Program Explanation: A General Perspective

Article *in* Analysis · March 1990

DOI: 10.2307/3328853

CITATIONS
188

READS
46

2 authors:

Frank Cameron Jackson
Australian National University
181 PUBLICATIONS **6,308** CITATIONS

SEE PROFILE

Philip Pettit
Australian National University
361 PUBLICATIONS **11,542** CITATIONS

SEE PROFILE

or something close to that. Because I reject B, I can find no plausible way of getting from A to the voluntariness of belief. B has been popular through the centuries, but it seems pretty clearly to be mistaken. If it were right, the following wish would be unintelligible:

> I wish that my arm would rise right now without my raising it, going up simply as an immediate consequence of my wanting it to go up.

This seems to me perfectly intelligible, and I offer that as one way of seeing that whatever we mean by doing something voluntarily it is more than, or different from, it happening as an immediate consequence of wanting it to happen.

Still, the threat is valuable. It warns us of further complexities in our concept of voluntary conduct. The question of whether or why belief is essentially involuntary may be unanswerable until those further complexities are understood.[22]

Syracuse University,
Syracuse, NY 13244-1170, U.S.A.

[22] My debt to friends who have helped me with this work goes well beyond what is indicated in previous footnotes. All of them gave me other help also, enabling me to improve several earlier sections of the paper. I am truly grateful for their assistance.

PROGRAM EXPLANATION: A GENERAL PERSPECTIVE

By FRANK JACKSON *and* PHILIP PETTIT

SOME plausible assumptions generate a serious problem about the role of properties in causal explanations. This paper sets out those assumptions, identifies the problem they generate and makes a case for a particular solution.

The paper is in three sections. The first section introduces three plausible assumptions about the role of properties in causal explanations. We will not be questioning these assumptions here. The second section shows how these assumptions generate a problem when combined with a fourth assumption that looks equally difficult to resist. The third section offers a solution to the problem by showing how, nevertheless, we can and should resist this fourth assumption.

The fourth assumption is that the only way in which a property can be causally relevant to an effect is by being causally efficacious

in its production. In order to resist that assumption, the paper draws on an account of causal relevance under which a property can be causally relevant without being causally efficacious. According to that account the realization of a property may program for the occurrence of an effect without actually contributing to its production. The paper, as the title suggests, provides another perspective on that programming account.[1]

I The Three Background Assumptions

The three following assumptions provide the backdrop for our discussion.

1. A causal explanation of something must direct us to a causally relevant property as opposed to a causally irrelevant property of the factor it identifies as explanatory: a property relevant to the causal production of the effect explained.
2. One way in which properties are causally relevant is by being causally efficacious. A causally efficacious property with regard to an effect is a property in virtue of whose instantiation, at least in part, the effect occurs; the instance of the property helps to produce the effect and does so because it is an instance of that property.
3. A property F is not causally efficacious in the production of an effect *e* if these three conditions are fulfilled together.
   (i) there is a distinct property G such that F is efficacious in the production of *e* only if G is efficacious in its production;
   (ii) the F-instance does not help to produce the G-instance in the sense in which the G-instance, if G is efficacious, helps to produce *e*; they are not sequential causal factors;
   (iii) the F-instance does not combine with the G-instance, directly or via further effects, to help in the same sense to produce *e* (nor of course, *vice versa*): they are not coordinate causal factors.

The first assumption hardly needs further paraphrase, though we should note that it is meant to hold good regardless of how causes are identified and individuated in particular cases. The second assumption is equally straightforward but it is worth remarking that the notion of efficacy it introduces is not tied to the view that causal efficacy is an irreducible feature of the world;

[1] This account is developed in Frank Jackson and Philip Pettit, 'Functionalism and Broad Content', *Mind* 97 (1988) 381–400, and 'Structural Explanation and Social Theory', in David Charles and Kathleen Lennon, eds, *Reductionism and Anti-reductionism* (Oxford: Oxford University Press, forthcoming).

it is compatible with a more or less debunking analysis of efficacy, say in terms of causal laws. One thing to note, however, is that no matter how the notion of causal efficacy is understood, it is distinct from the notion of instrumental effectiveness. A property will count as instrumentally effective *vis-à-vis* a particular effect, if it would have been a good tactic for producing the effect to realize that property. But such effectiveness does not entail efficacy: it does not mean that the effect occurred in virtue of the instantiation of the property.

