This is an accepted manuscript of an article published in *Philosophical Explorations* on 30 April, 2014. Available online: http://www.tandfonline.com/doi/full/10.1080/13869795.2014.910309

This manuscript contains typographical errors corrected in the published version.

Con-Reasons and the Causal Theory of Action

Jonathan D. Payton
Department of Philosophy, University of Toronto
jonathan.payton@mail.utoronto.ca

A con-reason is a reason which plays a role in motivating and explaining an agent's behaviour, but which the agent takes to count against the course of action taken. Most accounts of motivating reasons in the philosophy of action do not allow such things to exist. In this essay I pursue two aims. First, I argue that, whatever metaphysical story we tell about the relation between motivating reasons and action, con-reasons need to be acknowledged, as they play an explanatory role not played by pro-reasons (the reason the agent takes to count in favour of the action taken). Second, I respond to an argument recently developed by David-Hillel Ruben to the effect that a causal theory of action – still known as 'the standard story' – can't account for con-reasons. His argument attempts to show that a fundamental principle of the causal theory can't be reconciled with the role con-reasons play in a certain kind of imagined case. I first argue that a causal theorist is not, in fact, committed to the problematic principle; this argument has an added benefit, since the principle has been taken by many to show that the causal theory generates a puzzle about the possibility of weak-willed action. I then argue that a causal theorist has good reason to reject the possibility of Ruben's imagined cases. If successful, my arguments make clearer the commitments of the causal theory, and show that it can accommodate con-reasons in the way I think they ought to be accommodated.

1. The Nature of Con-Reasons, and a Case for Their Existence

At this moment there are (at least) two courses of action open to me: I could continue to work on this paper, or I could take a nap. If I deliberate about what to do, weigh my reasons for and against

each course of action, and conclude that I should keep working on this paper, then if I act on this conclusion my action is subject to what philosophers of action call a 'rationalizing explanation', or a 'reasons-explanation'. That is, we can explain why I continued to work on this paper by giving my reasons for doing so; if asked why I continued to work, I might say 'Because I want to have a new draft done by the weekend.' The reasons which figure in such explanations are called 'motivating' reasons, because they are the reasons which actually moved me to act. These stand in contrast to 'normative' reasons, which may count in favour of pursuing a certain course of action, but which need not play any role in an account of *why* I did what I did. For instance, if I've forgotten that I'll be visiting my spouse's parents for the next few days, and so won't be able to work on this paper over the weekend, then the fact that I'll be making this visit might be a *normative* reason to work on this paper now rather than putting it off, but it isn't a *motivating* reason. Since I've forgotten about the visit this fact doesn't figure in my deliberations or motivate me to act; it is a *merely* normative reason.[1] Similarly, I may actually remember that I'll be visiting my spouse's parents, and so that fact may be a reason that I *have* in the sense of having it available for use in deliberation, but it may not move me one way or the other, and hence may not be a *motivating* reason.

It's common practice in the philosophy of action to focus on the reasons in favour of a course of action, rather than those which count against it, in providing rationalizing explanations of agents' behaviour. For instance, in his now-classic exposition of the distinction between normative

[1] I have distinguished the two kinds of reason in such a way that a normative reason can still count as a motivating one. If I actually remember that I'll be visiting my spouse's parents for the next few days, then nothing I've said rules out that this fact can be a motivating reason. I have thus drawn the distinction so as to be neutral between those who think the categories are mutually exclusive (Smith 1994: chs. 4 & 5) and those who allow normative reasons to be motivating as well (Dancy 2000).

and motivating reasons, Michael Smith characterizes the latter as consisting of (a) a desire and (b) a belief that the course of action taken will satisfy the desire. Leaving aside any non-Humean scruples you might have – not to mention concerns that, if normative reasons are facts, then *motivating reason* and *normative reason* are being treated as mutually exclusive categories (see note 1) – the feature of this account that concerns us is that it leaves those reasons which figured in my deliberations, but which counted *against* the action taken, completely out of the picture. Any pair consisting of a desire and a belief that my eventual course of action prevents that desire from being satisfied won't count as a motivating reason on Smith's account. Similarly, standard statements of the role of rationalizing explanations suggest that such explanations refer only to the reasons the agent took to count in favour of doing what he eventually did.

> "When we explain an action in terms of the agent's reasons, we credit him with psychological states given which we can see how doing what he did, or attempted, would have appeared to him in some favourable light." (McDowell 1978, p.79)

> "A reason rationalizes an action only if it leads us to see something the agent saw, or thought he saw, in his action – some feature, consequence, or aspect of the action the agent wanted, desired, prized, held dear, thought dutiful, beneficial, obligatory, or agreeable." (Davidson 1963, p.3)

It seems that it would suffice, to show how an agent's behaviour could have appeared to him or her in a favourable light, to appeal only to those reasons which the agent took to count in favour of his or her action. I can show you why it seemed like a good idea to continue working on this

paper today by telling you that I wanted to have a draft done by the weekend; I need not mention any of the reasons I took to count in favour of taking a nap. To put this point in the terms recently introduced by David-Hillel Ruben, standard practice in the philosophy of action seems to be to treat motivating reasons as if they were all *pro-reasons* rather than *con-reasons*. Any reason I had not to do what I eventually did isn't usually considered to be a reason which motivated me to act.

