Causing Human Actions

New Perspectives on the Causal Theory of Action

edited by Jesús H. Aguilar and Andrei A. Buckareff

A Bradford Book
The MIT Press
Cambridge, Massachusetts
London, England

This book was set in Stone Sans and Stone Serif by Toppan Best-set Premidia Limited. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Causing human actions : new perspectives on the causal theory of action / edited by Jesús H. Aguilar and Andrei A. Buckareff.
p. cm.
"A Bradford Book."
Includes bibliographical references and index.
ISBN 978-0-262-01456-4 (hardcover : alk. paper)—ISBN 978-0-262-51476-7 (pbk. : alk. paper)
1. Act (Philosophy). 2. Action theory. 3. Intentionality (Philosophy).
I. Aguilar, Jesús H. (Jesús Humberto), 1962–. II. Buckareff, Andrei A., 1971–.
B105.A35C38 2010
128′.4—dc22

2010003187

10 9 8 7 6 5 4 3 2 1

Contents

Preface

This volume brings together essays by some of the leading figures working in action theory today. What unifies all of the essays is that they either directly engage in debates over some aspect of the causal theory of action (CTA) or they indirectly engage with the CTA by focusing on issues that have significant consequences for the shape of a working CTA or the tenability of any version of the CTA. Some of the authors defend this theory, while others criticize it. What they all agree on is that the CTA occupies a central place in the philosophy of action and philosophy of mind as the "standard story of action." Two of the essays in this volume have appeared elsewhere recently. (Chapters 8 and 9 by Carolina Sartorio and Randolph Clarke, respectively, previously appeared in *Noûs*. They appear with the permission of Wiley-Blackwell, and have been lightly edited for consistency.) The remaining essays appear in this volume for the first time.

Editing this volume, though not an easy task, has been a labor of love for us. We are convinced that foundational issues in the philosophy of action, such as the issues explored in this volume, deserve greater attention. It is our hope that the publication of this collection of essays will serve to elevate the prominence of the debates the essays range over in future research on human action and agency. This volume, then, is in part an effort to promote exploration of foundational issues in action theory and especially to encourage further work on the CTA by defenders and critics alike.

Work on this volume would have been more difficult if not impossible without the support of a number of people and institutions. First, Philip Laughlin, Marc Lowenthal, and Thomas Stone from MIT Press deserve a special debt of gratitude for supporting this project. Also from MIT Press, we would like to thank Judy Feldmann for her fantastic editorial work. Thanks to Wiley-Blackwell for giving us the permission to publish the essays by Carolina Sartorio and Randolph Clarke which originally appeared

12 Causal and Deliberative Strength of Reasons for Action: The Case of Con-Reasons

David-Hillel Ruben

An agent's having of a reason for an action (hereafter, simply "a reason") is often said to be among the causes or causal conditions of the action for which it is a reason (in this wide sense, "action" includes many cases of inaction).[1] Hereafter, this view is referred to as *causalism*, or (1). Causalism as here understood is a thesis about causation, not about causal explanation.

Causation and causal explanation might come apart in both directions, as it were. There might be causal explanations that do not cite any or only causes, under any description. In any case, it is clear that there can be true causal statements such that the cause is not explanatory of its effect because of the description of the cause in that statement, but there might even be causes such that, under no description of the cause, does the cause explain its effect. In this essay, I remain agnostic about all such possibilities.

An agent can, of course, have a reason for an action without its being a cause of that action, but in the case in which the agent performs the action because of that reason, the reason is said, according to causalism, to cause the action. If we conflate just for the moment questions about causation and causal explanation, Donald Davidson's well-known argument attempts to demonstrate just this: "If . . . causal explanations are 'wholly irrelevant to the understanding we seek' of human actions then we are without an analysis of the 'because' in 'He did it because . . .,' where we go on to name a reason" (Davidson 2001a, 86–87). The reasons Davidson focuses on, in this argument and elsewhere, are reasons which function as "pro-reasons," in two closely connected senses: (a) these are reasons that are relevant to, or bear upon, the action the agent does in fact take; (b) they are reasons that favor that action. Davidson's reasons consist of a belief and a pro-attitude toward the kind of action the agent does. Other views might identify intentions or some other mental items as these

reasons, but precisely which mental items count, on a causalist view, as reasons will not concern me here (von Wright 1978, 46–62). I also take no view here about the nature of such mental items; they might or might not be token-identical to brain or other physiological states.

The reasons that (1) is about are sometimes called "explanatory" reasons in contradistinction to "justifying" reasons, or "motivating" reasons in contradistinction to "normative" reasons, or "internal" reasons in contradistinction to "external" reasons. These are three different distinctions, related to one another in complicated ways. The reasons required by (1) are surely at least motivating: these are the reasons that have actual psychological "purchase" on the agent, and are not *just* there and merely in principle available to the agent in some wholly objective sense. A well-known view, which is also adopted by the position I am considering, asserts that the only reasons that could play the causal role required by the idea of motivation, and hence explain why the agent did what he did, are the agent's internal psychological states. So the reasons required by (1), as I construe it, are internal, explanatory, and motivating, although these reasons, just insofar as they are *reasons* for action, are being asked to play something of a weakly normative role as well.