The third assumption is less intuitively obvious than the other two and requires additional commentary. First, some elucidation. One way that F might be efficacious only if G is so, is by the F-instance and the G-instance being respectively more remote and more proximal causes of *e*; if they were sequential factors of this kind, both properties might be causally efficacious consistently with the truth of 3(i). 3(ii) is designed to eliminate this case. Similarly, 3(iii) is designed to make it clear that 3(i) is not true, just because the instances of F and G are each necessary parts of a causally productive complex of factors; if they were coordinate factors of that kind, the two properties might be causally efficacious consistently with 3(i). Notice that for all that 3 says, the instances of the distinct properties F and G may be identical. If G is efficacious in such a case, being a property in virtue of whose instantiation *e* occurs, still that will not make F efficacious: the instance of F will help to produce *e* but not because it is an instance of F; instead it will do so because it is an instance of G.[2]

What sort of situation would make 3(i), 3(ii) and 3(iii) together true? Well, here are some examples. In all of them the F-factor can be thought of as higher order and the G as lower order.

(A) A fragile glass is struck and breaks. Why did it break? First answer: because of its fragility. Second answer: because of the particular molecular structure of the glass. The property of fragility was efficacious in producing the breaking only if the molecular structural property was efficacious: hence 3(i). But the fragility did not help to produce the molecular structure in the way in which the structure, if it was efficacious, helped to produce the breaking. There was no time-lag between the exercise of the efficacy, if it was efficacious, by the disposition and the exercise of the efficacy, if it was efficacious, by the structure. Hence 3(ii). Nor did the fragility combine with the structure, in the manner of a coordinate factor, to help in the same sense to produce *e*. Full information about the structure, the trigger and the relevant laws would enable one to predict *e*; fragility would not need to be taken into account as a coordinate factor. Hence 3(iii).

[2] On related matters see Cynthia and Graham Macdonald, 'Mental Causes and Explanation of Action', *Philosophical Quarterly* 36 (1986) 145–58.

(B) I try and fail to fit a square peg in a round hole of diameter equal to the side of the square. Why did it not go through? First answer: because of the squareness of the peg. Second answer: because of the impenetrability of this overlapping part of the peg. The property of squareness was efficacious only if the overlap-cum-impenetrability-property was efficacious: hence 3(i) is true. But 3(ii) is also true, for the squareness did not help to produce the overlap-cum-impenetrability in the way in which it, if efficacious, helped to produce the blocking of the peg: there was no time-lag of the sort that such an influence would seem to require. And 3(iii) is also true, for the squareness did not combine with the overlap-cum-impenetrability to help in the same sense to produce the blocking; one could have predicted the blocking without reference to the squareness. As we might put it, the overlap-cum-impenetrability did not need any extra help from the squareness to produce the blocking.

(C) The water in a closed glass container reaches boiling temperature — the mean molecular motion is at such and such a level — and the container cracks. Why did it crack? First answer: because of the temperature of the water. Second answer, in simplified form: because of the momentum of such and such a molecule (group of molecules) in striking such and such a molecular bond in the container surface. (We are supposing that the case is one where the container breaks because of the internal pressure, not because of the temperature gradient between the water and the container.) The temperature-property was efficacious only if the momentum-property was efficacious: hence 3(i). But the temperature of the water — an aggregate statistic — did not help to produce the momentum of the molecule in the way in which it, if efficacious, helped to produce the cracking: hence 3(ii). And neither did the temperature combine with the momentum to help in the same sense to produce the cracking: one could have predicted the cracking just from full information about the molecule and the relevant laws. Hence 3(iii).