But you might think that a complete explanation of what I did would need to refer to all of the reasons which figured in my deliberation and led to my decision. Jonathan Dancy makes the point as follows:

> "Are my defeated motivators [ie. the reasons I took to be outweighed in my deliberation] to be thought of as among the reasons which motivated me? They are not among my reasons for doing what I did. In that sense, then, they are not among the reasons that motivated me to act…But still I was influenced by them, and they do figure in my motivational economy." (Dancy 2000, p.4)

If I decide to work on this paper today, it's not as though I didn't recognize any of the reasons that I had not to work on it, and the suggestion that a complete explanation of what I did could be given without any reference to them suggests that I did my deliberating with blinders on, that I looked *only* for reasons to work, or that those were the only reasons I cared about. That suggestion, if true, would undermine the rationality of my behaviour; it's quite *ir*rational to consider only the reasons that count in favour of an action when deciding whether or not to do it. So, an explanation which both reveals the rationality of my behaviour and which is *complete* should refer to both my pro- and con-reasons.

In objection to this point, one might argue that what we do depends only on our pro-reasons, not on our con-reasons. A rationalizing explanation works only if the occurrence of my action depended on the reasons given in the explanation, and one might argue that my act of continuing to work on this paper only depended on my pro-reasons, not on my con-reasons, as follows. If we hold fixed all of my pro-reasons and suppose that I had lacked all of my con-reasons, it seems plausible that I still would have continued to work on this paper; after all, the balance of reasons would have been quite obviously in favour of doing so. Hence, my action doesn't counterfactually depend on the existence of my con-reasons. By contrast, if we hold fixed all of my con-reasons and suppose that I had lacked all of my pro-reasons, it seems plausible that I wouldn't have worked on this paper, and would have done something else instead; after all, the balance of reasons would have been quite obviously against working on the paper. Hence, my action counterfactually depends on the existence of pro-reasons. Given that my action counterfactually depends *only* on my pro-reasons, and not on my con-reasons, the latter can play no role in a rationalizing explanation of my behaviour. They may be reasons that I *had*, in the sense that they were part of my cognitive economy and not just out there in the world unbeknownst to me, but they weren't motivating reasons.

The counterfactuals given in the preceding argument concern only what would happen if certain reasons were present or absent. The argument thus misses a crucial element of reasons-explanations, namely that they function by reference, not just to the existence of certain reasons, but also to their *strength* or *weight*. When deciding what to do, I didn't just need to consider what reasons I had for and against working on this paper; that would have left me with a mere list of considerations. I also needed to consider how much each one counted for or against a particular course of action. So, the complete explanation that I have in mind appeals not just to the fact that

I had such-and-such reasons to work on this paper and such-and-such reasons not to, but to the fact that I had such-and-such reasons of such-and-such a strength to work on this paper and such-and-such reasons of such-and-such a strength not to. If I took my pro-reasons to outweigh my con-reasons, then the explanation of my working on this paper is successful; if not, then we know that something has gone wrong, and that I must be suffering from some kind of practical irrationality, such as *akrasia*. Notice that such irrationality wouldn't be revealed by a mere list of my reasons.

Having made this point, we can see that the counterfactuals which must be used to test a rationalizing explanation concern, not just what would have happened if I had lacked some of the reasons that I had, but what would have happened if my reasons had been stronger or weaker. We need to ask what I would have done if my reasons to work on this paper had been weaker or had my reasons to take a nap been stronger. There may be limits as to how much we can change the strengths of my reasons and still have a good explanation – if my reasons to work on this paper could only have been weaker if the world had been drastically different, then that possibility may not be intuitively relevant to an explanation of what I did – but if, within those limits, a change in the strength of my con-reasons results in a change in my behaviour, then those reasons play a role in explaining what I did. And it seems clear that this is the case in the example we're considering, or at least that it could be. If my reasons for working on this paper aren't *drastically* outweighed in the actual scenario, then a relatively moderate weakening of those reasons, together with a moderate strengthening of my reasons not to work on this paper, would probably result in my doing something other than working on this paper. This is why my con-reasons figure in a complete explanation of my behaviour: what I did counterfactually depends, if not on their mere existence, then at least on their strength relative to that of my pro-reasons. Only if this weren't the case would an appeal to my con-reasons be explanatorily redundant.

It should be noted, before we proceed, that the discussion so far has been pursued independently of any commitments regarding the metaphysics of action. In particular, it has been pursued independently of the causal theory of action, which is the concern of the remainder of this paper.[2] On that view, rationalizing explanation is a species of causal explanation: to say that I Φ-ed for a particular reason is to either refer directly to a cause of my action, or to provide information about mental or physical events which, while not being *reasons*, lie in its causal history. A causal theorist will naturally interpret the counterfactual dependence of my action on my con-reasons as showing either that con-reasons cause my action or that, while con-reasons are not themselves causes of my action, describing them gives information about what *did* cause it. Proponents of other metaphysical views are free to interpret the counterfactuals as needed. The point is simply that con-reasons play a role in rationalizing explanations that isn't played by pro-reasons.

