Sound Advice and Internal Reasons

by

Ariela Tubert

Abstract: Reasons internalism holds that reasons for action contain an essential connection with motivation. I defend an account of reasons internalism based on the advisor model. The advisor model provides an account of reasons for action in terms of the advice of a more rational version of the agent. Contrary to Pettit and Smith's proposal and responding to Sobel's and Johnson's objections, I argue that the advisor model can provide an account of internal reasons and that it is too caught up in the psychology of the actual agent to be able to account for anything other than internal reasons.

Reasons internalism is the view that reasons for action bear an essential connection with motivation. In his now classic paper "Internal and External Reasons," Bernard Williams introduces the distinction between internal and external reasons in terms of two conditions: (1) an explanation condition and (2) a subjectivity condition. The *explanation condition* sets as a requirement that reasons be capable of motivating and so of explaining action, while the *subjectivity condition* requires that reasons be properly related to the agent's motivational set (desires, wants, projects, commitments, etc.). On this account, reasons that meet both conditions are internal while reasons that do not meet one or the other are external. In what follows, I am concerned with the prospects for such a view of internal reasons. That is, an account of reasons that are both capable of explaining action and in some way dependent on the agent's motivational set.

The paper proceeds in three parts. In the First Part, I propose an account of internal reasons in terms of the advice of a more rational version of the agent in question and I argue that some but not all aspects of the actual agent's psychology need to be held constant by the advisor as part of the circumstances faced by the agent. Then, in the Second Part of the paper, I discuss

Pettit and Smith's model of reasons as presented in their paper "External Reasons." In this paper, Pettit and Smith propose an account for all reasons – internal and external – on the basis of the advisor model. I argue that their proposal for making sense of external reasons fails because the advisor model is too caught up in details about the agent's psychology for it to deliver anything other than internal reasons. Finally, in the Third Part of the paper, I respond to objections that the advisor model does not allow the proper connection between reasons and motivation and thus fails to be an appropriate model for internalism. I clarify the explanation condition and use this refined notion of the explanation condition together with the lessons of the discussion of Pettit and Smith's account to respond to the objections and show that the advisor model can be a proper model of internalism and can provide an account of both the subjectivity and explanation conditions.

SECTION I

I.1. Williams on Internal and External Reasons: The Subjectivity and Explanation Conditions

Williams' distinction between internal and external reasons can be understood partly in terms of whether the reasons depend in some way on a given agent's *subjective motivational set*. By "subjective motivational set" Williams means an agent's desires, wants, 'dispositions of evaluation, patterns of emotional reaction, personal loyalties, and various projects, as they may be abstractly called, embodying commitments of the agent' (Williams (1979), 105). So, for instance, if someone aspires to be a great writer, loves to jog, is a fierce patriot, or has a craving for chocolate ice cream, these are all things that go into her subjective motivational set.

When Williams claims that an agent 'A has a reason to φ only if there is a sound deliberative route from A's subjective motivational set to A's φ-ing' (Williams (2001), 91), he is claiming that having a reason is dependent on the types of motivations we would have *if we deliberated properly*. Deliberation may involve processes like revising false beliefs or discovering that a certain course of action is the means to a given end. And in fact, the deliberative process may add or subtract elements from the agent's motivational set. So for instance, suppose I have a desire to spend the day at the beach, and I find out that taking the Number 5 bus is the most efficient way to get there. Then upon sound deliberation, my motivational set will come to include a desire to take the Number 5 bus. Given my desire to go to the beach then, I have an internal reason to take the Number 5.[1]

Or, to have a different sort of example – one involving false beliefs – consider Williams' famous case of the person who desires to drink from a glass which he wrongly believes contains gin (Williams (1980), 104). What this person wants is a gin and tonic but what he does not know is that the glass in front of him is actually filled with petrol – not gin. So then, though he presently desires to drink from the glass, if he knew what the glass contained he would lose his desire. Thus, he does not really have a reason to drink from the glass. According to Williams then, claims about what we have reason to do depend on there being a sound deliberative route from our motivational set, where *sound* deliberation cannot be based on false beliefs and it may sometimes require modifications to our current motivational set.

What these examples help show is that internal reasons as conceived by Williams are not meant to be straightforwardly dependent on an agent's current motivational set. We can have an internal reason even if the reason does not refer to a motivation that we presently have, provided that we would acquire such a motivation if we deliberated soundly. Nonetheless, Williams takes

it to be a constraint on reasons that they be related (through sound deliberation) to the agent's current motivational set. This constraint is what I call the *subjectivity condition*.

By contrast, claims about external reasons, by definition, need not depend on there being a sound deliberative route from the agent's current desires, projects, and interests; that is, they need not be dependent on the agent's motivational set. Williams argues that this notion of external reasons makes no sense – thus, he concludes, all reasons are internal. The guiding idea behind Williams' argument against external reasons, however, is that reasons are the sort of things that must be capable of explaining our actions, and this takes us away from the subjectivity condition and to the explanatory condition. His primary objection to external reasons is that 'no external reason statement could by itself offer an explanation of anyone's actions' (Williams (1980), 106). If a consideration provides a reason, then it needs to be the case that the person could act for that reason.