Is assumption 3 plausible? Well, given that the F and G properties do not relate as sequential or coordinate causal factors, it is certainly plausible that F cannot be efficacious in the same sense as G. Moreover it is plausible that if both are efficacious, then F is efficacious only in a derivative sense. The relation between the instantiation of F and the occurrence of *e* is secondary to the relation between the instantiation of G and that occurrence; other things being equal, the obtaining of the latter relation ensures the obtaining of the former, and not *vice versa*. But is it reasonable to go beyond these two plausible claims and endorse the claim in 3, that the F-property is not efficacious in any sense in producing *e*?

This is reasonable, and in two ways: strategically and theoretically. It is theoretically reasonable, because on any account of efficacy, it is Pickwickian to describe the F-property as efficacious,

given that any efficacy it is alleged to have exercised would have been screened off by the influence of the G-property. No conception of efficacy, no matter how debunking, should allow that efficacy can be exercised across such a screen. Whatever the conception of efficacy in play, it must be admitted that if the instance of G helped to produce *e* because it was an instance of G, and if the instance of F was not a sequential or coordinate factor, then F cannot have been efficacious in the production of *e*: the instance of F cannot have helped to produce *e* or if it did — if it was identical, say, with the instance of G — then it cannot have done so because it was an instance of F. And apart from those considerations, there is also this: that if higher-order properties are countenanced as efficacious, it seems we can invent efficacious properties at will. Not only was the fragility, the property of having a suitable molecular structure, efficacious in the breaking. So was the property of having such a property — if you like, meta-fragility; so indeed was the property of having that sort of property in turn — meta-meta-fragility; and so on into absurdity.

But it is strategically as well as theoretically reasonable to assert 3, rather than just admitting a derivative and a primitive sense of efficacy. Either being derivatively efficacious is a way for a property to be causally relevant or it is not. If it is not — if primitive efficacy is the only mode of causal relevance — then the problem to be raised in the next section remains, as will be there apparent. If derivative efficacy is a mode of relevance on the other hand, then while our problem goes away, it is replaced by a counterpart that requires the same sort of solution; it would require a solution like the proposal in Section III below. The counterpart problem is how to relate the two modes of relevance, how to make sense of the way different explanations of the same event can invoke properties that are efficacious at different levels: at the bottom level, primitively efficacious, and then efficacious at progressively more derivative levels.[3] For these reasons we shall not concern ourselves further with the notion of derivative efficacy and will take assumption 3 as given.

II The Problem

Our three background assumptions have a devastating impact once combined with the following proposition.

4. The only way for a property to be causally relevant to the production of a certain effect is by being causally efficacious in the process of production.

[3] This is a problem that arises, for example, for the theory presented in Peter Menzies, 'Against Causal Reductionism', *Mind* 97 (1988) 551–74.

Assumption 2 says that one way to be causally relevant is to be causally efficacious; this assumption adds that that is the only way. The combined impact of the four assumptions bears first on the possibility of causal explanation outside basic science and secondly on the possibility of causal explanation within.

The four assumptions entail that causal explanation in terms of causally relevant properties is only to be found within the realms of basic science: presumably, physics. For it is only in those realms that we seem to confront the sort of property which escapes being shown to be causally irrelevant by the combination of 1, 2, 3, and 4. This means that the four assumptions will drive us to dismiss the claims of the special sciences, and of course the claims of common sense, to be able to provide causal explanations in terms of causally relevant properties. For consider the sorts of properties invoked in those areas: the property of a group that it is cohesive; of a mental state that it is the belief that p; of a biological trait that it maximizes inclusive fitness. For each of these properties it is plausible that there is a property G lower down, so to speak, which is such that the higher up property F is efficacious only if G is, and yet the F-instance and the G-instance are neither sequential nor coordinate causal factors. But then by assumption 3, F is not causally efficacious, and by assumption 4 is not causally relevant. We shall have to regard all such properties as inefficacious, and so as irrelevant, and so as incapable of playing a role in causal explanation. The only properties with any claim to causal relevance and a proper place in causal explanation will be properties like mass and charge.