The literature in action theory has tended to overlook the fact that reasons function in another way too.[2] (a) One can have reasons for an action that one does not take; (b) one can have reasons that disfavor an action taken (a "con-attitude," to parallel Davidson's "pro-attitude"). Sometimes a person has conflicting reasons for acting, one set of which is a set on which he does *not* act, and both sets of conflicting reasons can be rationally or deliberatively relevant to, or bear upon, the same choice situation and rationally or deliberatively relevant to the same action finally chosen. The con-reason is relevant to the action taken, in the sense that that it is the action it disfavors. Both the pro- and the con-reason, as I shall continue to say, are "rationally or deliberatively relevant" to, or "bear upon," the same eventual choice made or action taken. The pro-reason is relevant to the action taken because it favors it; the con-reason is relevant to the action taken because it disfavors it. Each set of reasons justifies or supports a different proposed action on the agent's part, and the agent is not able to perform both actions, because it is impossible to act in both ways at the same time.

The reasons might strongly conflict, in the way in which a reason to do some token act of type *A* (or, as I sometimes elliptically say, a reason to *A*) and a reason not to do any token act of type *A* conflict; or the reasons might weakly conflict, in the way in which a reason to do some token act of type *A* and a reason to do some act of type *B* (or, as I sometimes elliptically say, a reason to *B*) conflict on any occasion on which one cannot as a matter of fact do both. On an occasion on which one cannot do both, a reason to do an act of type *B* must also be a reason not to do an act of type *A*, but only modulo the additional information that one cannot do both acts in the circumstances. In the case of strong conflict, no additional information is similarly required.

Suppose that *X* has a reason to do a token action of type *A* and a reason to do a token action of type *B*, where the two reasons weakly conflict on the particular occasion. Suppose further that *X* chooses to perform a token action of type *A*. The first reason, which favored doing *A*, was rationally or deliberatively weightier (in the circumstances, of course) than the second, which favored doing *B*; the first counted for more, as far as *X* was concerned. As we say, all things considered, *X* chose to perform a token action of type *A*. (I write as if choice precedes every action, but my argument would be unaltered if choice were not ubiquitous in this way.) The single final choice made or action taken is made or taken *because* of the pro-reason and *in spite of* the con-reason. Indeed, that is what "all things considered" must mean: because of the one set of reasons and in spite of the other.[3] As Dancy (2004, 4) says about these con-reasons: "But still I was influenced by them [the con-reasons] and they do figure in my motivational economy."

We use the language of reasons, weights, and strength, in describing our deliberations. Such language is metaphorical, but, metaphorical or not, it certainly seems irreplaceable. We can order reasons for action by their comparative strengths: one reason for action can be stronger than another, weaker than a third. Whether or not there are nonrelational truths about the deliberative strength of reasons, on which the relational truths about them supervene, it is relational information about reasons that is crucial for understanding the deliberative story.[4] (Note that if there is such supervenience, the same relational story can supervene on different nonrelational stories. What supervenience rules out is the possibility of different relational stories supervening on the same nonrelational or intrinsic story.) If we have both pro- and con-reasons, we want to know which reason wins the deliberative contest. It is only the relational information about the rational strength of reasons for action that interests us in cases of choices between actions.

The notion that reasons are causes, causalism, ties rationality and causality, in some way yet to be discerned. This would allow us to distinguish two kinds of strengths for reasons: rational or deliberative strength (as

above) and causal strength. Deliberative or rational strength is a measure of the extent to which a reason for action supports that action. Rational strength is an epistemologically normative idea, sometimes called "weak normativity." There is an analogy between this idea and the idea of the degree of support given by evidence to a conclusion in an inductive argument. A reason's comparative or relational rational strength compares it to the support competing reasons give to the alternative actions they support.

The problem of weakness of the will arises for the causalist as an issue because it challenges the natural tie, on the causalist program, between rational and causal strength of reasons for action. For the causalist, weakness of the will appears to force a wedge between the normative/rational and the causal/motivational ideas of a reason. Davidson himself says that

(2) "if reasons are causes, it is natural to suppose that the strongest reasons are the strongest causes" (Davidson 2001a, xvi).

Any causal view is going to have to address the question: how are rational and causal strength related? It is this question that gives rise to the problem that the causalist faces with weakness of the will. In a case of weakness of will, the rationally stronger reason is not the reason that causes or motivates the agent to act, if reasons do in fact cause actions. The agent acts on a rationally weaker reason, but one that is causally strong enough, where the rationally stronger reason is not causally strong enough. "Causally strong enough or not so" just means: the rationally weaker reason causes the action it supports, and the rationally stronger reason does not cause the action it supports.

I do not know whether the causalist can really successfully deal with the phenomenon of weakness of the will. But I want, in this essay, to address a different issue, unconnected to weakness of the will. Let's start by trying to trace out the causal chains that lead from the con-reasons, for on the causalist view con-reasons must have some effects, whatever they might be. The thought that there are events, con-reasons, which have no effects at all, is not one likely to appeal to the causalist. To be part of the causal order is surely to have both causes and effects.[5]

There are two importantly different cases that I want to distinguish (from the causalist point of view). In cases of Type I, the pro-reason and the con-reason jointly cause the same effect; in cases of Type II, they have separate and causally independent effects.