The explanation condition should not be taken to imply that to have a reason, the agent must actually be motivated to act on that reason. We can imagine a situation in which the person does not know some relevant piece of information, that there is petrol in the glass for example, and so that there is petrol in the glass cannot explain her actual actions. This does not mean that the explanation condition is not met in this case – Williams himself thought that the person has a reason not to drink from the glass. Insofar as the person could act on that consideration, the consideration in question could provide the agent with a reason for action.[2]

I.2. Advisor or Exemplar

As we have seen above, given internalism's focus on the motivation and beliefs of the agent, a potential problem for the view is that motivations based on false beliefs could be

wrongly thought to give rise to reasons. To solve this problem, defenders of internalism often appeal to the motivations of more rational versions of the agent in question. However, the resulting view of internal reasons can be criticized on the basis that it commits the *conditional fallacy* and thus that it cannot account for all reasons.[3] The idea behind this objection is that an agent may have reason to seek information or to become more rational, for example, but a fully rational version of the agent would not do or desire to do either, since that agent will already possess all information and be fully rational. Thus internalism faces the criticism that it cannot properly provide an account of the reasons that we have only insofar as we are less than fully rational.

In response to this type of problem some defenders of internalism, like Peter Railton and Michael Smith, have introduced a distinction between two internalist models: that of the *exemplar* and that of the *advisor*.[4] While the former may be subject to problems like the conditional fallacy, the latter is not. So, one way of understanding Williams' claim that an agent A has a reason to φ only if there is a sound deliberative route from A's subjective motivational set to A's φing is to consider whether a more rational version of A would desire to φ. This is the exemplar model. Alternatively, in "Internal Reasons" Smith argues instead that the best way of understanding claims about reasons is by considering what a fully rational *advisor* version of oneself would desire that one do in the circumstances. This advisor model, then, considers the advice of a rational version of oneself rather than what the exemplar model considers – namely, what the rational self would desire to do.[5]

To help understand the contrast between the two models consider Smith's variation on Gary Watson's example of someone who suffers a humiliating defeat in a game of squash and gets so angry that she wants to smash her opponent in the face with her racquet.[6] When we

consider this example on the exemplar model, what we do is imagine a fully rational version of the squash player and ask what they would desire to do in those circumstances. This fully rational agent would presumably desire to go up to her opponent in order to shake her hand. On this model then, the *actual* squash player has a reason to go up to her opponent in order to shake her hand. The problem with this conclusion is that, being less than fully rational, if the actual squash player were to go up to her opponent, she would be unable to resist smashing her opponent in the face with her racquet.

Next, consider the example on the advisor model. According to it, we again imagine a fully rational version of the squash player, but instead of asking what she *herself* would desire to do in the given circumstances, we ask what she would desire that the *actual* squash player do, where as advisor she bears in mind that the squash player is less than fully rational. The idea is that the advisor would recognize that the actual squash player is so angry that if she were to go up to her opponent, she would smash her in the face. So, instead of desiring that the actual squash player go up to her opponent, the advisor would desire that she instead walk off the court and calm herself down. As the example illustrates, the advisor model seems to make best sense of reason claims: The *actual* squash player in fact has a reason to walk off the court and calm herself down, not to go up to her opponent.[7]

I.3. Which aspects of our psychology form part of our circumstances?

An account of reasons based on the advisor model needs to take into account some aspects of the agent's psychology when considering the circumstances that the agent is in. That Jane enjoys reading Russian novels may give her a reason to read *War and Peace* over the summer but Mike who does not enjoy reading novels at all may not have such reason. Similarly, that Irina has a

strong craving for pickles may give her a reason to eat some but John having no such craving is unlikely to have such reason.[8]

So, some aspects of the agent's psychology, like the desire for pickles, needs to be taken into account as part of the circumstances that the advisor needs to consider. However, there is good reason not to include *all* desires as part of the circumstances – the advisor need not take into account my desire to drink from the glass containing petrol, for example. So, some desires need to be included as part of the circumstances while others need to be excluded. The question is, on what basis is a desire included or excluded?

I think that the way to draw the distinction is by appealing to the difference between motivated and unmotivated desires.[9] Motivated desires are those that are rationally dependent on other desires. Motivated desires are connected to other desires through reasoning, like means-ends or constitutive reasoning. So for example, my desire to take the #5 bus is based on my desire to go to the beach and my belief that the #5 takes me there. However if I stopped desiring to go to the beach or if I found out that the #5 does not go to the beach, then my desire to take the #5 should go away. On the other hand, I may have a desire to enjoy a day at the beach which is not itself based on other desires – I may have just woken up with a strong desire to be at the beach – this would be an unmotivated desire.

Similarly, my desire to drink from "this glass" may be motivated by my desire to drink a gin and tonic, but given that there is petrol in there, someone deliberating soundly would not desire to drink from the glass. However, my desire to drink a gin and tonic would be grounds for a desire to drink from a glass containing gin and tonic (though not from a glass containing petrol) as well as going to the store to buy gin. Or, to take a different example, I may desire to succeed at doing something challenging, believe that climbing the Aconcagua is the best way of fulfilling

this desire, and on that basis, come to have a desire to climb the Aconcagua. However, if I were deliberating soundly, I would realize how arduous an undertaking this is and given my other desires and not form the desire to climb the Aconcagua.

There are various ways in which a motivated desire can fail to provide a basis for reasons. Sometimes this happens because the motivating desire has dissipated even while the motivated desire lingers on, with the agent failing to notice that the motivated desire is now baseless. Sometimes it will happen because the motivated desire is causally connected to motivating desires only through false beliefs. Sometimes it will happen because the motivated desire is causally connected to motivating desires only through faulty means-ends-reasoning. The general idea is that motivated desires can be subject to rational scrutiny, and when they are not connected to one's unmotivated desires in a rationally approvable way, they will fail to give rise to reasons. Unmotivated desires cannot be rationally criticized in the same way, since by definition they are not dependent on other desires.[10] That is why the rational advisor is compelled to treat one's unmotivated desires as part of the circumstances to be held constant when giving advice, and it is why motivated desires are to be treated differently.