But even those who would happily turn their backs on common sense and the special sciences are going to be troubled by the combination of the four assumptions. Suppose that I explain the noise made by some mechanism by the property of the mechanism that some of its parts are loose. That property relates as F relates to G to the following more specific property: that this and that particular part are loose. It is the property, after all, of instantiating some such specific property, perhaps this, perhaps that, perhaps another one. Thus the explanation involving existential quantification — the reference to an indeterminate some — cannot be a proper explanation: it does not invoke an effacious property and so does not invoke a property that is relevant to the noise. The lesson holds quite generally, so that many of the explanations which physicists would endorse must look suspect. We cannot claim to explain the presence of vapour near the surface of boiling water by the fact that some of the water molecules have broken free. And we cannot claim to explain the radiation emitted by a piece of uranium by the fact that some of its atoms are decaying.

How might we hope to live with such results? They would mean that if in any area we gain access to a lower-order explanation of something, then we should jettison higher-order explanations in its favour. That in itself seems unattractive. We might want to give

up the fragility explanation of why the glass breaks if we had access to the account in terms of molecular structure; that is unsurprising, since anyone who had access to the latter account would have all the significant information at his disposal which is offered by the fragility explanation. But would we want in parallel circumstances to give up the squareness explanation, the temperature explanation, the special-science explanations or explanations that make use of existential quantification? It seems highly doubtful, given that they offer us information which is not available just in virtue of having access to the lower-order counterparts. Someone who knows that the impenetrability of this part of the peg stopped it going through that hole does not necessarily know all that is known by someone who explains the blockage by the squareness of the peg. And so too for the other examples.

But we do not have access to lower-order explanations in many of these cases, so perhaps the assumptions are going to be palatable after all. Not so. The assumptions mean that we must view higher-order explanations, even when they are the only accounts available, as rough-and-ready stand-ins for explanations proper. The relation between such a stand-in and an explanation proper will be like the relation we might have thought existed between these two accounts of why the house is down: that it is because of the meteor that struck it yesterday; and that it is because of the event reported in today's newspaper. As we might have said that someone with knowledge only of the second account is certainly ignorant of the property of the explanatory factor in virtue of which it explains the destroyed house — the momentum of the meteor — so we will have to say that someone with access only to higher-order explanations of things will be constitutionally ignorant of the properties that explain the matters with which he is concerned.[4] For the properties he knows about are literally irrelevant.

All of this hardly makes for a satisfactory scenario and it hardly answers to our sense of how we are actually placed. Hence a problem. The problem is to find a way out, in particular to find a ground for rejecting the only one of our four assumptions which looks questionable: assumption 4. Can we offer an account of causal relevance which allows an inefficacious property to be relevant to the production of an effect?[5]

[4] Perhaps a more useful analogy is with two accounts of the property in virtue of which the meteor knocked the house: the false, folk account which holds the weight responsible and the proper account which refers us to the momentum of the meteor.

[5] For some recent discussions of this sort of problem see Simon Blackburn 'Losing Your Mind' in John Greenwood, ed., *Proceedings of the Greensboro Conference* (New York: Oxford University Press, forthcoming); Ned Block 'Can the Mind Change the World?' in George Boolos, ed. *Meaning and Method: Essays in Honor of Hilary Putnam* (Cambridge: Cambridge University Press, forthcoming); and Fred Dretske, *Explaining Behaviour* (Cambridge, Mass.: MIT Press, 1988).

III A Solution

We can. In order to motivate the solution, consider a higher-order explanation involving existential quantification. Consider the explanation of why a piece of uranium emitted radiation over a certain period, which invokes the property of the uranium that some of its atoms were decaying: this, rather than the more specific property that such and such particular atoms were decaying. By our assumption 3, the property involving existential quantification cannot have been efficacious in producing the radiation. If it was efficacious, that is only because the more specific property was efficacious. And yet the instance of the abstract property did not relate to the instance of the more specific in the manner of a sequential factor or a coordinate one. So is there any other way in which the abstract property can have been causally relevant to the radiation, given that it was not causally efficacious?