2. A Problem for the Causal Theory

2.1. The Two Roots of the Problem

So, we should want to accommodate con-reasons whatever our metaphysical theory of action is. But that doesn't mean that all metaphysical theories will allow us to pull this off. In fact, in two of his recent papers, David-Hillel Ruben has argued that the causal theory of action – which from here on I'll express in simplified form, as the view that 'reasons are causes', leaving the qualification that describing my con-reasons may serve to provide information about something

[2] I should note one qualification about the causal theory of action. For our purposes I'll focus on bodily actions, ie. actions which consist of an agent's body moving in certain ways or being positioned in certain ways. For simplicity I ignore the possibility of mental action. Likewise, I ignore a causal-volitionist view on which all actions are mental acts of deciding or willing, which cause bodily movements. Presumably, Ruben's arguments apply to mental actions just as well as to bodily ones.

else which causes my action implicit – can't accommodate them.[3] According to this theory, con-reasons can only play a motivating role if they play a causal role, and Ruben argues that there is simply no plausible causal story to be told here. Thus, the causal theory – or 'causalism', as he calls it – is false, because not all motivating reasons are among the causes of our behaviour.[4]

This is an interesting and original argument against the causal theory, but how does it work? What prevents a causalist from giving con-reasons the same kind of causal role that she gives to pro-reasons? According to Ruben, the problematic feature of the causal theory – one of the two 'roots' of the problem of con-reasons – is that it "ties rationality and causality" together. (2009, p.64; 2010, p.169) Now, there's one way of reading that sentence on which it comes out as obviously true. The causalist ties rationality and causality together by insisting that rationalizing explanations either are or depend on causal explanations: whether my action was rational or not depends on what my motivating reasons were, which in turn depends on what the causes of my action were. However, Ruben has an even stronger claim in mind, which is found in the Introduction to Donald Davidson's *Essays on Actions and Events*:

> **(D)** "[I]f reasons are causes, it is natural to suppose that the strongest reasons are the strongest causes." (2001, p.xvi)

Several philosophers of action sympathetic to the causal theory have taken **(D)** on board, so the claim seems not to be an idiosyncratic feature of Davidson's view,[5] but what does it amount to?

[3] See (Ruben 2009; 2010)

[4] I assume that Ruben's arguments can be transposed so as to be more directly relevant to a causal theory on which rationalizing explanation is causal explanation, but reasons are not literally causes.

[5] See, for instance, (Mele 1983; 2003) and (Bishop 1989).

The claim seems to involve two notions of strength. The first, which we may call *rational* or *deliberative* strength, is the one we've been using until now. A reason is stronger or weaker in this sense as the agent takes it to count more or less in favour of a particular action. The second, which we may call *causal* strength, is a bit more difficult to tease out. The obvious thought is that, since for Davidson reasons motivate our actions by causing them, *causal strength* should be interpreted as *motivational strength*. Although Davidson doesn't discuss the notion of motivational strength in great detail, Alfred Mele (1998) does, and he does so in a way which seems (for the most part) congenial to a causal theory like Davidson's.[6] His discussion is framed in the language of desires (specifically, desires that agent has to the effect that that very agent performs an action of a particular sort) rather than that of motivating reasons, but he understands 'desire' broadly enough that it might actually encompass everything I refer to as a motivation reason. (25) He says that a desire to Φ "inclines the agent, in some measure, to Φ intentionally, or to try to Φ, or to try to put herself in a position to Φ. The desire's strength is the strength of this inclination." (27) Every desire of this sort (or rather, on Mele's view, every physical realizer of such a desire) contributes to the production of an attempt to satisfy it. (33)

This explication of causal/motivational strength certainly seems to fit the intentions with which proponents of the causal theory put forward the principle **(D)**. **(D)** apparently requires that the rational strength of a reason be reflected in its causal strength in a very specific way: namely that to the degree that an agent takes a reason to count in favour of Φ-ing, to that degree the reason should motivate her to Φ. Moreover, those who adopt the principle take this to entail that, if we split my reasons to Φ and my reasons not to Φ, the set with the greatest total rational strength

[6] Mele actually develops his notion of motivational strength to argue *against* certain elements of Davidson's view of practical reasoning and motivation as presented in "How is Weakness of Will Possible?" (Mele 1998: 29-31)

suffices to bring about the action that it recommends. Ruben sums up this line of thinking in a helpful passage:

> "[I]f an agent has a pro-reason to Φ and a con-reason to Ψ, and does an action token of type Φ because of that first reason, *the strength of that rationally or deliberatively stronger reason on which he acts is reflected by the fact that it is that reason that is the reason that causes an action, his Φ-ing, and the relative rational weakness of the reason on which he does not act is reflected by the other reason's failure to cause an action, in its failing to lead to his Ψ-ing*." (2009, p.64, emphasis added. The passage has been modified slightly so that the terminology is consistent with that used throughout this essay.)

This is why causalists typically think that **(D)** makes the existence of *akrasia* problematic.[7] An agent acts akratically by Φ-ing while knowing (or at least believing) that his best reasons count against Φ-ing.[8] Putting this in causal terms apparently requires that an akratic agent is caused to Φ by reasons which are not his rationally-strongest reasons, and that his rationally-strongest reasons

[7] See (Bishop 1989), (Davidson 2001), and (Mele 1983; 2003).

