LSE CPNSS

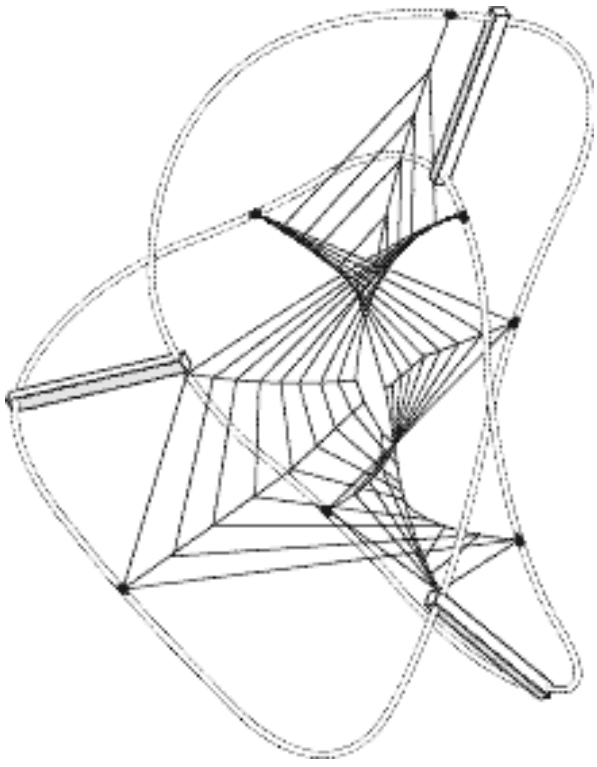
# Centre for Philosophy of Natural and Social Science

## Causality: Metaphysics and Methods

Technical Report 14/03

# *Defining Causal Strength*

Robert Northcott



Editor: Julian Reiss

# Defining Causal Strength

Robert Nortcott
Department of Philosophy, Logic and Scientific Method
London School of Economics
Houghton St
London WC2A 2AE
*r.d.northcott@lse.ac.uk*

October 2003

# 1 A Unified Scheme

*Introduction*

Did Iraqi children starve because of sanctions or because of Saddam Hussein's government? Should I take the left turning to avoid the roadworks, or take the right turning to avoid the multiple traffic lights? Is IQ due more to nature or nurture? It is commonplace to observe that causal modelling is ubiquitous in science, and indeed everyday life. But almost as ubiquitous is the notion of causal strength, and the related one of some causes being more or less important than others. This issue might appear fundamental, but is in fact relatively neglected in the philosophical literature. Whereas the general metaphysics of causation has received vast coverage, this subsidiary challenge of comparing two causes' strengths has not. Yet meeting it turns out to be a surprisingly intricate task.

Our account of causal strength will aim to satisfy the following desiderata:

1) The concept should be *univocal*, that is deliver unambiguous results. It turns out that there can seem to exist at least two distinct notions of causal strength, so satisfying this requirement needs more work than might initially have been suspected.
2) Results should be (sufficiently) *objective*.
3) Results should also be *quantitative*.
4) The definition should be widely *applicable*. For example, it should be able to deliver results even in cases of causal interaction.
5) Results should carry some *normative* force.

Note from the start two aspects of our investigation. First, we are concerned only with calculating the strengths of causes that are already *given*. Therefore we say nothing about the venerable issue of how to define or identify those causes in the first place. Second, our concerns here are primarily conceptual rather than epistemological. All we seek is a satisfactory definition of what it means to say that one cause is stronger than another.

*Two kinds of causal strength?*

It might seem that our task is a pretty straightforward one: can we not just define the strength of a cause by the quantity of effect it leads to? But consider the following story. Suppose that Holmes shoots Moriarty, but that if he had not then Watson would have done so anyway. What strength should we then assign to Holmes's shot as a cause of Moriarty's death? One analysis runs: Moriarty was killed by the bullet fired by Holmes, therefore his death was a direct consequence of Holmes's shot, therefore Holmes's shot should be assigned maximum causal strength. To use different words, Holmes's shot had full causal *potency* here. (Throughout, we shall use 'strength', 'potency', and 'importance' interchangeably.) Call this sense of causal strength 'potency-magnitude', or PM.

But there also exists a second analysis, which runs: given that Watson would have shot Moriarty in any case, in fact Holmes's shot *made no difference*. Whether Holmes fired or not, Moriarty would still have died either way. Accordingly, we should not assign Holmes's shot any causal strength after all. Call this second sense of causal strength 'difference-magnitude', or DM.