Once we include some desires as part of the circumstances, we realize that this modified advisor model can account for the connection between reasons and motivation and that reasons are connected to the actual agent's motivation through sound deliberation. So, my proposed model for internalism is as follows:

(i) Agent A has a reason R to φ in circumstances C iff

A+ (the version of A that deliberates soundly) would advise that A φ's in C,

where C includes both what A is capable of doing as well as A's unmotivated desires.

I believe that this model captures the "sound deliberative route from the agent's motivational set" idea from Williams.[11] If A, given her desires and motivational capacities would deliberate soundly about what to do, she would advise herself to φ in the circumstances she is actually facing.[12]

What I have reason to do in the circumstances I am facing, then, depends on what I would recommend that I do from a perspective in which I make no mistakes in reasoning and have access to relevant information – keeping in mind that the advice is to take into account as part of the circumstances what I am capable of doing as well as my unmotivated desires.

SECTION II

In the previous section, I introduced an account of reasons for action based on a combination of Williams' internalist view and the advisor model. I believe that the advisor model supports Williams' internalism, his claim that all reasons are internal reasons. In "Internal and External Reasons," Pettit and Smith give an account that is similar to the one provided above in that it is also based in a combination of Williams' view and the advisor model. However, in a somewhat surprising move, Pettit and Smith argue that their account applies to both internal and external reasons. In this section, I will argue that Pettit and Smith fail to deliver an account of external reasons, as the advisor model is too caught up in certain features of the agent's psychology to deliver anything other than internal reasons. This failure is instructive because it helps to reinforce something close to Williams' original argument against external reasons. In Section III, I will be drawing on the lessons from the discussion of Pettit and Smith's view to respond to some possible objections to the model of internal reasons I proposed in Section I.

Using the advisor model, Pettit and Smith reconstruct Williams' account of internal reasons as composed of three independent conditions, of which the following two are relevant here (155):

> (a) There is the claim that an agent A has internal reason to φ in certain circumstances C only if A+ wants A to φ in C, where A+ has and exercises all of the capacities that ensure that her desires conform to principles of reason.
>
> (b) There is the further claim that A has an internal reason to φ in C only if A has – not would have, but has – the capacity to recognize and respond to the fact mentioned in (a).[13]

The first condition is where the advisor model comes in: whether someone has an internal reason depends on what the fully rational version of that person would, as advisor, desire that she do. The second condition says that to have an internal reason one must actually be able to recognize and respond to the considerations that provide the reason. This second condition draws on Williams' requirement that reasons are the sort of thing that should be able to explain behavior. Pettit and Smith's idea is to use the distinction between the two conditions to argue that (a) by itself, once we eliminate the occurrences of 'internal' within it, can provide a *general* account of reasons, whether internal or external. Internal reasons would then be those which satisfy *both* (a) and (b), while external reasons would be those which satisfy (a) but *not* (b).[14]

On Pettit and Smith's picture, then, the reasons I have – whether internal or external – correspond to the desires that the fully rational version of me has regarding what I, myself (a less than fully rational person), should do given the circumstances I find myself in. Internal reasons are those reasons which correspond to a certain subset of those desires that the fully rational version of me has, namely, those which I myself am capable of recognizing and responding to. External reasons, on the other hand, are those reasons which correspond to a different subset of

those desires that the fully rational version of me has, namely, those which I myself am *not* actually capable of recognizing and responding to.

However, as I am about to argue, Pettit and Smith's attempt to provide a coherent account of external reasons is ultimately unsuccessful. Again, Pettit and Smith claim that external reasons are those that satisfy condition (a) but do not satisfy condition (b). I think that Pettit and Smith give good reasons to accept the advisor model. And so, assuming that their condition (a) provides a general account of reasons, what I will be arguing is that it is implausible that condition (a) could be satisfied while condition (b) is not, *given how Pettit and Smith themselves understand the advisor model.* That is, I will be arguing that Pettit and Smith's own defense of the advisor model undermines their account of external reasons.

To see the problem, let's reconsider the squash player example, this time with a special focus on conditions (a) and (b). On the advisor model, again, we are to consider not what the rational version of A, A+, would desire herself (A+) to do in the circumstances she (A+) finds herself; rather, we are to consider what A+ would desire that A do in A's circumstances. In the case of the squash player, we are to suppose that the actual player who loses the game of squash is so upset that she has an overwhelming desire to smash her opponent in the face with her racquet. By contrast, the fully rational version of this squash player, who is just like the actual squash player but who has and exercises all her rational capacities, would desire to go up to her opponent and shake her hand.

In reflecting on what the actual squash player has reason to do, Pettit and Smith write, '…on the advice model, A has no reason at all to walk over and shake her opponent's hand. Instead what she has reason to do is leave the court without saying a word and calm herself

down' (148). The thought is that the actual squash player has a reason to walk away because the fully rational version of herself would desire that she, the actual squash player, walk away given her circumstances. This, again, is how Pettit and Smith, echoing the earlier argument by Smith, argue for the advisor model over the exemplar model.