[8] It has been suggested to me that if we accept con-reasons, then it might be possible to act akratically by acting on a con-reason. That is, can I Φ on the grounds that *p*, when I take *p* to count *against* Φ-ing. Perhaps the obvious example of an agent Φ-ing for a reason they see as counting against Φ-ing is Milton's Satan: perhaps when he declares 'Evil, be thou my good', he doesn't see the evilness of Φ-ing as genuinely counting in favour of it, but he does nonetheless Φ on the grounds that doing so would be evil. If Satan really does preserve his judgment that Φ-ing would be evil, and that all things considered he shouldn't Φ, then he satisfies the general conception of *akrasia*. On the other hand, it is hard to make sense of his declaration on these grounds; on the contrary, he seems to think that, all things considered, he should be *evil* rather than good. Unfortunately, I can't deal with the possibility of this sort of akratic action here, at least not in the detail it deserves.

count against Φ-ing. If '*x* is the causally strongest reason (or set of reasons)' means that *x* actually suffices for the action it recommends, then **(D)** is in direct conflict with the existence of *akrasia*.

There is one difficulty with this construal of causal strength, however. As Davidson and Mele present it, the motivational strength of a desire is directed at the object of that desire. So, the motivational strength of my desire to work on this paper is directed at the act of working on this paper; it is *not* directed at the act of taking a nap. This makes it seem as though, if we follow Davidson and Mele, it is *impossible* for an action to be motivated by con-reasons. Fortunately, this problem is easily solved. As discussed in Section 1, the motivational and explanatory role of con-reasons is reflected in the fact the existence, nature and timing of my action counterfactually depends on their existence and/or their deliberative strength: if you vary the deliberative strength of my reasons to take a nap, then you will either affect the nature and timing of my act of working on this paper, or make me refrain from taking a nap altogether, and work on this paper instead. These reasons are perfectly capable of playing this explanatory role if we understand their motivational strength in terms of the degree to which they incline me to take a nap, for to that same degree they incline me *away* from working on this paper, and this latter inclination has an obvious role to play in grounding claims about what I would have done had my con-reasons varied in strength or been absent altogether.

On its own, then, **(D)** doesn't seem to prevent a causalist from acknowledging con-reasons. But **(D)** is only one of the roots of Ruben's argument. The other comes from his distinction between two kinds of case in which con-reasons play a causal role:

> **Type I Cases:** These are cases in which both pro- and con-reasons are causes of the same effect, namely the action that I perform.

Type II Cases: These are cases in which my pro-reasons are causes of my action, but my con-reasons are not.

Ruben's own example of a Type I case is a variant on Buriden's Ass, where the ass needs to choose, not between two equally attractive piles, but between two piles one of which is more attractive than the other. In this imagined case, the ass is able to see that its reasons for choosing pile A rationally outweigh its reasons for choosing pile B, and so chooses pile A. In making his case that both pro- and con-reasons are causally relevant to the ass's behaviour, Ruben appeals to counterfactuals of the sort I introduced in the last section: "If pile B, although unattractive, had not been available, or had been even less attractive, but pile A had remained the same, the ass would have chosen pile A more quickly, earlier, more determinedly, with less hesitation, or some such." (2010, p.171) In addition to Ruben's own counterfactual claims, I would add that if pile B had been more attractive, and pile A less attractive, then the ass might not have gone for pile B at all; the reasons may have balanced out, leading to a more traditional Buriden's Ass problem, or the reasons favouring pile B might have actually outweighed those favouring pile A, with the result that the ass chose the former over the latter.

Type I cases are apparently unproblematic for a causal theorist. The occurrence, timing and manner of the agent's action counterfactually depend on the existence and strength of his or her reasons, both pro and con. Thus, we can run the same sort of argument that I did in the previous section to make a case for the causal relevance of both the agent's pro- and con-reasons.

A causal theorist's troubles come from Type II cases. In Type II cases, the occurrence, timing and manner of the agent's action depend only on the existence and strength of his or her pro-reasons, and not on the strength of his or her con-reasons.

> "[A]lthough the agent has a con-reason, or con-reasons, that in some sense 'weigh' with him, the relative weighting of the pro- and the con- is clear and obvious. Deliberation is not necessary. In those sorts of cases, [Ruben] can see no reason to believe that the con-reason, if it has causal influence, must display that causal influence by affecting the character or time of the action actually taken." (2010, p.173)

The ass might be so determined to get to pile A that the attractiveness of pile B (which, after all, is not negligible) plays no causal role whatsoever: if pile B had been absent, the ass would have still chosen pile A, and would have done so at exactly the same time and in exactly the same manner as in the actual scenario. So, the ass's behaviour doesn't counterfactually depend on its con-reasons.

2.2. The Problem

We now have both roots of the problem in front of us: first, the causalist's apparent commitment to **(D)**; second, the apparent existence of Type II cases. The causalist is apparently required to provide a causal role for con-reasons to play in Type II cases, but on reflection it can seem impossible for her to do so. If my con-reasons 'weigh' with me in a Type II case, and are genuinely *motivating* reasons, playing some role in my deliberation or in a rationalizing explanation of what I do, then a causalist is committed to saying that they have some effect(s). Since by hypothesis my con-reasons don't cause my *action* in a Type II case, the causalist must find something *else* for them to cause. For instance, if my reasons to take a nap don't help cause me to work on this paper, they might cause thoughts of regret later on, when I think about how I wasn't able to satisfy all of

the reasons that I had, both pro and con. The problem here is that if we have "two distinct causal chains, each of which leads to a different result," then deliberative strength is no longer reflected in causal strength; my pro- and con-reasons both have deliberative strength with respect to what I do, but only the pro-reasons have causal strength with respect to it. (2010, p.175) If deliberative and causal strength come apart in this way, then **(D)** is violated.[9] The causal theory of action thus seems to be incapable of reconciling **(D)** with the possibility of Type II cases, so long as con-reasons are allowed into the story of an agent's motivating reasons in the way I argued for in Section 1.