Notice immediately that the PM sense of causal strength is unable to distinguish

between the case in which Watson is present and the case in which he is not, whereas of course DM can. Hence the two senses are indeed distinct and so may diverge, as in this example. Accordingly, we must investigate whether causal strength really can be given a univocal understanding after all.

*Preliminary definitions*

Begin by stating more formally what we understand by DM and PM, starting with DM. What difference does a cause make? The definition of a DM must include some specification of what the world would have been like if the cause in question had *not* operated. For instance, if Holmes had not fired, then Watson would have done so anyway. Label the cause at issue 'C', so in this example C = Holmes's shot. Label the relevant effect 'E', so here E = Moriarty's death. We want some specification of the 'alternative' counterfactual cause, i.e. of what would have happened had Holmes not fired. Label that 'D', so here D = Watson's shot. Lastly, we shall need some term to represent all the implicit background conditions, such as that Holmes and Watson knew how to fire their guns, that Moriarty did not have on a bullet-proof vest, and so on. Label these assumptions 'W', for the state of the whole world just excluding our specific causes of interest C and D. Then we can define the DM as follows:

DM of C relative to the counterfactual D = E(C&W) - E(D&W)

In words, the DM of C relative to D is the effect given C minus what the effect would have been given D instead. Note that any assignation of DM is therefore only ever relative to some choice of counterfactual D. There is no such thing as some 'absolute' DM defined independently of counterfactual context (or rather to the extent that there is, this is what we call PM - more on which presently). This is desirable, since the idea of a cause 'making a difference' surely presupposes some context of comparison - made a difference relative to *what*?

In our Holmes-Moriarty example, if Watson would have shot Moriarty anyway then Moriarty dies whether or not Holmes fires. So, taking Moriarty's death to be E = 1 and his survival to be E = 0, the DM of Holmes's shot is given by the formula as: E(C&W) - E(D&W) = 1 - 1 = 0. That is, Holmes's shot indeed made no difference.

The formula is clearly readily extendable to cases where there is more than one counterfactual. For example, suppose that if Holmes had not shot then either Watson would have shot or else Inspector Lestrade would have entered and shot instead (cause 'L'). Each of these possibilities could be given some weighting in the formula and the DM then calculated. For instance, if we used weightings of k1 and k2 for Watson and Lestrade respectively, corresponding, say, to the probabilities of them being the ones to fire the alternative shot, and if Watson would have hit and killed Moriarty whereas Lestrade would have missed him, then the DM of Holmes's shot would now be: E(C&W) - k1 x [E(D&W)] - k2 x [E(L&W)] = 1 - k1. (k1 and k2 in this formula are constants, and serve as multiplying coefficients of the effect functions E.) Thus, because of the possibility of Moriarty otherwise surviving, Holmes's shot now did make some difference after all, in proportion to the chances of the back-up shot being fired by the errant Lestrade rather than the reliable Watson.

Move on now to our other kind of causal strength, PM. We have just seen how the

values our formula yields for DM depend in part on our choice of counterfactual. By contrast, the concept of causal potency, i.e. PM, intuitively seems to be intrinsic and local (on which more in section 3). Nevertheless, I propose that for our purposes PM is also adequately definable by using this same counterfactual technique. In particular, the potency of a causal input can be defined by reference to *the specific counterfactual of that input being totally absent*, with no other input taking its place. So the 'choice' of counterfactual here is no real choice at all - it is always the possible world exactly the same as the actual one in all respects except that the particular cause in question is absent. Note immediately that sometimes it will not be clear just how to interpret a cause's 'absence' - this issue will be addressed shortly. But for now, for E = effect, C = cause, and W = the rest of the world in addition to C, define the causal potency of C as follows:

PM of C = E(C&W) - E(W)

This is really quite intuitive when applied to everyday examples. For instance, to determine the causal strength of throwing a brick at a window, we would compare the window with the brick thrown at it ('E(C&W)') with the window with no brick thrown at it ('E(W)'). In words, a cause's strength is just the quantity of its impact on the effect, holding all other causes constant. (We shall ignore here the issue of other possible formulations, such as using the quotient of the effect terms instead of their difference.)