The reason I return to this example is because I want to point out that the circumstances that the actual squash player is in – specifically, those circumstances that the fully rational version of the squash player, acting as adviser, is supposed to take into account – seem to include *the actual squash player's current psychological capacities*. The fully rational version of the squash player will desire that the actual squash player walk away, according to Pettit and Smith, because otherwise the actual squash player will be unable to control herself and will hit her opponent with her racquet. But, by building it into the advisor model that the rational version (A+) of an agent A is to take into account A's current capacities, Pettit and Smith seem to be building in conditions which potentially assure that if (a) is satisfied then (b) will be satisfied as well. That is, they seem to be building in conditions that potentially assure that on their account, no reasons are external.

To make the point clear, let me restate part of condition (a) while emphasizing the relevant aspect: an agent A has a reason to φ *in certain circumstances* C only if A+ wants A to φ in C. If A+ were to desire something which A is incapable of doing – such as walking over to her opponent and shaking her hand – then A+ would not be properly taking into account A's circumstances. But, if what A has a reason to do depends on what A+ would desire A to do in A's circumstances, while part of A's circumstances include what A is capable or incapable of doing, then it will *never* be the case that A has a reason to φ even though she is not capable of

recognizing or responding to this fact. That is, it will *never* be the case that A has an external reason.

Here might be a helpful way to picture the present point. Imagine that upon initially considering how to advise A, A+ comes to the view that A should φ – so initially, A+ is inclined to advise A to φ. But then, upon further reflection, A+ realizes that A herself is not in a position in which she is capable of recognizing or responding to those considerations which persuade A+ that A should φ. If so, then that A is incapable in this way would seem to be part of the circumstances C which A is in – circumstances which A+ must take into account in her role as advisor. Just as the fully rational version of the squash player must take into account in her role as advisor that, being less than fully rational, the actual squash player would smash her opponent in the face if she were to go up to her, it would seem that A+ must take into account what sorts of considerations A is capable of recognizing or responding to. The upshot of this is that once A+ fully takes A's circumstances into account, including what A is incapable of, A+ will no longer be inclined to advise A to φ. That is, A+ will not desire that A φ's.

The problem this poses for the proposed account of external reasons is that if what A has a reason to do depends on what A+ desires A to do in her circumstances, while those circumstances include what A is capable or incapable of doing, then it will never be the case that A could have a reason to do something even though she is not capable of recognizing or responding to this fact. But this is just to say that A will never have a reason which satisfies (a) but does not satisfy (b). So, if external reasons are supposed to be reasons which satisfy (a) but not (b), it then follows that there are no external reasons.[15]

If the fully rational version of the agent is to properly do her job as advisor, she will need to take psychological details of the agent into account as part of the circumstances the agent finds herself in, including details about what sorts of considerations the agent is capable of recognizing and responding to, and then advise accordingly. As a result, the advisor model is unable to transcend facts about the agent's psychology in the way external reasons are expected to. Furthermore, this shows that the advisor model can *only* be a model of internal reasons and if this is right, the considerations which lead to the advisor model as opposed to the exemplar model in the first place, namely the conditional fallacy, may also be able to provide an argument against the existence of external reasons.

SECTION III

I want to conclude by considering and responding to two objections to my above proposal. First, given the particular critique of Pettit and Smith that I have advanced, why not just drop the advisor model altogether?[16] The idea here would be that if deliberation is constrained by one's capacities, as I argue in Section II, then we can simply focus on the considerations that one could reach through sound deliberation—that is, focus on the (counterfactual) sound deliberation of the agent, rather than on the advisor. The question is pressing for my view because I add the capacities of the actual agent to the relevant circumstances. This addition seems to take care of the problem raised by the conditional fallacy, that the actual agent may have reasons to do things that the rational agent does not. But solving this problem was the main motivation for turning to the advisor model to begin with. Perhaps once that problem is solved, we can simply drop the advisor model. In short, the question is, what is added by the advisor model that is not already there in the idea of sound deliberation?

My reply is that the advisor model is still needed to avoid the kind of problems addressed by Railton and Smith that I discussed in Section I. There are cases in which the relevant circumstances faced by an actual agent A include that the agent's actual capacities get in the way of her deliberating soundly. It is not possible for us to consider a counterfactual where the agent is picked out by the property of deliberating soundly and reason about what that sound deliberator would do when they are not deliberating soundly. That is, if we drop the advisor model, we would have to imagine what the sound deliberator would do when she is not deliberating soundly – but it is the notion of being a sound deliberator that is essential to our considering what she would do to begin with.[17] The advisor model provides a way around this problem by allowing us to distinguish between the sound deliberator who provides advice and the actual person who is a less than ideal deliberator.[18]

The second objection is that the advisor model does not allow for the proper connection between reasons and motivation thus it fails to be an appropriate model for internalism.[19] Johnson (1997) criticizes the advisor model as a model of internalism:

> …it is misleading to present the advice model as a model of the internalism requirement. The latter connects reasons to motivation, but the advice model does not; it connects reasons to advice. Indeed, it makes no mention at all of motivation of any kind. For the desires to which reasons are connected on the advice model are not desires to do anything, and so are not motivations to do anything either. Reasons on this model are what, were we rational, we would desire that we do in our less than rational circumstances. This is what we might call an "advising" desire: My having a reason implies that my more rational self desires that I do certain things. But this does not imply that anyone at all desires to do those things, rational or not. (Johnson, 621).