In the following section, I propose to undermine this argument by digging up its roots. I will argue, first, that a causal theorist is not committed, *qua* causalist, to **(D)**; she can thus safely reject it and avoid any trouble from con-reasons. However, one still might find **(D)** to be a plausible principle to include in one's theory of action, and so I go on to argue that a causalist not only fails to be committed to the possibility of Type II cases, but can actually present a principled case *against* their possibility.

3. Solving the Problem

3.1. Rejecting (D)

Recall that **(D)** requires that a motivating reason's deliberative strength be matched by its causal strength, and furthermore that a motivating reason make a causal contribution to the occurrence of the act that it recommends, regardless of whether the agent actually performs that action or not. At

[9] Ruben actually raises a second difficulty for the causalist in his essays: not only must a causalist find something for my con-reasons to cause, she must say that they cause the action they recommend, which is precisely what is ruled out in Type II cases. I won't deal with this difficulty here; since it is also motivated by **(D)** and the possibility of Type II cases, it requires no separate treatment.

least, the principle must be read in this way in order to create a problem with respect to con-reasons: **(D)** is supposed to be violated if we allow a con-reason to play a causal role with respect to other events without playing a causal role with respect to the action that it motivates. That a reason has its causal strength with respect to something other than the act which it favours is apparently something a causalist can't allow.

Understood in this way, **(D)** looks like a very strong claim; so strong, in fact, that a causalist could be forgiven for wondering why she should be thought to be committed to it. First of all, as noted, many philosophers of action take **(D)** to pose a problem for the existence of *akrasia*, or weakness of will. To the degree that we find it plausible that human beings sometimes act in ways that seem unreasonable even to themselves, we should demand a justification for **(D)** with at least that degree of plausibility. Second of all, **(D)** so understood seems to be incompatible with the existence of Type I cases, in which pro- and con-reasons both work together to cause the action that I perform. To the degree that we find it plausible that Type I cases ought to pose no problem for the causal theory of action, we should demand a justification for **(D)** with at least that degree of plausibility.

The only direct argument we've seen for **(D)** is that it is supposed to be a natural thought for a causal theorist that, if reasons are causes, then the strongest reasons are the strongest causes. But that thought looks no more natural than this one:

(B) If bachelors are males, then the tidiest bachelors are the tidiest males.

The discovery that there are some married men who are tidier than the tidiest of bachelors would do nothing to undermine the claim that bachelors are males, because **(B)** is so obviously

wrongheaded. In general, that *X*s are *Y*s doesn't entail that the *F*-est *X*s are the *F*-est *Y*s. If **(D)** is of that form, then it's no better off than **(B)**. In fact, however, **(D)** might actually be *worse* than **(B)**. **(B)** attributes the same property, tidiness, to bachelors and males, so we can legitimately represent it as an instance of 'If *X*s are *Y*s, then the *F*-est *Y*s are the *F*-est *Y*s.' **(D)**, by contrast, describes a relation between two distinct properties: deliberative strength and causal strength. Thus, **(D)** looks to have the form 'If *X*s are *Y*s, then the *F*-est *X*s are the *G*-est *Y*s,' which is a plainly invalid inference.

One might think that there must be more going for **(D)**, since it's supposed to be a quick statement of the problem Davidson is addressing in "How is Weakness of the Will Possible?" Perhaps, the thought goes, we can find a more fully-developed and plausible argument for the principle in that essay.

If this suggestion were correct, then "How is Weakness of the Will Possible?" would be an attempt to reconcile a commitment of the causal theory of action with the existence of *akrasia*. It's quite common to read the essay in that way, not least because Davidson himself presents the essay that way in his Introduction to *Essays on Actions and Events*. But this isn't the correct way to read it. In the original essay, Davidson is concerned to reconcile the existence of weakness of will with the truth of two principles:

> (P1) "If an agent wants to do *x* more than he wants to do *y* and he believes himself free to do either *x* or *y*, then he will intentionally do *x* if he does either *x* or *y* intentionally." (1969, 23)
>
> (P2) "If an agent judges that it would be better to do *x* than to do *y*, then he wants to do *x* more than he wants to do *y*." (ibid.)

Since an agent exhibits weakness of will precisely by doing *x* when she (a) judges it better to do *y* and (b) believes herself free to do either *x* or *y*, (P1) and (P2) are apparently incompatible with the existence of weak actions. Admittedly, Davidson says that these principles "derive their force from a very persuasive view of the nature of intentional action and practical reasoning," (1969, p.31) and later interpreters have been happy to read his later claim that he was concerned primarily to defend a causal theory of action back into this passage; the result is that Davidson is now read as claiming that (P1) and (P2) derive their force from a *causal* theory of action and practical reasoning.[10] This is puzzling, since neither principle contains any explicitly causal claims. They seem to be principles one can adopt or reject independently of one's commitments on the metaphysics of action. That is, you can accept or deny that agents always want most what they judge to be best, and act so as to achieve what they want most, without needing to worry about the metaphysical question of what acting so as to achieve something actually consists in, ie., what separates intentional actions from things that 'just happen'. How, then, could these principles get their support from a causal theory of action?