*Some technical wrinkles*

We need to iron out and clarify some technical wrinkles. First, in general the effect term may be an event or a variable. So far, we have been assuming the latter and hence expressing E as a function of causal input and background conditions. This is fine if, for example, E is air temperature. But suppose that the effect is more naturally thought of as an event that dichotomously either does or does not happen - for instance, if E is getting cancer. In that case, we may want instead to speak in terms of *probabilities* and adjust our notation accordingly. For example, the PM of some carcinogen C with respect to the effect E of getting cancer should be written:

p(E|C&W) - p(E|W),

where p denotes a probability function, the probabilities concerned being conditional ones. The formula for DM would be adjusted similarly.

Next, a more tricky issue - what exactly do we mean when citing the 'absence' of a cause C in the right-hand side of our PM formula? It is true that in many cases the interpretation of such an absence will be natural and unproblematic. For instance, if C is throwing a brick at a window then the absence of C would simply be not throwing the brick. However, just like effects, causes too may be either an event or a variable, and in the latter case problems arise. For example, suppose the cause of interest is a hot air temperature, say 50 degrees. What would be the 'absence' of such a cause? We could hardly speak of the absence of *any* air temperature, but at the same time there is no immediately obvious fallback point to adopt as our baseline reference temperature. Choosing freezing point, for instance, may sometimes seem odd - 'how much did the hot day cause me to sweat?' seems if anything to imply a contrast with average rather than

freezing ambient temperature. Yet on other occasions, such as the query 'how strong an effect does air temperature have on the speed of evaporation of a puddle?', the reference temperature might be freezing point after all. There seems to be no obvious general answer.

I propose to follow [Humphreys 1990] on this issue, and to appeal in general to what he terms the *neutral level* of causal input. This he defines (p38), in the case of a variable, as 'the level of the variable at which the property corresponding to that variable is completely absent.' A key point is that this neutral level is likely to depend on the exact effect of interest and on the exact context. For example, it may be that our focus of interest is the *change* in the level of our variable, in which case the neutral level would of course just be the original level before the change.

The same problem can be relevant even if we interpret C to be an event. For example, suppose C is ambient air pressure at sea level and we were interested in its effect on the optimal volume of a lung. We might interpret the 'absence' of this C to be a vacuum, but of course a vacuum would definitely not be neutral with respect to this particular effect, since presumably a vacuum would leave no role for a lung at all. Depending on our interest, we might instead want to use as comparison the air pressure at higher or lower altitudes, or perhaps the rate of change at sea level of air pressure with respect to altitude. Again, the point is that it is not enough simply to cite in the formula the 'absence' of the cause C, since the interpretation of this will be unclear. We must instead always define a neutral state in which, to repeat Humphreys's formulation, 'the property corresponding to that event is completely absent'.

Another example of a neutral state that can be awkward for simplistic accounts is when it is suggested that we set the absence of C to be just the level of cause that leads to *zero* effect. But sometimes even in the absence of C the level of effect is non-zero, and it is this latter level that we should use as our baseline. For example, the probability of getting lung cancer for non-smokers is greater than zero, so when calculating the causal strength of smoking we should take as a baseline this non-zero level.

A key point which Humphreys stresses is that this neutral level is *objective*. By this he means that, once *given* the specification of our cause, effect and context of interest, the neutral level is then defined objectively. The pragmatics only enter, as it were, in setting the background parameters; after that, the definition of the neutral level follows automatically and unambiguously.

We mention two further issues, each concerning the specification in our formula of the background conditions 'W'. First, we should note the danger here of phenomena such as Simpson's paradox. In particular, if we do not hold fixed all other causal parents of an effect while varying our cause of interest, any results obtained for causal strength may be misleading. This is really just the logic of controlled experiment. For us, this boils down to a reminder that our definition of causal strength can only be as good as the prior specification of causes using which it is applied. For example, if C is smoking and E is lung cancer, it may still be that C's causal strength is different with respect to one person than with respect to another, perhaps because the two individuals differ with respect to their genetic predispositions or diets. Thus smoking will have a great many different causal strengths with respect to lung cancer, depending on the exact state of all other causally relevant factors, and it is important that these different strengths are not conflated.

Second, we are interested in the impact on our effect of changing a causal input. But of course, in general changing a causal input will likely also have an impact on other aspects of the world besides our particular effect. Thus, strictly speaking, 'W' in our formulas will be different in each term, which we shall recognise in our notation. Note though that since these differences are precisely those that do not impact on the effect of interest, no Simpson's paradoxes crop up and the force of our definition is unaffected.