The idea is that unlike the exemplar model, which connects reasons with the motivations of rational agents, the advisor model connects reasons with the advice (or desires) of the rational advisor with respect to what the less than rational agent is to do. The advisor model, it seems, does not connect having a reason to φ with having the motivation to φ but with the advice of a

wiser self to φ. Given that internalism holds that there is a necessary connection between reasons and motivation, it would seem that the advisor model is not a model of internalism at all.

Sobel (2001) argues that the advisor model is not able to account for the explanation condition.

> 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. (Sobel, 230)

Sobel argues that there are cases where an agent may have a reason to φ and yet there is no version of that agent (whether rational or not) that is motivated by that consideration to themselves φ. The cases at stake are exactly the kind Smith uses to motivate the advisor model, cases like the squash example where there is a reason (e.g. to get off the court) that the less than fully rational person has which does not motivate the agent (because they are unaware of it, for example) but that the rational version of the agent would not have and would thus not be motivated by either.[20] So, the advisor model cannot account for the explanatory power of reasons as the considerations providing a reason cannot always explain the actions of the less than fully rational person and, unlike on the exemplar model, they cannot always explain the actions of the rational person either.[21]

I believe that, drawing on the lessons of Section II, the model of internal reasons proposed in Section I can account for the explanatory power of reasons. On this model, reasons are connected to motivation because the relevant advice takes into account both what the agent is capable of doing and the unmotivated desires of the agent. This is what I take the explanation condition to require: that the consideration which is to provide a reason be such that although the agent may not actually be moved by the consideration in question, the motivation is in some

sense *already present* in the agent. Notice that I may not be moved to take the # 5 bus because I may not realize that it provides the fastest way to get to the beach. However, I may desire to get to the beach quickly and so have the motivation that would justify taking the # 5. In that case, my desire to get to the beach quickly is something that could motivate me to take the #5 bus, even if in fact it does not. The general idea is that an appropriate model of internal reasons for action should keep the circumstances, including the basic motivations, constant but idealize the reasoning that connects them with action.

So, Johnson objects that the advisor model 'makes no mention at all of motivation of any kind' and that the fact that my more rational advisor would recommend that I do something 'does not imply that anyone at all desires to do those things, rational or not' (621). However, on the present model, motivations play a role in the model because the advisor needs to take unmotivated desires into account in providing advice. Thus the connection between reasons and motivation is established.

Sobel claims that the advisor model can't account for the explanation condition because it 'does not support the claim that C could explain A's or A+'s φ-ing' (230). But on the modified advisor model, the advisor is to take into account what the advisee is capable of doing as well as some of her actual motivations. And as I am about to explain, the proper understanding of what the advisee is capable of doing entails that if a consideration provides a reason, then it could explain A's actions even if it does not explain A's actions in the actual world.

As explained in Section I, the explanation condition says that reasons need be capable of explaining action and thus that the agent *could* act on those reasons.[22] In Section II, we saw Pettit and Smith's characterization of the condition in terms of what the agent is *capable* of recognizing and responding to. However, it is not clear what sense of 'could', or 'capable' we

are working with. If the terms are understood to apply too widely, covering any possible world, then the explanation condition loses its force and provides little or no constraint. If they are understood too narrowly to be only what the agent actually recognizes and responds to, then they are too restrictive by not allowing that one can have a reason that one is not aware of and thus raising doubt on whether the explanation condition can be a condition on reasons at all.

I propose that the explanation condition be understood as follows: Consideration S provides agent A a reason to φ in circumstances C only if there is a possible world where A φ's in C and A's φing is explained by A being motivated by S.[23]

Sobel (2001) argues against a similar version of the explanation condition. His argument is partially based on the claim that it can't play the role Williams wants it to play as part of the argument against externalism.[24] However, we could find that Williams was correct in providing an account of internal reasons while mistaken in his argument against external reasons. In addition, the explanation condition as characterized here together with the subjectivity condition may be able to play a role in an argument against external reasons. As we noted in the discussion of Pettit and Smith's view above, the advisor model of reasons can only provide a model for internal reasons. As long as we take the advisor model seriously then we will find that there are no external reasons, that is, that there are no considerations that can be given as advice and yet that the agent cannot recognize and respond to.[25] So, the considerations that lead us to accept the advisor model over the exemplar model (namely the threat of the conditional fallacy) may be used to support an argument in favor of internal reasons.

Although the idea of dropping the explanation condition can seem attractive, I believe it is a mistake. The explanation condition can help to capture the idea, familiar from Davidson, that reasons are also causes.[26] The account of the explanation condition given here can give

content to the requirement that for a consideration to provide a reason for an agent it must be the case that the consideration can cause an action by the agent or in other words, that the agent can act for that reason. When we think about it in these terms, we can also make sense of the requirement that the advisor take into account the unmotivated desires and recognitional capacities of the agent. The requirement that for a consideration to be a reason for A it needs to be able to cause an action in A would lose meaning if no existent motive is kept constant (in some sense anything could cause A to do something, but not all such cases would be examples of A acting) but if it is so restrictive as to only allow A's existing motives, then it cannot account for the reason giving force of the consideration (my existing motivations may already contain some irrationality.) Taking into account one's unmotivated desires and recognitional capacities makes sense because it allows for the idea that reasons are considerations that could cause an action in this particular individual in the circumstances the person is facing. By holding constant the unmotivated desires and recognitional capacities, we guarantee that there is some motivation already in the agent that could causally lead to the action. I believe this makes sense of the constraint that the reasons for someone should be capable of causing that person's actions.