One might try to derive (P2), at least, as follows. Causalism requires that motivating states cause one's behaviour, and one might think that such a claim stands in need of some phenomenological support. On this way of thinking, if causalism requires that motivating reasons *cause* one's behaviour, then there must be something about their phenomenology which indicates that they play this causal role. E. J. Lowe, for instance, seems to take this view when he writes,

[10] See, for instance, (Mele 1983, p.345) and (Mele 2003, pp.76-77).

> "[w]hen we take ourselves to be acting rationally it never *seems* to us that we are being caused to act in the ways we do by our beliefs and desires, and once we suppose that, on a given occasion, we were in fact caused to act in a certain way by our beliefs and desires, we find ourselves obliged to withdraw any claim to have acted rationally on that occasion." (Lowe 2010, p.189)

(P2) seems like it could be a response to this worry, since one might think that, while judgments about what is best don't have the kind of phenomenology which indicates a causal role, desire do. (P2) satisfies the phenomenological demand by ensuring that judgments as to what is best, or as to what one ought to do, will have an 'oomphy' phenomenology. Unfortunately, Davidson rejects this phenomenological argument. He follows Anscombe in thinking that "the primitive sign of wanting is *trying to get,*" (Anscombe 1963, p.68; Davidson 1969, p.22) and I can pursue a goal even if I lack a desire with an 'oomphy' phenomenology.[11] So this style of argument isn't to be found in Davidson's essay.

The obvious place to look next for Davidson's argument that causalism motivates (P1) and (P2) is the passage quoted earlier, which many now read as saying that (P1) and (P2) get their force from a plausible *causal* theory of practical reason and action. That passage continues:

> "When a person acts with an intention, the following seems to be a true, if rough and incomplete, description of what goes on: he sets a positive value on some state of affairs (an end, or the performance by himself of an action satisfying certain conditions); he believes (or knows or perceives) that an action, of a kind open to him

[11] For discussion, see (Schueler 1995), pp.28-34.

> to perform, will promote or produce or realize the valued state of affairs; and so he acts (that is, he acts *because* of his value or desire and his belief)." (1969, p.31)

This is Davidson's reading of Aristotle's practical syllogism: the agent sets a value on *x*, believes that Φ-ing promotes *x*, and straightaway Φs (this is no doubt a simplistic reading of Aristotle, but that's not to the present point). The idea is that when we act intentionally, we act as we do because we see something good or valuable about acting in that way; this isn't a causal theory of action, but rather what Sergio Tenenbaum calls 'the scholastic view', that we desire only what we conceive to be good, and always act in the light of some apparent good.[12] This reading of Davidson gets even further support from an earlier passage where he first introduces the conceptual difficulty posed by *akrasia*: "There seem to be incontinent actions…The difficulty is that their existence challenges another doctrine that has an air of self-evidence: that, in so far as a person acts intentionally he acts, as Aquinas puts it, in the light of some imagined good." (1969, p.22) Now, Davidson notes that this view of practical reasoning, as stated, doesn't entail that there are no weak or incontinent actions; as long as the agent sees *some* good in what she does, she will satisfy this account of practical reason, even if she sees some other course of action as better than the one she takes. However, Davidson thinks that whatever reasons we might have for thinking that people always act in light of some imagined good, they ought to count in favour of the view that people always act as they judge it would be best to act. (1969, p.22) After all, if people can intentionally

[12] Tenenbaum (2007, p.1, n.2) accurately cites "How is Weakness of the Will Possible?" as evidence of how popular this thesis was at the time of its writing. It's important, when reading Davidson's essay, to remember just how widespread the view was in comparison to the causal theory of action, which was only beginning to see a revival in the wake of "Actions, Reasons and Causes."

act against their own judgment as to what course of action is best, why suppose that they can't act against their own judgment as to what courses of action are *good?*

I conclude that one is not committed to **(D)** simply by virtue of adopting a causal theory of action. This is good news, not only with respect to the problem of con-reasons, but with respect to the problem of how weakness of will is possible. If **(D)** is a principle motivated by a theory of practical reasoning rather than by a metaphysical account of the relation between motivating reasons and actions, then despite what many causalists have thought, commitment to the causal theory of action does not, on its own, generate any puzzle about weakness of will.

The strength of the foregoing argument shouldn't be overstated. I haven't shown that **(D)** is false, or even that there is no reason whatsoever to adopt it. A causalist might find Davidson's account of practical reasoning congenial, and so build **(D)** into her metaphysics of action as a consequence of that account. The point is simply that a causalist is free to solve the problem posed by con-reasons by rejecting **(D)**. I now turn to the stronger of my two arguments: a causalist can effectively argue against the possibility of Type II cases.

3.2. Rejecting Type II Cases

We should note, first of all, the dialectical difficulty that Type II cases pose. The causal theory of action is a theory of *motivating* reasons, rather than a theory of (merely) *normative* reasons or of reasons that I *have* but which don't motivate me. The central claim behind the causal theory is that the reasons which motivate us to act cause our actions, where the reasons which motivate us are those which figure in complete rationalizing explanations of our behaviour. If the causal theorist is committed to telling some causal story about con-reasons in a Type II case, then, this can only

be because those reasons help explain what the agent did, and hence play a motivating role with respect to the agent's behaviour in that case.