Yes, there is, and the answer is more or less obvious. Although not efficacious itself, the abstract property was such that its realization ensured that there was an efficacious property in the offing: the property, we may presume, involving such and such particular atoms. The realization of the higher-order property did not produce the radiation in the manner of the lower-order. But it meant that there would be a suitably efficacious property available, perhaps that involving such and such particular atoms, perhaps one involving others. And so the property was causally relevant to the radiation, under a perfectly ordinary sense of relevance, though it was not efficacious. It did not do any work in producing the radiation — it was perfectly inert — but it had the relevance of ensuring that there would be some property there to exercise the efficacy required.

How are we to describe the relationship between such a property and an effect? The realization of the property ensures — it would have been enough to have made it suitably probable — that a crucial productive property is realized and, in the circumstances, that the event, under a certain description, occurs. The property-instance does not figure in the productive process leading to the event but it more or less ensures that a property-instance which is required for that process does figure. A useful metaphor for describing the role of the property is to say that its realization programs for the appearance of the productive property and, under a certain description, for the event produced. The analogy is with a computer program which ensures that certain things will happen — things satisfying certain descriptions — though all the work of producing those things goes on at a lower, mechanical level.

The solution proposed for the problem we have been confronting is that in each case the higher-order, inefficacious property is

causally relevant to the event produced, because its realization programs for the realization of a lower-order efficacious property and, in the circumstances, for the occurrence of the event in question. The lower-order efficacious property may not be the lower-order property mentioned in each case but, if it is not, it will be one for which the realization of that property directly programs, one for which the realization of a property programmed for by that property directly programs, or whatever: it will be a property for which the original property programs indirectly, via the programming of intermediate properties.

The only way to bear out the proposed solution is to show that it plausibly applies in all the sorts of cases considered earlier. The fragility of the glass ensures, by the very meaning of what it is to be fragile, that the glass has a molecular structure — maybe this, maybe that — sufficient in the circumstances to produce the breaking. The squareness of the (impenetrable) peg ensures, as a matter of elementary geometry, that there will be an impenetrable part of the square end to obstruct its passage through the hole and again it may be this part or that which provides the obstruction. Finally, the temperature of the water more or less ensures — it makes it probable to a point approaching certainty — that a suitably situated molecule will have a momentum sufficient to break a molecular bond in the container and therefore to produce a cracking. What if the lower-order property mentioned is not itself an efficacious one? If it is not, then the story will go a level or more deeper until we find an efficacious property for which the original higher-order property programs indirectly, via the programming of the intermediate features.

It appears then that there are at least two distinct ways in which a property can be causally relevant: through being efficacious in the production of whatever is in question, or through programming for the presence of an efficacious property. The general problem raised in the last section can be solved by means of this observation, for the observation gives the lie to the troublesome assumption 4. It suggests with the sorts of cases used to illustrate the F–G relationship, for example, that the property playing the F-role has programmatic relevance to the occurrence of the event *e*.

Does the solution extend to other cases at which we have gestured? Well, it is clearly going to work with any explanations which are higher-order in virtue of using existential quantification. And equally, though we shall not explore the cases here, it seems to extend to explanations in common sense and the special sciences: for example, to explanations in sociology which invoke a property like group-cohesion, to explanations in psychology which invoke attitudinal contents as causally relevant properties, and to explanations in biology which appeal to a property such as that of maximizing inclusive fitness. In all such cases it is hard to see how the explanations can have the interest they clearly possess other

than through being of the programming variety. The Yiddish *modus tollens* applies. If not this, what?[6]