*Final definitions*

Putting together the resolutions of all these wrinkles then, we can give final versions of our two definitions of causal strength. Let C be the cause of interest, D be another cause, and E be the effect of interest. Let C0 be the neutral level of C and D0 be the neutral level of D. Then let W1 be the background conditions given C but just excluding C (i.e. the state of the rest of the world given that C obtains, but not including C itself), and given D0 but just excluding D0 - in other words, W1 is the state of the rest of the world when C is 'activated' from its neutral level and D is not. Similarly, let W0 be the background conditions given C0 and D0 but just excluding C0 and D0, and let W2 be the background conditions given D and C0 but just excluding D and C0. Then, if C is an event and E is a variable:

1) the PM of C = E(C&W1&D0) - E(C0&W0&D0); and
2) the DM of C with respect to a second cause D = E(C&W1&D0) - E(D&W2&C0)

Analogous definitions hold as we vary between events and variables. For instance, if E is still a variable and now so is C, then we should re-label so that C1 is the actual level of causal variable and C0 the neutral level. This yields:

1) the PM of C1 = E(C1&W1&D0) - E(C0&W0&D0); and
2) the DM of C1 with respect to a different level C2 of C = E(C1&W1&D0) - E(C2&W4&D0) - for W4 defined appropriately; and:
   the DM of C1 with respect to a level D1 of a second cause D = E(C1&W1&D0) - E(D1&W2&C0).

If E is an event, we can rewrite these formulas as probability functions instead, in the way described earlier. Other variations, such as DM defined with respect to weighted averages over a range of counterfactuals, or cases where some of the causes are events and some variables in different combinations, can also be expressed in a similar way.

It follows from our definitions above that causal strengths are, so to speak, highly sensitive. They will vary with choice of cause C, of course. They will also vary with choice of effect term E - the same thing may be a strong cause of one effect, but a weak one of another. Again, this is obvious. Causal strengths will also vary with changes in background conditions W. For instance, striking a match will cause light if the atmosphere contains sufficient oxygen, but not otherwise. As noted earlier, the specification of W is also our way of capturing the sensitivity of causal strengths to all causally relevant factors. Finally, the score for causal strength will of course be relative to the right-hand term in the formula. In the case of PM, this means relative to the choice of neutral level - although note again that, once given a specification of C, E and W, this

choice is determined objectively. In the case of DM, this means that the causal strength will be relative to choice of counterfactual, and the latter presumably will be interest-relative.

Obviously, such a definition of causal potency is hardly particularly original. As [Sober et al 1992] points out, complications arise once we try to use it to *compare* causal potencies. However, I do not think that these complications turn out to be at all disturbing [Northcott 2003a].

Clearly, our definitions can yield negative as well as positive values for the causal strengths, but I do not see this as being particularly problematic. In a similar way, there is no objection to allowing 'negative causation' generally, that is - in [Humphreys 1990]'s terminology - to acknowledging counteracting as well as contributing causes.

The precise choice of unit of effect will tend not to be crucial here, since our aim is to compare the impacts of different causes on a common effect E. So long as our units are the same for each calculation, these *comparisons* of impact will typically be independent of the precise choice of unit. For example, the mass displaced by one cause would still be twice as much (say) as that displaced by another, regardless of whether that mass were measured in ounces, grams or tons. We can therefore happily define our causal strengths in whatever units scientists themselves naturally use for E anyway.

Strictly speaking, however, this unit-independence will not quite hold always. In particular, if one choice of unit is *non-linear* with respect to the other, problems can arise. For example, suppose our competing units of mass were grams and logarithm-of-grams. Then if one cause was assessed to be twice as potent as another using grams displaced, it would in general *not* be assessed twice as potent using logarithm-of-grams displaced. It may well be that in practice such cases are rare, that is to say that controversies about assignments of relative causal strength only rarely if ever hinge on choice of unit in this way. Nevertheless, there seems no way to rule out the possibility in principle and accordingly if such a case did arise we would indeed be forced to concede that the comparison of causal strengths was choice-of-unit-dependent. This issue would become rather more serious if choice of unit determined also the *qualitative* ranking of causal strengths. This recalls the classic Miller problem concerning the language-dependence of approximate truth. But as with that, so long as no units actually in scientific use generate such ranking reversals then so long can the issue legitimately be disregarded [Northcott 2003c].