However, the causalist isn't committed to attributing just *any* causal role to my motivating reasons; to say that my reasons to work on this paper motivated me to work on this paper but played no causal role with respect that action, only helping to cause something *else*, would clearly not be to expound a causal theory of action. The causal theory of action says that my motivating reasons cause the very same things that they motivate, namely my actions. If properly understood, this commitment of the causal theory of action ought to raise a worry about Type II cases. If Type II cases are going to be of any concern to a causalist, then they must involve motivating reasons; thus, they must be cases in which my con-reasons motivate my action without causing it. But the crux of the causal theory of action is just that motivating reasons motivate our actions *by causing them*. Forget any concerns about con-reasons in particular; if Type II cases exist, so that a reason can motivate an action without causing it, then the causal theory of action is automatically false. If the existence of Type II cases entails the falsity of causalism, then the assumption of their existence runs the risk of being question-begging.

One might miss this point if one thinks of causalism as Ruben does, namely as the view that motivating reasons "must have some effects, whatever they might be." (2010, p.170) His idea seems to be that a causalist is committed to the existential claim that my con-reasons cause *something* in the scenario in which they motivate my action, and to the claim that what they cause is my action, but that the latter isn't the causalist's ground for the former. If the core commitment of the causal theory were simply that motivating reasons have some effects, then the core commitment of the causal theory wouldn't be refuted by the existence of Type II cases. What would be refuted is the supposedly natural corollary **(D)**, and worries about question-begging seem

less pressing if one's thought-experiment only violates a natural corollary of the theory one is attacking, rather than the core of the theory itself. But this is not the right way to understand a causalist's commitments. She commits herself primarily to the claim that my motivating reasons cause my action, and then infers that they cause *something* by existential generalization. If she has no reason to believe that a reason motivates my action, then *qua* causalist she has no reason to attribute to it *any causal role whatsoever*. Thus, the worry about question-begging stands. The claim that a reason motivates an action without causing it is just one premise in Ruben's argument, but it alone entails that the causal theory of action is false. As with the argument of Section 3.1, however, this point shouldn't be overstated. For all I've said so far, Type II cases might be possible. A causalist will be in a much stronger position if she can reject the existence of Type II cases on principled grounds, rather than simply by appeal to the fact that her theory entails that they don't exist.

What, then, might we say to motivate the claim that Type II cases exist? Earlier, I simply said that there seem to be possible cases in which an agent would have behaved in exactly the same manner as they actually did, had their con-reasons been absent. Ruben fleshes out this claim more fully:

> "[a Type II] case might be one in which the agent has, and acknowledges that he has, a weak moral reason that he does consider in his deliberations, but the weak moral reason has in the end no actual effect on his eventual choice or behaviour. Or a case in which the ass considers both hay piles in its deliberation but is so determined to get to hay pile A that he would make exactly the same choice regardless of what acknowledges to be the lesser but negligible attractiveness of hay pile B, and so the

> ass would make the identical choice – a choice qualitatively identical in character, timing, and so on – had hay pile B not been available at all." (2010, p.172)

There are two claims here. *First*, the agent's con-reasons 'weigh' with him; this is supposed to be evidence that those reasons played a motivating role, and hence will figure in a complete reasons-explanation. *Second*, the occurrence of the agent's action doesn't counterfactually depend on the existence of those reasons; had the agent lacked those con-reasons altogether, he would have performed an action identical in all respects to the one he actually performed. This is supposed to be evidence that those reasons didn't play a causal role.

Beginning with the first point, in what sense are the con-reasons supposed to 'weigh' with the agent? It can't be simply that the agent acknowledges that he has those reasons, since this is consistent with their being *merely* reasons that he has, and not genuinely motivating reasons. The agent's con-reasons need to figure in his deliberations, or, if there was no deliberation, they must nonetheless help to explain his action. Regarding the first possibility, if the ass's con-reasons actually figured in his deliberations then you would think that at least the timing of his action would have been different than it actually was if those con-reasons had been absent. Deliberation takes time, and the more things you have to consider, the longer it takes. If the ass needs to deliberate about his con-reasons in the actual scenario, but not in the counterfactual one, then you would expect the counterfactual deliberation to take less time than the actual one. If that's so, then we shouldn't think that the ass would have made a qualitatively identical choice if his con-reasons had been absent.[13] Regarding the second possibility, suppose that the ass can just *see* that pile A is

[13] Ruben actually considers this objection, but his response seems to be, not to answer it directly, but to shift the burden of his argument to cases in which the agent doesn't undergo any conscious, explicit deliberation at all (2010, p.174).

much better than pile B, and so chooses the former without having to think about it. Then perhaps there wouldn't even be a time-difference between his actual action and his action in the counterfactual circumstance where pile B is absent.[14] But if there is no difference between what the ass does in the actual and counterfactual scenarios, then in what sense does the presence of his con-reasons contribute to an explanation of his behaviour? We should insist that the presence of con-reasons only helps explain an agent's behaviour only if the absence of those con-reasons would bring along with it a change in the agent's behaviour. If the ass does *exactly* what he would have done had his con-reasons been absent, then their presence can't help to explain what he did.