Type I: These cases are the ones that will most naturally spring to mind on a causalist view, but I believe that such cases are more limited than one might unreflectively assume. Let's start with an analogy from natural science. In circumstances *c*, a ball falls toward the Earth because of gravitational attraction and despite the presence of an upward wind. One could say: the cause of the ball falling to Earth is jointly both the gravitational attraction and the relative weakness of the wind's counteracting force. The two causal factors jointly contribute to the same result, the ball's final trajectory. The relative weakness of the wind is a causal factor in the ball's actual fall to Earth. If there had been no, or even less, weak counteracting force of the wind, the ball would have fallen to Earth faster, somehow differently, or some such. Perhaps the difference is temporal: if there had been no or less a counteracting force of the wind, the ball would have fallen to Earth sooner, earlier. Both factors were parts of the full cause; had either been absent, the result would have been (or probably would have been) different, or at least differently placed temporally.

A Type I example in the case of action would also not be hard to find, and, as I said above, it is these cases that spring most readily to mind: suppose Buridan's ass is drawn to hay pile A because it is more attractive, hay pile *B* is less attractive. (In a real Buridan's ass case, the piles are equally attractive, but not so in the case I am now imagining.) No starving ass here. Both the attractiveness of *A* and the relative unattractiveness of *B* result in the ass's choosing hay pile *A*. We are all such asses much of the time; many cases are of this type. Again, the same causal factors jointly contribute to the same result, the ass's final choice. The relative unattractiveness of *B* is a causal factor in the ass's actual choice of *A* over *B*. If pile *B*, although unattractive, had not even been available, or had been even less attractive, but pile *A* had remained the same, the ass would have chosen hay pile *A* more quickly, earlier, more determinedly, with less hesitation, or some such. Since both factors, the degree of attractiveness that pile *A* had and the degree of unattractiveness that pile *B* had, were both parts of the full cause of the ass's token choice of *A*, had either been absent, the result would have been (or probably would have been) different, either in character or in time. For example, absent pile *B*, the ass would have still chosen pile *A*, but the token choosing would have been different in some significant way from the actual choosing, as a result of the difference in its cause.

In examples of cases of Type I, it is assumed that both causal factors jointly influence the same outcome, so the outcome would have been different, or probably would have been different, if either of the causal factors had been different. Both causal factors matter to the same outcome. Although something like this might be true, and indeed no doubt is true

in many cases, it need not be. That is, on the causalist program, there could be other cases (Type II) in which the presence of the con-reason has some effect, but has no effect of any sort on the actual action taken, but instead has some effect on something else. I do not know how to prove that there must be such cases for the causalist, but it seems to me intuitively clear that there could be.

What would the causalist have to hold, in order to deny this claim that I have just made? Since in cases of joint causation of a single effect by multiple part causes, it follows that the effect would have been different, or probably would have been different, had any one of the part causes been different (or altogether absent), the causalist who wishes to deny the possibility of cases of Type II would have to say that:

(3) In every case of action for which the agent has both pro- and con-reasons that figure into his deliberations, had the agent not had such a con-reason, the action he took would have been different or altered, or probably would have been different or altered in some way, or occurred at a different time.[6]

I just don't think that (3) could be true. I can envisage many cases in which, were the con-reason absent, the action taken could be qualitatively the same (in all nontrivial respects) as the action that was actually taken. We can say of such cases: "the agent had, and acknowledged that he had, a less weighty reason not to do something, which figured into his deliberative activity, but that less weighty reason did not at all causally influence his eventual choice to do what he did, in any way." Perhaps such a 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 behavior. 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 he acknowledges to be the lesser but not 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. That is, had the agent not have had the con-reason, his actual choice would have been (or probably would have been) qualitatively the same in all relevant respects. Cases of Type II already presume that reasons and causation even on the causalist program can part company to this extent: a con-reason must cause something, but the con-reason might not be a part-cause of the same effect that the pro-reason causes or part-causes.

The case of temporal location might merit special consideration. Perhaps the action chosen in the face of both pro- and con-reasons could be intrinsically qualitatively the same as the action the agent would have chosen had he only had the pro-reason. But at least won't the time of the actions be different? If the agent had had no con-reason, he could not have deliberated and weighed up pro- and con-reason, and however short a time that deliberation might have taken, had he had no con-reason, the action would have occurred just that much sooner.

It might be so, but then again, it might not be so. Consider cases in which the con-reason, although it remains a con-reason, is so obviously (to the agent) weaker than the pro-reason that the agent has no need to go through some actual deliberative process. The con-reason is not strong enough to act as the countervailing wind did in my earlier example, as a force dragging and delaying the decision to act. So the action taken and the action the agent would have taken had he had no con-reason would display no difference in time of occurrence.