*PM and DM unified*

Recall again our two definitions (for when all causes are events and the effect is a variable):

1) the PM of C = E(C&W1&D0) - E(C0&W0&D0); and
2) the DM of C with respect to a second cause D = E(C&W1&D0) - E(D&W2&C0)

Notice that, although there is in it no ambiguity or choice about which counterfactual to consider, nevertheless the definition of PM is still in form similar to that of DM - how much difference does the cause make compared to its not being there at all? Strictly speaking, the only difference is that with PM we set the choice of counterfactual 'D' to be D0. This enables us to see now how our two senses of causal strength can be *unified*. In

particular, the key insight is to see a DM as always being just the difference between two PMs. Alternatively put, if we take PM to be our core definition of causal strength, then a DM can always be seen as a *relative* causal strength. It would follow that the two senses are analytically unified. Analogously, the existence of both relative and absolute *speed* does not imply that there are really two distinct senses of 'speed'. In the same way as we need only one definition of speed, so, I shall argue, we need only one of causal strength.

The key fact is that, formally, the DM of C with respect to D is always just the PM of C minus the PM of D:

DM of C with respect to D
= E(C&W1&D0) - E(D&W2&C0)
= [E(C&W1&D0) - E(W0&C0&D0)] - [E(D&W2&C0) - E(W0&C0&D0)]
= [PM of C] - [PM of D].

Thus any DM can always be re-expressed in terms of two PMs. For example, the DM of Holmes's shot with respect to Watson's is equal just to the PM of Holmes's shot minus the PM of Watson's. Note that the two PMs are each defined with respect to the *absence* of the other cause. The DM of Holmes's shot given that Watson had fired relative to Watson's shot given that Holmes had fired, would be trivially zero since we would just be comparing E(C&W3&D) with itself. Hence the only DM we would be interested in here is that of Holmes's shot given that Watson had *not* fired relative to Watson's given that Holmes had *not* fired, which implies the comparison of PMs as per our calculation above.

Note that our formula specifically picks out C and D. Why pick out them rather than any of the other relevant causes in the actual world? Purely pragmatic reasons suffice - we happened to be interested in the PMs of C and D, and in particular in the DM of C with respect to D. The latter implies that we are interested in particular counterfactuals defined in terms of C and D rather than any other causes. That is, our choice of DM of interest automatically picks out particular counterfactuals, and the interesting bit is that those counterfactuals also happen to be the ones involved when defining the PMs of those causes. Hence, a DM can indeed always be expressed as the difference of two PMs.

It is also worth noting that, formally, a PM can in turn always be understood as a limiting case of DM:

PM of C
= E(C&W1&D0) - E(C0&W0&D0)
= [E(C&W1&D0) - E(C0&W0&D0)] - [E(C0&W0&D0) - E(C0&W0&D0)]
= [PM of C] - [PM of no cause]
= DM of C with respect to no cause, by our previous result.

Thus the PM of some C can be described as the PM of C minus the PM of no cause at all, in other words described as a DM. This immediately suggests that perhaps, instead of speaking of DM being cashed out in terms of PM, we could equally well think of things the other way round - namely, of it always being possible to cash out PM in terms of DM. Why should it be PM that is taken to be the more fundamental? We shall argue later (section 3) that there does exist some reason for preferring PM in this regard, but in any

case, for many purposes it will not matter whether we talk in terms of PM or of DM. The important point is that, expressed whichever way, there is only one independent notion in play here. Despite initial appearances to the contrary in the Holmes-Moriarty example, there do not exist two kinds of causal strength; rather, there is really only the one.

*Two examples*

Imagine a Newtonian particle with a gravitational force on it. In this example, the DM/PM distinction can sometimes seem to collapse; let us see why. What is the potency of gravity here? The PM of C where C = gravity and C0 = no gravity would be yielded by:

PM of C = E(C&W1) - E(C0&W0)
= (the particle's motion with gravity) - (the particle's motion with no gravity).

Next, suppose we ask how much difference does gravity make? It is often natural to interpret this as being relative to no gravity at all, in which case the DM of gravity would be:

DM of C relative to the counterfactual C0 = E(C&W1) - E(C0&W0)
= (the particle's motion with gravity) - (the particle's motion with no gravity)

In other words, here DM and PM are exactly the same.