One may wonder with respect to the squash example discussed in Section II, whether A+ would advice A to go up to the opponent and shake her hand, regardless of whether A is capable of doing that.[27] To the extent that there is any content to the idea of advice, I think that the capacities of the advisee have to be taken into account.[28] We wouldn't advise a typical human being to open their wings and fly to work so as to get there on time, similarly, we wouldn't advise her to act on considerations that she could not respond to. In "Internal Reasons and the Obscurity of Blame," Williams defends a similar idea on the basis of the connection between what one is capable of doing and blame. He says,

> … if 'ought to have' is appropriate afterwards in the modality of blame, then (roughly) 'ought to' was appropriate at the time in the modality of advice. Now, 'ought to' in the modality of advice implies 'can', because advice aims to offer something as a candidate for a deliberative conclusion. If φing is not available to the agent, 'you ought to φ' cannot function as a piece of advice about what he should now do; when it is a matter of what I am to do, manifestly 'I cannot' act as a stopper." (Williams (1995a), 40)

Regardless of whether we want to maintain the connection with blame after the fact, it seems right that 'I cannot' acts as a stopper for advice. While one may wish that a friend could fly to work or avoid getting uncontrollably angry after losing a game of squash, realizing that she is incapable of doing it would have an effect on the advice one would give her and thus the advice model is properly a model of internalism – even if internalism is false and the advice model in the end is not a proper model for all reasons.[29]

Department of Philosophy
University of Puget Sound

References

Darwall, Stephen L. (1992). "Internalism and Agency," *Philosophical Perspectives* 6, pp. 155-172.

Davidson, Donald (1963). "Actions, Reasons, and Causes," reprinted in *Essays on Actions and Events*, Oxford: Oxford University Press (2001).

Johnson, Robert Neal (1997). "Reasons and Advice for the Practically Rational," *Philosophy and Phenomenological Research* 57: 3, pp. 619-625.

Johnson, Robert Neal (1999). "Internal Reasons and the Conditional Fallacy," *Philosophical Quarterly* 49: 194, pp. 53-71.

Nagel, Thomas (1979). *The Possibility of Altruism*, Princeton: Princeton University Press.

Pettit, Philip and Smith, Michael (2006). "External Reasons", Macdonald, C. and Macdonald, G. eds., *McDowell and His Critics*, Oxford: Blackwell.

Railton, Peter (1986). "Moral Realism," *Philosophical Review* 95, pp. 163–207.

Robertson, Teresa (2003). "Internalism, (Super)fragile Reasons, and the Conditional Fallacy," *Philosophical Papers* 32: 2, pp. 171-184.

Rosati, Connie (1995). "Persons, Perspectives, and Full Information Accounts of the Good," *Ethics* 105, pp. 296-325.

Shope, Robert (1978). "The Conditional Fallacy in Contemporary Philosophy," *Journal of Philosophy* 75, pp. 397–413.

Smith, Michael (1994). *The Moral Problem*, Oxford: Blackwell.

Smith, Michael (1995). "Internal Reasons," *Philosophy and Phenomenological Research* 55: 1, pp. 109-131.

Sobel, David (2001). "Explanation, Internalism, and Reasons for Action", *Social Philosophy and Policy* 18: 2, pp. 218-235.

Sobel, David (2003). "Reply to Robertson," *Philosophical Papers* 32: 2, pp. 185-191.

Thomas, Alan (2006). *Value and Context*, Clarendon: Oxford University Press.

Van Roojen, Mark (2000). "Motivational Internalism: a Somewhat Less Idealized Account," *The Philosophical Quarterly* 50:199, pp. 233-241.

Watson, Gary (1975). "Free Agency," *Journal of Philosophy* 72, pp. 205-220.

Wiland, Eric (2000). "Good Advice and Rational Action," *Philosophy and Phenomenological Research* 60: 3, pp. 561-569.

Wiland, Eric (2002). "On the Rationality of Desiring the Forbidden," *Analysis* 62: 4, pp. 296-99.

Wiland, Eric (2003). "Some Advice for Moral Psychologists," *Pacific Philosophical Quarterly* 84: 3, pp. 299-310.

Williams, Bernard (1981). "Internal and External Reasons," *Moral Luck*, Cambridge: Cambridge University Press.

Williams, Bernard (1995a). "Internal Reasons and the Obscurity of Blame," *Making Sense of Humanity, and other philosophical papers*. Cambridge: Cambridge University Press, pp. 35-45.

Williams, Bernard (1995b). "Replies," Altham and Harrison eds., *World, Mind, and Ethics*, Cambridge: Cambridge University Press, pp. 185-224.

[1] Notice that this is a pro-tanto reason not an overall one. It may be that I also have a reason to exercise and this may give me an overall reason to walk to the beach. And so, even if I have *a reason* to take the Number 5 bus, this does not guarantee that I *should* do it.

[2] I will be discussing the explanation condition in more detail in Section III. In particular, I will attempt to give more content to the constraint that an agent "could act on that consideration."

[3] See for example Shope (1978) and Johnson (1999).

[4] See: Railton (1986); Smith (1994) and (1995).

[5] The notions of what the rational version of oneself would advice and what the rational version of oneself would desire that one do are used more or less interchangeably here. The relevant notion is that of an advising desire – a desire about the actions of the less rational self that would be expressed in sincere advise.

[6] Watson (1975). Smith argues for the advisor model on similar grounds in Smith (1994) and so do Pettit and Smith (2006).

[7] One may think that perhaps this isn't so, that the agent may have a reason to go up to their opponent and shake their hand despite their inability to do so. I will discuss this example again in Section II and come back to this particular point in Section III.