I don't think these cases are at all far-fetched. If we focus on cases in which decisions are difficult, in which pro- and con-reasons are finely balanced, it will seem that the con-reason, if causalism is to be believed, must exert some sort of causal influence on the action taken, even if only a difference in temporal placement. But we are not normally so conflicted in our choice of action. Think instead of cases, and I suggest that these will be the vast majority of cases, in which, although 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, I 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.

So, let it just be stipulated that we are considering a case of Type II, in which the agent does something in the circumstances in which he does have conflicting pro- and con-reasons, but that he would have also done that action in an intrinsically qualitatively identical manner, and at the same time, had he only had the one set of pro-reasons. His "opposing" con-reason does not make him hesitate, or dither, in doing whatever it is that he does, in any way. There must, therefore, be cases in which the con-reason is not causally necessary for the actual action taken, if it is, as the causalist insists it is, a cause at all.

But if causalism is true, the con-reason must be a cause or part-cause or causal factor of something else other than the choice or action actually taken. In such cases, if causalism is true, the pro-reason will initiate a causal

chain leading to the action taken, and the con-reason must initiate a different, independent causal chain that leads to something else. One thing that the con-reason can certainly not cause is the action it favors, since that action never happened and therefore nothing could cause *it*. To be sure, *that* something does not occur can have a cause, but what does not occur can have no cause since it does not exist.

If we assume that the con-reason does not also contribute to the causation of the action it disfavors, but rather would have to cause something else, there is any number of possible candidates for the effects of such con-reasons available to the causalist. Perhaps a person's con-reason *directly* causes regret (Williams 1981, 27ff.), or causes some other change in his mental landscape (his dispositions to act, for example), or causes some psychological illness in him. He does the action favored by the pro-reason, but since he had reasons against it, his con-reason ends in him regretting what he did, or some such. Or perhaps the effect of the con-reason is not even at the personal level at all. Might its effect not be some physiological or brain event of which the actor is perhaps ignorant or unaware?[7] (Or, "some further physiological or brain event," if the having of a con-reason is such a physical event too.)

The important feature of all these candidate effects for cases of Type II is that they require a second causal chain, in addition to the one that goes from the stronger pro-reason to the action taken. If so, there would be one causal chain leading from his having a pro-reason to his subsequent action. There would be another quite independent causal chain leading from his con-reason to his subsequent regret, or illness, or to some (further) physiological or similar event. The causal chains would not converge causally on the final choice or action, as they would if both pro- and con-reasons causally contributed to the same action taken, as we sketched above in cases of Type I.

On this rather simple picture, the pro-reason initiates a causal chain leading to the action; the con-reason initiates a wholly independent, second causal chain, leading to the regret or brain state or whatever. One thing to note about this view is that it might not permit us to capture causally the idea that both pro- and con-reason are rationally or deliberatively relevant to the same token final choice or action. The con-reason might not be a reason against acting in a certain way in virtue of whatever causal role it plays. A con-reason could not be the con-reason it is (a reason not to do what was done) in virtue of its causing something else other than that action. At the level of reasons for choice and action, the two reasons bear differently on (one favors and the other disfavors) the same choice or action, but the causal story might not mirror this in any way. There would be just two distinct causal chains, each of which leads to a different result; one leads to an action, the other to some psychological or neurophysiological or dispositional state. But perhaps a causal model of how pro- and con-reasons work in choice situations need not capture within the causal model this fact about the rational significance that both types of reasons have to the same action or choice, one in favor of it and one against it, so I don't take this as a decisive objection to the suggestion under discussion.

My argument now focuses only on cases of Type II, such that pro- and con-reason initiate independent causal chains. I do not deny that there can be many cases of Type I, but these are not the ones I wish to consider. In the cases on which I now focus, the pro-reason initiates one causal chain and the con-reason initiates another, whatever it might be and to wherever it leads.

In Type II examples, consider the actual situation, *c*. In *c*, the pro-reason to *A* is rationally weightier for the agent than the con-reason to *B*. Causally, assuming (2), if there is no weakness of will, it is the pro-reason that causes the agent to *A*, rather than the con-reason causing the agent to *B* (so the con-reason causes something else, whatever that might be). But now consider a counterfactual situation, *c**.

*c** is just like *c*, save in one feature, and whatever is a causal consequence of that one feature: in *c**, although the pro-reason retains the same deliberative weight that it has in *c*, the con-reason becomes much weightier. This sort of scenario is very common. At a later time, an agent can assess a reason as having more "gravitas" than he earlier imagined it had. It might weigh more with him than it did before. So in *c**, the con-reason counts more for the agent. The agent does not judge that the reason to *A* has become less strong than it was; it is just that the reason to *B* has become deliberatively stronger, and so stronger than the reason to *A*.

The reason to *B* now rationally outweighs the reason to *A* in the agent's deliberations, so the agent now *B*s rather than *A*s. In *c**, the reason to *B* has become the pro-reason and the reason to *A* has become the con-reason. Something about the reason to *B* has changed, and consequently the ordinal information about relative strength of reasons has changed. But nothing about the reason to *A* need have changed, other than certain relational, ordinal truths about its deliberative strength.