The two could have diverged if in the DM calculation we had adopted a different choice of counterfactual. Suppose we were comparing the strength of gravity on Earth with that on the moon. Let the Earth's gravity be C1 and the moon's gravity C2, so that the counterfactual C2 was now some lesser but non-zero alternative level of gravity, corresponding to its strength on the moon. Now the calculation would run:

DM of Earth's gravity C1 relative to the moon's gravity C2
= E(C1&W1) - E(C2&W4)
= (the particle's motion with Earth gravity) - (the particle's motion with moon gravity)
= [E(C1&W1) - E(C0&W0)] - [E(C2&W4) - E(C0&W0)]
= [PM of Earth's gravity] - [PM of moon's gravity].

There are two different questions here: 'how much difference does the Earth's gravity make compared to some other level of gravity?', and 'how much difference does the Earth's gravity make compared to no gravity at all?' The difference between the questions is entirely down to choice of counterfactual - either moon or zero. Often, as in the way we originally set the example up, the implicit choice of counterfactual will be zero anyway, in which case DM and PM will coincide and there will not be even the appearance of ambiguity. Perhaps this is why the issue of causal strength seems so unproblematic in the Newtonian particle case, and indeed in many everyday contexts too.

Turn now to a second example (like the first one, adapted from [Sober 1988]) - do genes or environment have the most causal impact on the height of an individual corn plant? On one view, the situation is more problematic here because both genes and

environment are necessary inputs for a plant to achieve any height at all, which makes it seem impossible to assign either factor a greater importance than the other. But there is a second way of looking at it too. Suppose that we have a traditional genetic input and the option of a new one (a new plant breed, say), and similarly a traditional and new environmental input too (a new fertiliser, say). Suppose further that switching to the new plant breed increases average plant height by 2cm, but that switching to the new fertiliser increases it by 5cm. There is now a clear sense in which, with respect to these particular options, switching the environmental input has more causal potency than does switching the genetic one. So overall we have something of a paradox: on the one hand, the causal strengths of genes and environment seem inextricably intertwined and therefore necessarily equal; on the other, it seems that one of them can be seen as more important than the other after all.

First, our DM can capture this second sense. How much difference does the new fertiliser make relative to the old one? This is given by the subtraction E(new fertiliser) - E(old fertiliser), which is just a DM. Likewise, we can calculate a DM for the new plant breed, and then compare the two DMs to see which is bigger and hence which factor more important.

What of the first sense of causal strength here, according to which genes and environment must be adjudged equal? I claim that this sense is the one captured by our other notion, PM. Now biologists standardly say that to speak of causal potencies here is meaningless, and [Sober 1988] agrees with them. But let us work through our definition. For C = genes and D = environment, with neutral levels C0 and D0, with W3 defined appropriately and with the other notation for W as before:

PM of C = E(C&W3&D) - E(C0&W2&D)
= (corn plant's height with both environment and genes) - (corn plant's height with environment but no genetic input)
= (corn plant's actual height) - 0
= corn plant's actual height.

And for environment, we get an exactly analogous calculation:

PM of D = E(D&W3&C) - E(D0&W1&C)
= (corn plant's height with both genes and environment) - (corn plant's height with genes but no environmental input)
= (corn plant's actual height) - 0
= corn plant's actual height

(We assume in the calculation that each PM is calculated with respect to the presence of the other input. If the other input was instead absent, then each of genes and environment would have been awarded zero potency. But either way, our basic point here would still hold.) So our definition implies that: (1) genes and environment *each* has maximum causal potency here - the plant's height goes from zero to full with their (individual) presence compared to their absence. (2) Therefore each has the *same* degree of PM, as desired.

Intuitively, conclusion (2) seems fine. I would also defend conclusion (1) - intuitions

that each of genes and environment could only have perhaps a causal potency of 'a half' reflect, I suspect, an intuition that the total potencies of two inputs should not add up to more than the total effect. But such an intuition would be misplaced here. These potencies are being calculated individually, i.e. for each input while assuming the other input is already in place. Were we to calculate the 'joint potency', i.e. where C = (genes & environment) in our PM formula, then the joint potency would again just be the plant's actual height, and no more than that. So under no circumstances is any PM ever calculated to be more than the total effect. If there are many jointly necessary causes, then it is surely no weakness of our scheme if any one of those causes when taken individually is found to have maximal potency - *given* that all the other causes are already present. (See also section 2 for discussion of this point.)