[8] Smith, for example, claims that certain aspects of our psychology are part of the circumstances that the fully rational self should take into account when giving advice, while other aspects are not. He says, 'Suppose, for example, that you and I differ in our preferences for wine over beer. Preferring wine, as you do, you may tell me that there is a reason to go to the local wine bar after work for a drink, for they sell very good wine. But then, preferring beer, as I do, I may quite rightly reply "That may be a reason for you to go to the wine bar, but it is not a reason for me".' (Smith (1995), 122) So Smith accepts that one's particular preferences, like the preference for wine over beer, may be part of one's circumstances. Yet, as I will observe below, he denies that all of one's desires or motivations are part of one's circumstances and so that they are part of the considerations the fully rational agent has to take into account in giving advice as to what to do. However, he does not provide a way of distinguishing between those aspects of the agent's psychology that should be considered as part of the circumstances and those that should not.

[9] Although put to a different use, this distinction is drawn by Nagel in *The Possibility of Altruism*.

[10] Admittedly, unmotivated desires can depend on false beliefs. For example, I might have an unmotivated desire for a round square, which depends on my (false) belief that obtaining a round square is possible. But this is different: in the case of a motivated desire based on a false belief, part of the rational criticism of the motivated desire is that satisfying it wouldn't really satisfy the unmotivated desires. Whereas, in the case of an unmotivated desire dependent on a false belief, this can't be the rational criticism—the rational criticism has to be something else, like that the unmotivated desire in question just can't be satisfied at all. Notice that although the advisor needs to take these unmotivated desires dependent on false beliefs as part of the circumstances, they are not likely to give rise to reasons as the advisor is unlikely to advise that the agent act on those desires.

[11] Although for the most part, Williams himself seems to be implying the exemplar model, his remarks about blame (cited in the last section of the paper) for example, suggest that the advice model can capture his view and if my argument here is correct, it can do a better job at capturing his view than the exemplar model. Furthermore, in his reply to McDowell, he seems to point to the same kind of problem for the

exemplar model that is being presented here. With regards to the view that holds that what A has reason to do in certain circumstances is what the *phronimos* would have reason to do in those circumstances, Williams says: 'But, in considering what he has reason to do, one thing that A should take into account, if he is grown up and has some sense, are ways in which he relevantly fails to be a *phronimos*. Aristotle's *phronimos* (to stay within that model) was, for instance, supposed to display temperance, a moderate equilibrium of the passions which did not even require the emergency semi-virtue of self-control. But if I know that I fall short of temperance and am unreliable with respect even to some kinds of self-control, I shall have good reason not to do some things that a temperate person could properly and safely do.' (Williams (1995b), 190.)

[12] With respect to rationality, the advisor need not be thought of as fully rational (as Smith, for example, claims) but, in the relevant respects, a better and more informed deliberator. I think the term "sound deliberation" captures what is important: deliberation not based on false beliefs, takes into account all the relevant information (this includes having the imagination to think of the relevant options), and avoids mistakes in reasoning (follows the rules of logic, for example). The idea is that the advice does not come from someone so different as to be a different person but it is the equivalent of a wiser self, trying to discern what one should do in the circumstances one is actually facing. That the advisor need not be thought of as fully rational may help to respond to the criticism of the advisor model presented by Rosati (1995.) She is concerned with the fully rational person being so different from the actual person that their motives and advice might just as well be from a different person. I think this may be right if full rationality is required, however not so if what is required is just sound deliberation in the limited sense portrayed here where A+ need only deliberate soundly on what is relevant for the consideration in question. Van Roojen (2000) argues for a less than ideal advisor on the basis that it provides a way of responding to the conditional fallacy.

[13] The third condition concerns the rational capacities that A+ possesses. I have here renumbered the conditions for expository purposes and I am using "A+" rather than their "Å" for the ideal advisor.

[14] See for example, '… to summarize the main philosophical points that we think emerge from our considerations: (i) Normative reasons, internal or external, should be identified by an amendment of Williams's claim (a). A has reason to φ in certain circumstances C, we might say, only if Å desires A to φ in C, where Å has and exercises all of the capacities that ensure that her desires conform to principles of reason. … (ii) Internal reasons should be identified more narrowly by reliance on Williams's claim (b). We might describe them as those normative reasons which A has the capacity to recognize and to which A has the capacity to respond….' (Pettit and Smith, 167)

[15] This is in a sense unsurprising since Pettit and Smith are basing their account of reasons on Williams' account of internal reasons, it would indeed be surprising if they could provide an account of both internal and external reasons on such basis.

[16] I thank an anonymous referee for calling this possibility to my attention.

[17] Williams seems to be getting to a similar point in his reply to McDowell. He points out that it will not help to just try to add into the circumstances the limitations of the agent because "If the circumstances are defined partly in terms of the agent's ethical imperfection, then the *phronimos* cannot be in those circumstances." (Williams (1995b), 190.)

[18] We need not take the advisor model too far though. The idea of an advisor stands for the counterfactual situation, the possible world, where the agent is deliberating soundly about what the agent in the actual world should do. So, the idea of the advisor and the idea of the agent counterfactually deliberating soundly are closely connected. But the advisor model allows for the proper separation between the

capacities of the actual agent and the capacities of the counterfactual sound deliberator so as to avoid the problem of limiting the circumstances to those that sound deliberators can find themselves in.