Does the solution mean that in all of these cases we are entitled to provide the explanation we offer only if at some level there is an efficacious property at work for which the property programs? Yes, given there are efficacious properties, as our second assumption implies; but if there is an infinite progression of levels downward and therefore no effacious properties — by our third assumption — then the program story will have a different significance, bearing on relations between equally non-efficacious levels. Does the solution mean that we must be able to identify the efficacious property in question: that we must be able to provide the corresponding process explanation? No; it will do to have grounds for believing just that some such property must be in operation. In fact a little reflection suggests that perhaps most of the explanations we are ever likely to offer will be program explanations. Presumably we only reach potentially efficacious properties in physics. And presumably we will have access to these other than via existential quantification only in dealing with specific microphysical particles. But we will rarely be in a position to deal with specific particles and so most of the explanations we are ever likely to offer will be of the program variety.[7]

In conclusion, there is one point which it will be useful to emphasize. The notion of a programming property does not just explain how an inefficacious property can be relevant to the causation of an event. It also shows how a program explanation can have a significance that remains in the presence of an explanation invoking the corresponding efficacious property — the corresponding process explanation — and more generally in the presence of a lower-order explanation, whether it is of the program or process variety. A program explanation of an event e may provide information which the corresponding process explanation does not supply. Thus, it may be an explanation which the process explanation does not supersede.

The fragility case, as we have seen, is one where the claim fails: under natural background assumptions, we can say that someone who understands the molecular structure which accounts for the breaking of the glass understands all that is grasped by someone who offers the fragility account. The reason is that all it means to be fragile is to be such — say, to have such a molecular structure —

[6] For more detailed argument in the psychological and sociological cases respectively, see the papers mentioned in footnote 1.

[7] This is to take the program-process distinction as absolute. It can also be treated as a relativized distinction, with an arbitrary level of explanation being designated as involving causal process and then higher levels being cast as programming; in this sense, psychological properties would program for neurophysiological process, even if neurophysiological properties are not strictly efficacious. We take this approach in 'Structural Explanation and Social Theory'.

that breaking occurs under the relevant sort of knock. But the fragility case is exceptional. Someone who understands the lower-order explanations relevant in the other sorts of cases — even if they are process explanations — does not necessarily grasp the information available to someone who has access to the program explanations. If the molecular structure of a glass causes it to break in suitable circumstances, then the glass is fragile. But a part of the peg can stop it going through the hole without the peg's being square and the momentum of a water molecule can crack a container without the water's being at boiling temperature. Thus to know that the squareness and the temperature are explanatory, programming for the results in question, is to have information which is not available from the corresponding process explanations.

The point we are emphasizing can be put in other terms. According to David Lewis, to explain something is to provide information on its causal history.[8] Let us interpret the causal history as the process, involving such and such efficacious properties, that leads to the event or whatever in question. A program explanation provides a different sort of information from that which is supplied by the corresponding process account and therefore a sort of information which someone in possession of the process account may lack. The process story tells us about how the history actually went: say that such and such particular decaying atoms were responsible for the radiation. A program account tells us about how that history might have been. It gives modal information about the history, telling us for example that in any relevantly similar situation, as in the original situation itself, the fact that some atoms are decaying means that there will be a property realized — that involving the decay of such and such particular atoms — which is sufficient in the circumstances to produce radiation. In the actual world it was this, that and the other atom which decayed and led to the radiation but in possible worlds where their place is taken by other atoms, the radiation still occurs.[9]

Research School of Social Sciences,
Australian National University,
Canberra ACT 2601

[8] See the paper 'Causal Explanation' in his *Philosophical Papers,* Vol. 2 (Oxford: Basil Blackwell, 1988).

[9] This paper was written while Philip Pettit was a Visiting Fellow at Corpus Christi College, Oxford, with visitor's facilities at Nuffield College. He is grateful to both institutions for their support. We are also grateful for comments received on an earlier draft from Simon Blackburn, John Campbell, Jennifer Hornsby, Cindy and Graham Macdonald, Peter Menzies, David Miller, Michael Tooley, Tim Williamson and the members of a discussion group at Oriel College, Oxford.