So the question of how much contribution each of genes and environment made, is now well-defined. And although the answer we get may be trivial, it seems to me that, contrary to biological orthodoxy, it is nevertheless certainly not meaningless. The simplicity of the issue in the Newtonian particle case compared with the apparent dichotomy of senses of causal strength in the biological one, leads [Sober 1988] to conclude that 'there is no such thing as the way science apportions causal responsibility; rather, we must see how different sciences understand this problem differently, and why they do so' (p304). But I think that *both* cases can be analysed using our same DM/PM framework. Therefore, given that DM and PM in fact boil down just to the single underlying notion, the suggestion is that a unified understanding across science of the notion of causal strength *is* possible after all.

Part of the confusion here stems from the fact that the second sense of causal strength in the biology example is usually analysed using the statistical technique of ANOVA rather than using our DM formulation. My own view is that the use of ANOVA to calculate causal strengths in this way, here and elsewhere, is both unnecessary and mistaken [Northcott 2003b]. But the important point for our purposes is merely that the DM/PM formulation can indeed be applied successfully to these apparently difficult cases.

*Causal strengths in group problems*

So far, we have defined causal strength only for singleton cases, as it were 'individual-PM' and 'individual-DM' for individual plants and individual Newtonian particles. But suppose, for example, we were interested not in whether an individual plant's height was due more to genes than to environment, i.e. not in the singleton PM and DM. Rather, suppose we wanted to know instead which of genes and environment was the more important cause of height across a whole *population* of plants - wanted to know, as it were, the 'group-PM' and 'group-DM'? Consider a new example, comparing smoking and asbestos. We could ask either which of the two factors is the more important carcinogen with respect to a particular individual person, or ask instead which is the more important across a society as a whole. The question is, can we unproblematically scale up our singleton definitions for use in these group cases?

As [Sober et al. 1992] shows, causal strengths across a population can be straightforwardly calculated as the sum of the potencies of each individual level of exposure, weighted for the frequency of that level of exposure in the population. For example, if the PM of smoking 10 cigarettes per day for 30 years is X, and if 10 per cent

of the population smoke this much, then this would contribute 0.1X to the eventual sum. And if the PM of smoking 20 cigarettes per day for 30 years is Y, and if 15 per cent of the population smoke this much, then this would contribute 0.15Y to the eventual sum. We could then sum in this way over all the levels of smoking present in society, yielding an eventual grand total, 'Z' say. This Z would be the 'group-PM'. It represents the quantity of cancer caused in this group or population by smoking, compared to if there had been no smoking at all. ([Sober et al 1992] calls this group-PM 'distribution-dependent' causation and contrasts it with 'potency', by which it means what we have been calling individual-PM.)

An alternative, perhaps simpler, way to view the calculation is as follows. First, take the grand total of the expected number of cancers in the population. Next, take the expected number of cancers in the population in the counterfactual case of there being a zero level of smoking. Finally, simply subtract this second number from the first, thus yielding the group-potency. This would again simply be a direct application of our potency formula - namely, the expected value of effect with the cause, minus the expected value without it.

Of course, the first method of calculation above eventually yields exactly the same answer as the second. The first method does carry the advantage of defining group-PM purely in terms of already-defined individual-PMs, but our second method will perhaps make it easier to see how to extend the analysis to the case of group-DM. Either way, we eventually arrive at the following definition of group-potency (notation as before):

> Group-PM of C = E(C&W1) - E(C0&W0), where E is the total effect across the group

What about group-DM? Suppose we wanted to calculate the group-DM of smoking. As with an individual-DM, this group-DM would only even be defined with respect to some choice of counterfactual. Suppose we select the counterfactual whereby everyone in the population who actually smokes cigarettes in fact smoked a pipe instead. We could then calculate the expected number of cancers in this latter case and subtract it from the current number of cancers, thus yielding the group-DM: how much difference does smoking cigarettes make to the number of cancers, relative to smoking pipes instead? Put generally and formally:

> Group-DM of C relative to the counterfactual D = E(C&W1&D0) - E(D&W2&C0), where E is the total effect across the group

How much difference smoking cigarettes made relative to a zero alternative level of smoking would just be its (group-) PM again. Thus the connection between PM and DM at the group level is exactly analogous to that we already saw at the individual level.