[19] See, for example, Johnson (1997), Wiland (2000), and Sobel (2001). A related but somewhat different objection to the ones I will be discussing below is raised by Eric Wiland. Wiland (2002) and (2003) argues against the advisor model as a model for internalism on the basis of the possible divergence between what an advisor may recommend and what she may desire that the agent do. Wiland (2003) argues that 'sometimes a person will advise another, despite knowing that the advisee will not deliberately accept the advice, because she knows that her advice will end up influencing the advisee's actions anyway' (303). So, Wiland claims, the advisor may advise the agent to do something that is not possible for the agent to do so as to motivate them to do something else. Wiland also points out that sometimes advisors recommend something other than what they want the agent to do in order to motivate the agent to do what they want them to do. One of Wiland's examples is that of a therapist who recommends that a couple abstain from sex so as to get the couple of have more sex (2002, 297). The idea being that what the advisor desires and what the advisor recommends may generally come apart. I believe that Wiland's examples do not present a real problem for the advisor model. The relevant notion for the advisor model is that of an advising desire – a desire that would be expressed in sincere advice. Insofar as we are interested in an account of reasons for action, the relevant advice is that which is directed to the rational part of the advisee. Arguably, this is part of the notion of advice. There are many things that someone may do in response to a request for advice, some of them may constitute advice and some of them may not. The therapist could attach strings to the patients' limbs and move them like puppets. This would not constitute advice. I believe that in Wiland's example, the therapist is providing treatment to the patients and is doing so by in some sense manipulating them. The therapist is not addressing the rationality of the couple but is instead attempting to modify their behavior. If effective, the couple would end up doing what the therapist wants them to do but this doesn't mean that the therapist is providing advice in the sense that is relevant here. The relevant notion of advice then, takes into account the motivational capacities of the advisee (including possible failures of rationality) but addresses the rational part of the advisee. I want to thank an anonymous referee for bringing this issue to my attention.

[20] Sobel (2001) puts it in terms of "fragile reasons": '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.' (231)

[21] Unlike Johnson who argues for rejecting the advisor model in favor of the exemplar model because the best versions of the advisor model collapse into the exemplar model, Sobel argues that the advisor model is preferable and thus that the explanation condition and internalism need to be rejected.

[22] This idea is drawn from Williams' objection to external reasons also quoted in Section I: 'no external reason statement could by itself offer an explanation of anyone's actions' (1980, 106).

[23] This formulation is partially based on the second option considered by Sobel (2001). Sobel's second option reads as follows: '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' (222). Sobel ultimately rejects the explanation condition. Although I think that this version of the explanation condition given by Sobel is not restrictive enough – there is some possible world in which I am motivated by anything – I think that we can modify the condition as I do above to make it explicit that the circumstances are to be held constant and that said circumstances include the agent's unmotivated desires and the agent's motivational capacities to get my own version above.

[24] Sobel also argues against it on the basis of the possibility of *superfragile* reasons. Superfragile reasons are reasons that are 'so fragile that the only vantage points from which one could appreciate the way in which φing furthers something in the actual agent's motivational set are vantage points in which one lacks a reason to φ.' (231) Sobel gives as an example of a superfragile reason, that of 'a distinctive taste that, once one has tasted it, one is glad to have done 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.' (231) More generally, Sobel thinks that superfragile reasons can be found in cases where the considerations that provide the reason require first hand experience to appreciate and so once one could only appreciate the consideration in question (the taste) once one has lost the motivation to act on the consideration (see his "Reply to Robertson" for this second, more general formulation.) While I am sympathetic to the idea that we only come to appreciate some experiences by having gone through them, it is not so clear to me that there are any experiences where this is literally true. That is, it seems to me that for all experiences, one could come to be convinced that a certain experience is worth having without having had it, even if we cannot exactly know what is like to have the experience without having it. That is, even if it were true that one sometimes comes to appreciate something after experiencing it, it seems less clear to me that one could not possibly have come to appreciate it in any other way, by being told what the experience is like by someone trusted, for example. And this possibility is all that is needed to satisfy the version of the explanation condition that I am working with here. Just to be clear, if there were considerations which the person could not recognize and respond to while maintaining one's circumstances constant then my account would say that they are not reasons. However, if one could recognize and respond to them, then they may be. For Sobel's examples to present a problem, they have to be examples of considerations that the agent could not recognize and respond to yet they clearly present a reason for that agent. I am skeptical that such examples can be found, either it would be the case that the agent could recognize and respond to the considerations in question (even if they haven't experienced them) or they will not provide reasons for that agent.

[25] For a related argument defending the view that the advisor model must satisfy the internalism constraint, see Thomas (2006, 85.)

[26] See, for example, Davidson (1963)

[27] I want to thank a member of the audience at the meeting of the Latin American Society for Analytic Philosophy and an anonymous referee for pressing me on this point.

[28] Smith discusses various platitudes about advice but he does not mention explicitly the platitude I am proposing here (1995: 151).

[29] I want to thank Geoff Sayre McCord for pressing me to elaborate on some of the ideas in Section II and the philosophy department at Puget Sound where I first presented those ideas some years ago. I also want to thank the participants at the Gothenburg Conference on Moral Motivation, especially Michael Smith, for very helpful discussion of an earlier draft of this paper. In addition, I want to thank David Sobel, the audience at the Buenos Aires meeting of the Latin American Society for Analytic Philosophy, and a reading group at Puget Sound that included Johanna Wolff, Matthew Parrott, Douglas Cannon, and Justin Tiehen.