At the level of decision, choice, and reason, this is all straightforward. But how should we represent the allegedly underlying causal facts of the matter in *c** (in order to obtain a coherent causalist story)? In *c**, the reason

to *B* is rationally weightier than the reason to *A*, so assuming that (2) is true and that there is no weakness of will, the reason to *B* will cause the agent to *B* instead of causing whatever it did cause in *c*. In *c**, there will now be a causal chain leading from the reason to *B* all the way to the agent's *B*-ing. But what does the reason to *A* now cause in *c**?

Before we try to answer that question, let's return for a moment to the question about the relational and intrinsic strengths of reasons for action that we briefly mentioned earlier on. Aside from the requirements of causalism, I do not know whether there are intrinsic truths about the deliberative strengths of reasons for action, on which their relational deliberative ordering supervenes. I would prefer to remain agnostic on that issue too. But it seems clear that causalism is committed to there being an intrinsic reality to the causal strength of reasons. On causalism, reasons are also causes (to put it succinctly), and there certainly must be an intrinsic reality to causal strength. The whole truth about causes and their relative strengths cannot be exhausted by only relational information. If c_1 is causally stronger than c_2, there must be something intrinsic about c_1 and c_2 that makes this relational fact true. One of the consequences of causalism's tying together rational and causal strength is that reasons must have strength both in a rational/normative and a causal/motivational sense, as I claimed earlier, and, as a result of that, reasons must also have causal strength in an intrinsic sense.

Now revert to our two circumstances, *c* (the actual situation in which the reason to *A* outweighs the reason to *B*) and *c** (the counterfactual situation in which the reason to *B* outweighs the reason to *A*). In *c*, the reason to *A*, the deliberatively strongest reason, caused the agent to *A*. Remember that we are supposing that the only difference between *c* and *c** is the fact that the reason to *B* in *c** rationally outweighs the reason to *A* and hence causes the agent to *B*. In *c**, the reason to *A* also undergoes a deliberative relational change, since in *c** it is outweighed by the reason to *B* but was not so outweighed before in *c*. But there is no reason to think that there must be some intrinsic causal change to the reason to *A*, simply in virtue of the deliberative and causal changes to the reason to *B*, remembering that in cases of Type II the causal chains initiated by the pro- and con-reasons are independent. Whatever changes the relational deliberative change to the reason to do *A* supervenes on, in the case described, they may not include any intrinsic causal change in the reason to *A*. The deliberative relational change to the reason to *A* (for it is now outweighed in *c**) may supervene only on changes, deliberative and causal, to the reason to *B*.

If so, then the reason to *A* should have the same causal strength in *c** as it had in *c* (even though it is now rationally outweighed by the reason to *B*), and since the reason to *A* caused the agent to *A* in *c*, then the reason to *A* should cause the agent to *A* in *c** as well (with one possible exception, described below). If the reason to *A* in *c* was strong enough to cause the agent to *A*, then it should still have the same causal strength in *c**, and therefore should be strong enough to cause in *c** whatever it caused in *c*, given that there are no causal nonrelational differences between *c* and *c** as far as the reason to *A* is concerned. If the reason to *A* has the same causal strength or power in both, then its effects should be the same in both circumstances. What it is strong enough to cause in one, it should be strong enough to cause in the other. The relational difference that in *c** the agent's reason to *A* is outweighed rationally by his reason to *B* can't make a difference to what the former is causally strong enough to do, since its causal strength is intrinsic.

So why doesn't the agent do *A* in *c**, just as he did in *c*? If the reason to *A* is able in *c* to cause the agent to *A*, and if it has the same causal properties in *c** that it had in *c*, then it should still cause the agent to *A* in *c**. True, the reason to *B* gains in deliberative strength in *c** (and so the relational facts about the relative strength of both reasons will change from *c* to *c**). Given the causalist's (2), what the reason to *B* causes, what its causal strength is, must have changed from *c* to *c**, a causal change on which its new deliberative strength supervenes. So the reason to *B* should also cause the agent to *B* in *c**. There should be, in *c**, as far as we can tell, a standoff: the agent should be caused both to do *A* and to do *B*.

To be sure, the agent can't do both *A* and *B*; by assumption, the agent is not able to do both on a single occasion. But in the counterfactual situation *c**, causally speaking, there should be no grounds for thinking that the con-reason will now win out "over" the pro-reason. The con-reason is now rationally and hence causally strong enough to cause the agent to *B*, but the pro-reason remains at the same intrinsic causal strength and hence, on the causalist view, should still be strong enough to cause the agent to *A*. So why should we expect the agent to do one or the other? Why doesn't the agent do *A* rather than *B*, even in the counterfactual situation, since his reason to do *A* remained in principle strong enough to cause him to do *A*, or why doesn't he do nothing at all, as in a true Buridan's ass example, since the two causes might cancel themselves out?

