavior is claimed to be a reliable guide to the weather.

[23] Stephen Darwall's formulations of *existence internalism* (Darwall, *Impartial Reason*, 55) and *metaphysical internalism* (Stephen Darwall, "Reasons, Motives, and the Demands of Morality: An Introduction," in Stephen Darwall, Allan Gibbard, and Peter Railton, eds., *Moral Discourse and Practice* [New York: Oxford University Press, 1997], 308–9) are both, like Williams's formulation of internalism, put in terms of necessary conditions for being a reason. Thus, these versions of internalism that Darwall describes are also subject to the importantly different interpretations mentioned in the text. Darwall briefly notes this ambiguity in the latter discussion.

Tracking internalism holds that one's informed pro-attitude toward ϕ-ing is similarly just a reliable guide to one's reasons, not what makes it the case that one has a reason to ϕ.[24] It is thus compatible with objectivism rather than subjectivism about reasons for action. Objectivism and subjectivism, in this context, are theses about what makes it the case that one has a reason to ϕ. If an account claims that the answer to this question is not to be found in the agent's contingent pro-attitudes, then it counts as a version of objectivism. On the other hand, *truth-making internalism* embraces the subjectivist's claim that what makes it the case that one has a reason to ϕ is that one has the relevant informed pro-attitude toward ϕ-ing. Although Williams's defense of internalism is compatible with either the subjectivist or objectivist interpretation, the spirit of his discussion makes clear that he is more inclined to embrace the subjectivist account.

Because of this, I am not inclined to investigate whether or not a successor notion of internalism that avoids the problems discussed above can be found. It seems to me that the interesting philosophical debate here centers on the acceptability of subjectivism rather than the acceptability of internalism. The most philosophically interesting aspect about the debate over internalism has been the debate over what makes it the case that one has a reason to ϕ. Christine Korsgaard, for example, argues that Kantian accounts (and indeed, any philosophically respectable accounts) of practical reason should embrace internalism. She writes, "Practical reason claims, if they are really to present us with reasons for action, must be capable of motivating rational persons. I will call this the internalism requirement."[25] I actually think that this version of internalism is mistaken for reasons completely different from those I have presented in this essay.[26] Yet the point to notice for the moment is that leading proponents of the two fundamentally different accounts of practical reason do not take themselves to differ over the thesis of internalism. The interesting dispute between neo-Humeans like Williams and neo-Kantians like Korsgaard is over what makes it the case that one has a reason for action. The interesting question is hence not whether to embrace internalism or externalism, but whether to embrace objectivism or subjectivism—a debate that may boil down to a dispute about the powers of practical reason to

[24] Michael Smith's account of reasons for action in *The Moral Problem* is best understood as a version of tracking internalism. He thinks that the desires of all ideally rational agents converging on certain things is necessary and sufficient for our having reasons, and in particular reasons to do what our desires converge on. According to Smith, the best explanation for such a convergence, if it occurred, would be that there are "extremely unobvious a priori moral truths" (Smith, *The Moral Problem*, 187). On his view, it is these truths that make it the case that we have reasons to do certain things; our ideally informed deliberations simply get our motivations to track these truths. I critique Smith's arguments for convergence in David Sobel, "Do the Desires of Rational Agents Converge?" *Analysis* 59, no. 3 (1999): 137–47.

[25] Korsgaard, "Skepticism About Practical Reason," 11.

[26] See Sobel, "Subjective Accounts of Reasons for Action."

bring about consensus in the motivations of people who start out with radically different motivations.

IV. Conclusion

Williams writes, "The whole point of external reason statements is that they can be true independently of the agent's motivations." That is, Williams thinks that external reasons would not be essentially relative to the agent's subjective motivational set. But Williams's claim here about externalism is not a necessary consequence of rejecting the explanation condition and internalism. There is room in logical space for resisting the thought that a true reason for A to ϕ must motivate A to ϕ after sound deliberation while accepting that what makes it the case that A has a reason to ϕ is that A "has some motive that will be served or furthered by his ϕ-ing."[27]

It is clear enough why it would seem natural, if one were positing a connection between motivations and reasons, to think that it is a constraint on having a reason to ϕ that one be motivated to ϕ after sound deliberation. After all, should it instead be a constraint on having a reason to ϕ that one be motivated to do something else, say X, after sound deliberation? Yet as we have seen, ideal advisor views better capture the wanted relationship between the sound deliberator and the reasons of nonidealized agents than do views that look to what the sound deliberator is motivated to do or concludes that she has reason to do. When, as a consequence, we embrace an ideal advisor account, we leave behind the thought that if consideration C grounds a reason for A to ϕ, it must be that C could motivate A to ϕ.[28] Thus, in searching for the best understanding of the pro-attitudes that have a fundamental connection to our reasons, we are forced to leave the explanation condition and internalism behind.[29]

Philosophy, Bowling Green State University

[27] Williams, "Internal and External Reasons," 101. This is Williams's casual and "very rough" characterization of internalism in the earlier paper. The formulation of internalism offered in the later "Internal Reasons and the Obscurity of Blame," which I cite at the beginning of Section II of this essay, is clearly intended to be his official "nonrough" characterization of internalism. This formulation is also the sort Williams invokes in Bernard Williams, "Replies," in Altham and Harrison, eds., *World, Mind, and Ethics*, 186–94. Furthermore, the later characterization is the one that has been picked up by subsequent writers on internalism such as Darwall and Korsgaard.

[28] Throughout this essay I have been treating the concepts of 'motivation' and 'desire' as unproblematic so as to focus on other issues. In fact, I find these concepts not yet satisfactorily analyzed. For some initial misgivings, see David Sobel and David Copp, "Against Direction of Fit Accounts of Belief and Desire," *Analysis* 61, no. 1 (2001): 44–53.

[29] Unfortunately, I did not read Robert Johnson's excellent "Internal Reasons and the Conditional Fallacy," *Philosophical Quarterly* 49, no. 194 (1999): 53–71, until it was too late to take it into account here. Johnson offers compelling arguments for some of the central conclusions that I urge in the second half of this essay.