So far we've only been concerned with a simple counterfactual claim about what would have happened if the ass's con-reasons hadn't been present at all. But there is a way to defend the claim that the ass' con-reasons are genuinely motivating, and hence explanatory, even if his behaviour is in all respects the same as it would have been without them. As we saw in Section 1, reasons both pro and con can explain our behaviour even if our behaviour doesn't change under the supposition of their absence; all that's required for reasons to play an explanatory role is for our behaviour to change under the supposition of a change in their strength. For instance, although the ass might have done exactly the same thing had pile B not existed, it seems plausible that he would have behaved differently if the relative attractiveness of piles A and B had differed. If pile A had still outweighed pile B, though not by nearly as wide a margin, then presumably the ass couldn't have just *seen* which pile was better, but would have needed to deliberate, and hence would have taken longer. If pile A had been less attractive, and pile B sufficiently *more* attractive

[14] Even this claim can be doubted. Even if both pile A and pile B are within the ass's visual field, you might think that in order to see that one is much better than the other, the ass must shift his attention from one to the other. If shifts of attention like this take time, then the ass will go for pile A more quickly in the counterfactual scenario where pile B is absent and no shifts of attention are necessary.

so as to outweigh pile A, then unless the ass suffered weakness of will he would have chosen pile B instead. These sorts of counterfactuals constitute strong evidence that the ass's con-reasons played a motivating role.

While these counterfactuals help motivate the claim that the ass's con-reasons are motivationally relevant, they also motivate the claim that those reasons are *causally* relevant in precisely the way that causalism says they must be. The right way to test whether a reason is causally relevant to an agent's behaviour in that way is to see what she does under variations of its strength, and not just to see what she does when that reason is absent. If the behaviour varies with the strength of that reason, then the strength of the reason seems to play a causal role with respect to the behaviour.

The causalist therefore has a principled reason for rejecting the existence of Type II cases. The very same counterfactuals which motivate the thought that my con-reasons play a motivational role in such cases also motivate the thought that they play precisely the causal role that a causalist will want to claim that they do.

4. Summing Up

I've argued that con-reasons deserve to have a place in rationalizing explanations of human action, and hence that a causal theory of action must somehow accommodate them. Moreover, I've argued that Ruben has given us no reason to think that con-reasons will pose any special problem for a causal theory, since his arguments hinge on commitments that a causalist can easily reject. The claim that there could be Type II cases, as Ruben describes them, begs the question against the causalist, and Ruben's arguments for their possibility are unconvincing. Moreover, the principle **(D)** is not something that anyone is committed to simply by adopting a causal theory of action. If

a causalist can reject both the existence of Type II cases and the commitment to **(D)**, then she can reject both of the roots of the problem apparently posed by con-reasons. A causal theory of action is therefore easily compatible with the story with which I began, according to which the reasons which count against the action I perform nonetheless helped to motivate me to do it.

Sources

Anscombe, G. E. M. *Intention, 2nd Edition*. Cambridge: Harvard University Press.

Davidson, Donald. (1963) "Actions, Reasons and Causes," in *Essays on Actions and Events*. New York: Oxford University Press. 2001. pp.3-20.

———————. (1969) "How is Weakness of the Will Possible?" in *Essays on Actions and Events*. New York: Oxford University Press. 2001. pp.21-42.

———————. (2001) "Introduction," in *Essays on Actions and Events*. New York: Oxford University Press. 2001. pp.xv-xxi.

Dancy, Jonathan. (2000) *Practical Reality.* New York: Oxford University Press.

Lowe, E. J. (2010) *Personal Agency: The Metaphysics of Mind and Action.* New York: Oxford University Press.

McDowell, John. (1978) "Are Moral Requirements Hypothetical Imperatives?" in *Mind, Value, and Reality*. Cambridge, Harvard University Press. 1998. pp.77-94.

Mele, Alfred R. (1983) "'Akrasia,' Reasons, and Causes." *Philosophical Studies* 44:3, pp.345-368.

———————. (1998) "Motivational Strength." *Noûs* 32:1, pp.23-36.

———————. (2003) "Philosophy of Action," in *Donald Davidson*, Ed. Kirk Ludwig. New York: Cambridge University Press. 2003. pp.64-84.

Ruben, David-Hillel. (2009) "Con-Reasons as Causes," in *New Essays on the Explanation of Action,* Ed. Constantine Sandis. New York: Palgrave MacMillan. pp.62-74.

———————. (2010) "Causal and Deliberative Strength of Reasons for Action: The Case of Con-Reasons," in *Causing Human Actions: New Perspectives on the Causal Theory of Action*. Eds. Jesús H. Aguilar and Andrei E. Buckareff. Cambridge: The MIT Press. pp.167-182

Schueler, G. F. (1995) *Desire: Its Role in Practical Reasoning and the Explanation of Action.* Cambridge: The MIT Press.

Sehon, Scott. (2005) *Telelogical Realism: Mind, Agency and Explanation*. Cambridge: The MIT Press.

Smith, Michael. (1987) "The Humean Theory of Motivation," *Mind* 96, pp.36-61.

———————. (1994) *The Moral Problem.* Malden: Blackwell Publishing, Ltd.

Tenenbaum, Sergio. (2007) *Appearances of the Good.* New York: Cambridge University Press.

Acknowledgements

Special thanks to David-Hillel Ruben for helpful comments on earlier drafts, discussion, and encouragement. Thanks to Sergio Tenenbaum and an anonymous reviewer for *Philosophical Explorations* for helpful comments on earlier drafts. Research for this paper was funded by the Social Sciences and Humanities Research Council, Canada.