I mentioned one possible exception, above, to the claim that "since the reason to *A* was strong enough to cause *A* in *c*, then the reason to *A* should be strong enough to cause the agent to *A* in *c** as well." We need to take

note of this qualification. Perhaps the causal chains initiated by the reason to *A* and the reason to *B* are independent in the actual circumstances, *c* (they are not joint causes of a single effect or joint effects of a single cause), but they might not remain independent in the counterfactual situation *c**. If the reason to *B* gains deliberative strength in *c**, this relational change might supervene on some changed intrinsic causal fact about it. Suppose that in *c** the reason to *B*, in addition to causing the agent to *B*, is now able to interrupt the causal chain that would otherwise lead from the reason to *A* to the agent's *A*-ing, and that explains why the agent does not, after all, do *A* in *c**. Let's consider this possible rejoinder to the difficulty we have detected.

There would be some flexibility in deciding just where, in *c**, the requisite inhibitor blocked or stopped the chain commencing with the reason to *A* from leading to its "natural" conclusion, *A*, as long as the chain did not get all the way to that action. For the sake of argumentative simplicity, let us suppose that the reason to *B* inhibited the very next link on the chain. On such a chain, let *m* be the node that would have followed immediately after the reason to *A*. So let us say that, in the counterfactual situation, what happened is that the reason to *B* inhibited or prevented *m* from occurring, prevented or inhibited the reason to *A* from causing *m*, and hence prevented the action *A*. That is why the agent does *B* instead of *A* in the counterfactual situation, and why his reason to *A* does not lead to his *A*-ing in *c**, an explanation entirely consistent with the causalist position.[8] (In fact, the same result would be achieved if something else other than his reason to *B* was the blocker or inhibitor, but the reason to *B* is going to prove the most likely candidate for that role.)

The problem with this solution is simply that it is not true to the phenomenological facts of the case. What this purported solution does is to try and construe an agent's not acting on a causally strong enough reason that he has as a case of having that reason blocked or impeded by a conflicting reason that he also has. The identification doesn't succeed.

Even apart from cases of weakness of the will, there is an indefinitely large number of ways in which an agent's wishes, wants, desires, and so on can be thwarted. Bad luck affects us all. A typical sign of this happening is agent frustration. In weak-willed cases, according to causalism, the agent's rationally strongest reason does not commence a causal chain leading to an action because it is not causally strong enough; a rationally weaker but causally strong enough reason does.

On the other hand, in the rather different case we are now considering, the causalist rejoinder has it that the agent's otherwise-causally-strong-enough reason commences a causal chain that simply gets blocked. If an agent's causally strong-enough reason does not lead to action only because the causal chain leading from it to action is blocked in some way, the agent will be and feel thwarted. Something that he is causally driven to do, as much as he is causally driven to do what he does do, gets closed off to him.

Recall that on this view, since the reason to *A* is meant to retain in *c** whatever causal properties it had in *c*, it should therefore cause the agent to act in *c** as well as in *c*. If the agent failed to *A* in *c**, only because the causal efficacy of his reason to *A* had been blocked, even if by another reason, the agent would feel this as some sort of failure. In one kind of frustration or failure, an agent might feel frustrated because he did not do what he thought was rationally the best thing to do. This kind of frustration is more properly, perhaps, thought of as a kind of defeat. It is the kind of frustration that an agent experiences in cases of weakness of the will. The agent knew that he had a stronger reason to do *A*, but did *B* instead. Alas, his reason to *B* produced more psychic turbulence than did his reason to *A*, in spite of the fact that rational or deliberative strength should have inclined him otherwise. This is a case in which the agent acts on the rationally less weighty, but the causally more effective, reason.

On the other hand, the case we would need to envisage if the solution being proposed worked is different but would also give rise to a kind of frustration. It is a case in which the agent would fail to act on the reason (albeit the less weighty reason, rationally speaking) that otherwise is (equally) causally strong enough to drive him to act. In the case we are considering, the agent has a rationally stronger reason to do *B*, and a rationally weaker reason to do *A*. But (according to my argument) both his reason to *B* and his reason to *A* should cause him to act, since each is causally strong enough to lead to the action it supports.

But of course he does *B* in *c**, not *A*, and we asked why this should be so. We asked: in *c**, why doesn't his reason to *A* lead him to act too? The suggested reply is that it does not lead him to act only because something blocks the path leading from this otherwise causally strong enough reason to the action for which he has that reason, thereby preventing him from so acting. Here too, the agent would experience a kind of frustration. It is not defeat or rational frustration, as in the case of weakness of the will. It is the inability to act on a reason that would otherwise be causally strong enough for an action, whatever its relational rational weight might be. It is more like being unable to scratch an itch severe enough to drive you to scratch (but, to be sure, when you have a rationally overriding reason not

to do so), because the frustration arises from the causal failure, not from a rational failure. At a rational level, the agent would be happy that the reason to *A* did not lead to his *A*-ing, since he had more reason to *B* than to *A*. But at a causal level, he was primed to do *A* just as much as to do *B*. But if he does not do what he is causally primed to do, he would feel frustrated. It would indeed be like not being able to scratch an itch, when the desire to scratch was as causally strong as any reason not to do so was.

How much weight should we put on these sorts of phenomenological facts in deciding metaphysical matters? It is, I think, too easy to be a skeptic about this. What we are deciding are not just metaphysical matters generally, but specifically issues in action theory. If some view in action theory attributes to agents various kinds of mental states, or has the consequence that they have those states, what better check is there than introspection? I think that many views in action theory can be judged in this way, for example, ones that attribute various second-order mental states to agents, or ones that require of the agent almost a limitless stock of beliefs (Ruben 2003, chap. 4). One might dispute my argument that the causalist view does imply that the agent would experience the kind of frustration that I claim. But if the argument about this implication is sound, then introspection is the only way I know to test the claim.

What the causalist was trying to do was to give a causal model for the case in which the agent *B*s, because his reason to *B* has become deliberatively weightier than his reason to *A*, even though his reason to *A* has retained its original nonrelational, causal strength. In this case, surely the truth of the matter is that nothing needs to be thwarted and the agent need feel no frustration. He gladly "surrenders" his reason to *A*, at least in the circumstances, to his now-superior-because-weightier reason to *B*. It is not true that his reason to *B* prevents or blocks him from acting on his otherwise causally strong enough reason to *A*. In the case at hand, he chooses not to do *A*, because he takes his reason to do *A* as relatively of less importance or weight than his reason to do *B*, and in the case as we have constructed it, I do not see how this fact can be modeled causally. There is a perfectly clear deliberative story about what goes on in this case, but it is a story for which the causalist can provide no convincing causal counterpart.

There is, I submit, no fully convincing way causally to model decision making that includes con-reasons, at least for cases of Type II. It is the element of relational deliberative weight, comparative strength, which cannot be captured causally, at least in those cases in which the con-reason does not contribute causally to the action taken. What matters in deliberation is the comparative or relative strength of reasons. If reasons were causes, there would be nonrelational truths about the causal strength of reasons. Because of these nonrelational causal truths, the two scenarios, the causal/motivational and the rational/normative, won't mesh. As long as one thinks only about pro-reasons for action causing the actions they favor, the point is not salient. But once con-reasons are introduced, it becomes clearer that there is no plausible causal modeling for all the ways in which con-reasons work in our deliberation scheme.

Acknowledgments

An earlier version of this essay appeared as "Con-Reasons as Causes" in *New Essays on the Explanation of Action*, ed. C. Sandis (Basingstoke: Palgrave Macmillan, 2009), 62–74. Richard Bradley, LSE, has helped in improving this essay from what it otherwise would have been.

Notes

1. (1) is understood here to speak of causation, not necessarily only of deterministic causation. The causation in question might be probabilistic or stochastic. I frequently add "or probably" to cover cases of probabilistic or stochastic causation. My discussion should apply equally whether the causation in question is deterministic causation or stochastic causation.

   What are causes? No essay can do everything and, with that, I intend to beg off any responsibility for explicating the idea of causation. I am presupposing a fairly standard account of causation, on which causes are token events or token states, and that a causal chain is a series of such.

2. Although Jonathan Dancy (2000, 4) notes their existence: "but still I will normally speak as if all the reasons that do motivate all pull in the same direction."

3. A con-reason is also a pro-reason in its own right for the action not taken, and is a con-reason only in the sense that it counts against the action that was taken. Similarly, a pro-reason is only a pro-reason for the action taken and is itself also a con-reason for the action not taken. In what follows, to simplify terminology, I will only use the idea of a pro-reason to be the reason that counts for the action one takes, and the con-reason to be the reason that counts for the action one does not take, the reason that gets outweighed. In the light of this, it would be wrong to think of pro-reasons and con-reasons as two different sorts of reasons. I was careful above only to say that reasons can function in these two different ways, depending on context.

4. Instead of "relational" versus "intrinsic" (or, "nonrelational"), I would have spoken of "ordinal" versus "cardinal," but there seem to be presuppositions built into the latter contrast that do not necessarily exist in the former contrast. Even apart from the demands of causation that I discuss below, I doubt whether the whole truth about reasons is exhausted by merely relational information, although nothing in this essay requires that to be true. In a theoretical syllogism, a set of premises can confer a nonrelational probability on a conclusion, making the conclusion rational to believe to some degree; similarly, reasons for action can confer nonrelational support on an action, making it a rational action to perform to some degree. Reasons have an intrinsic strength as well as a comparative strength relative to other reasons one has, if indeed one has any others.

5. I have often wondered why the principle "Every event has an effect" does not have quite the same intuitive appeal as "Every event has a cause." It might seem obvious that they should stand or fall together.

6. Of course, the choice to *A* that he would have made or the *A*-ing he would have performed had he not had a reason to *B* must differ from the choice to *A* that he did actually make or the *A*-ing he actually did do in at least one way, simply in virtue of the fact that it would have been a choice made in the absence of having a conflicting reason to *B*. The qualification "in some intrinsic way" is meant to exclude such trivial differences.

7. I do not think that one should underestimate the importance of the shift from the personal to the subpersonal level, in order to maintain (1) and (2), broadened to include con-reasons. It is a major concession on the part of the causalist. I do not intend to develop the point here, but certainly the hope that lay behind the causalist program for reasons for action was that reasons could be construed as causes, yet doing so was compatible with understanding reasons and actions in their own terms, sometimes called "the space of reasons." This program was not necessarily committed to construing reasons and actions as "really" about brain states and gross behavior (even if they turn out to be identical to brain states and gross behavior). The language of psychology and action was meant to have an internal coherence and integrity all its own. To that extent, this option can easily take the causalist program somewhere it had not intended to go.

8. Note that this example is not one of preemption, as some have suggested to me. If it were a case of preemption, one would have two reasons both favoring the *same* line of action, the first of which causes the action and the other of which did not cause the action but would have caused the same action, had one not had the first reason. In causal preemption, the inhibition or prevention is by the preempting cause of some node on the chain that would have led from the preempted cause to the effect. This is certainly not the case we are considering. But, arguably, all cases of preemption involve some sort of causal inhibition or prevention, as does the case we are considering.

13 Teleological Explanations of Actions: Anticausalism versus Causalism

Alfred R. Mele

Teleological explanations of human actions are explanations in terms of aims, goals, or purposes of human agents. According to one familiar *causal* approach to analyzing human action and to explaining instances thereof, human actions are, essentially, events that have appropriate mental items (or neural realizations of those items) among their causes.[1] Many causalists appeal, in part, to such goal-representing states as desires and intentions (or their neural realizers) in their explanations of human actions, and they take acceptable teleological explanations of human actions to be causal explanations. Some proponents of the view that human actions are explained teleologically regard all causal accounts of action explanation as *rivals*.[2] I dubbed this position "anticausalist teleologism" (AT for short; Mele 2003, 38) and argued against it (ibid., ch. 2).

I revisit AT in this essay. In section 1, after providing some background, I rehearse an objection raised in Mele 2003 to a proposal George Wilson (1989) makes in developing his version of AT. In section 2, I assess Scott Sehon's (2005, 167–171) recent reply to that objection. In section 3, I articulate a version of Donald Davidson's (1963/1980) challenge to anticausalists about action explanation and assess Sehon's (2005, 156–160) reply to Davidson's challenge.

1 Wilson's Proposal and My Objection

Are there informative, conceptually sufficient conditions for such things as a human being's acting in pursuit of a particular goal that do not invoke causation? In chapter 2 of Mele 2003, I argued that attempts to identify such conditions by leading anticausalists about action explanation—Carl Ginet, Scott Sehon, R. Jay Wallace, and George Wilson—are unsuccessful. In the present section, I review Wilson's proposal and my objection to it.

Wittgenstein, L. 1958. *Philosophische Untersuchungen*. Frankfurt: Suhrkamp.

Wittgenstein, L. 1972. *Philosophical Investigations*. Trans. G. E. M. Anscombe. Oxford: Blackwell.

Woodward, J. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Woodward, J. Forthcoming. Causal perception and causal cognition.

Wright, J., and J. Bengson. 2009. Asymmetries in folk judgments of responsibility and intentional action. *Mind and Language* 24:24–50.

Yablo, S. 1992. Mental causation. *Philosophical Review* 101:245–280.

Yaffe, G. 2009. Commentary on Moore's "Intention, responsibility, and the challenges of recent neuroscience." *Stanford Technology Law Review*. http://stlr.stanford.edu/2009/02/intention_responsibility_and_t.html.

Young, L., F. Cushman, R. Adolphs, D. Tranel, and M. Hauser. 2006. Does emotion mediate the relationship between an action's moral status and its intentional status? Neuropsychological evidence. *Journal of Cognition and Culture* 6:291–304.

Yuill, N. 1984. Young children's coordination of motive and outcome in judgments of satisfaction and morality. *British Journal of Developmental Psychology* 2:73–81.

Zhabotinsky, A. M. 1964. Periodic processes of malonic acid oxidation in a liquid phase. *Biofizika* 9:306–311.

Zimmerman, M. 1981. Taking some of the mystery out of omissions. *Southern Journal of Philosophy* 19:541–554.

Contributors

Frederick Adams Professor of Cognitive Science and Philosophy, University of Delaware

Jesús H. Aguilar Associate Professor of Philosophy, Rochester Institute of Technology

John Bishop Professor of Philosophy, University of Auckland

Andrei A. Buckareff Assistant Professor of Philosophy, Marist College

Randolph Clarke Professor of Philosophy, Florida State University

Jennifer Hornsby Professor of Philosophy, Birkbeck College, University of London

Alicia Juarrero Professor of Philosophy, Prince George's Community College

Alfred R. Mele William H. and Lucyle T. Werkmeister Professor of Philosophy, Florida State University

Michael S. Moore Charles R. Walgreen, Jr. University Chair, Professor of Law and Philosophy, Professor in the Center for Advanced Study, and Co-Director of the College of the Law Program in Law and Philosophy, University of Illinois College of Law

Thomas A. Nadelhoffer Assistant Professor of Philosophy and Contributing Faculty to Law and Policy Program, Dickinson College

Josef Perner Professor of Psychology, University of Salzburg

Johannes Roessler Senior Lecturer in Philosophy, University of Warwick

David-Hillel Ruben Director of New York University in London and Professor, Birkbeck College, University of London