The key point is that our singleton and group definitions of PM and DM are identical, save for the interpretation of the effect term. This is important because it may often be somewhat arbitrary whether we describe a problem as being group or singleton in the first place. For example, suppose we were comparing Britain's crime rate with that of the USA, and suppose (counterfactually) that the only pertinent difference between the two countries was their gun laws. Then we could get a DM of the American gun laws relative

to the British ones just by comparing the two countries' crime rates. Now we could also compare crime rate figures for many different countries, all assumed to have either American or British gun laws, and see how the figures varied across them all. That would enable us to work out a global group-DM for the causal strength with respect to crime rates of American-style relative to British-style gun laws, averaged across the 'population' of all these different countries. In this light, our original calculation comparing just Britain and the USA's figures alone, would be an individual-DM - it is worked out just with respect to a single datum point, so to speak, rather than with respect to the average across the whole population of different national crime rates.

Now suppose instead that we were investigating the effect of gun laws on the criminality of just a single individual, comparing the cases of if he had lived in Britain with if he had lived in the USA. This would yield the DM for American relative to British gun law for that one person. Alternatively, we could have taken the average crime rates across the whole populations of Britain and USA, and seen how they compared instead. It would be natural to take the latter calculation as yielding a group-DM, and the former an individual-DM. In other words, the *same* DM calculation - comparing Britain's with USA's crime rate - was described as an individual-DM in the context of the previous paragraph's global study, but now as a group-DM in the context here of a study of an individual person. It is exactly the same calculation each time; the only thing that has changed is our *description* of it as either 'singleton' or 'group'. So it is clearly a desirable thing that our definition would yield the same score for the DM each time, and hence is robust with respect to such arbitrary re-descriptions. (In passing, it one of the demerits of the ANOVA technique that it does not meet this requirement [Northcott 2003b].)

*Desiderata revisited*

1) *Univocal*. This was the whole thrust of our demonstration that DM and PM boil down to one and the same underlying notion.
2) *Objective*. As stated earlier, once given a specification of our focus of interest, to whit a specification of our cause, effect and background conditions, our definition of causal strength proceeds objectively.
3) *Quantitative*. Our definition is quantitative.
4) *Applicable*. For causal interaction, see section 2 in a moment. Generally, we have shown that the same approach can be applied to singleton and group cases, and also to cases from biology often claimed to be awkward. We show later (section 3) that using it we can non-trivially compare the strengths of any two causes of a given effect, no matter how apparently incommensurable those causes be.
5) *Normative*. Since we define causal strength to be the quantity of effect caused, and since quantity of effect is an empirical variable, it follows that our definition's scores for causal strength are validated empirically. For example, to say that (in a particular circumstance) a fertiliser has a causal strength of 5cm with respect to the height of a plant, implies that the addition of the fertiliser actually *would* increase that plant's height by 5cm. It follows that our scores for causal strength therefore do have instrumental normative force.

# 2 Causal interaction

*Introductory example*
Suppose we have two fertilisers, Green and Blue, and we are interested in the effects they have on a plant's height. Suppose further that the following table summarises these effects:

*Table - Plant heights and interacting fertilisers*

| | Blue fertiliser | Nothing |
|---|---|---|
| Green fertiliser | 14 | 4 |
| Nothing | 2 | 0 |

In words, Blue on its own scores 2 points, Green on its own scores 4, but Blue and Green together leads to a big positive interactive effect and a large score of 14. So what is the causal strength here of, say, Green? Intuitively, the issue seems confusing because it is not clear how - or whether - to incorporate the big interactive effect with Blue. Is Green's potency 4, or 12, or 4 plus an interaction of 8, or perhaps 4 plus some share of an interaction of 8, or maybe the question is unanswerable or meaningless?

The analysis of this paper indicates the following straightforward resolution: there are in fact *two* potencies involved here for Green, depending on whether or not the background conditions W are specified to include the presence of Blue. Recall that if W changes, then so in general does the potency - a match is a potent cause of fire if W includes sufficient oxygen in the atmosphere, but not otherwise. In this example, the notion that there is a single potency for Green is therefore in effect a misplaced attempt to define a single potency across several relevantly different circumstances. (For the wrong-headedness of this in general, see [Northcott 2003a].) Rather, we should say just that Green has a potency of 12 when Blue is present and a potency of 4 when it is not. There is no need, we shall argue, to invoke an 'interactive strength' separate from those of the two inputs, or to declare the issue intractably confused. And the context-specificity of the very notion of causal strength means also there is no need to be embarrassed by Green being awarded more than one value for its potency. Two different contexts means two different potencies, just as with lighting the match.

Suppose we had had a population of ten plants, five treated with Blue fertiliser and five not. In half the cases therefore, Green would have been awarded a potency of 12, and in the other half only 4. The total potency across this population would have been (5 x 12) + (5 x 4) = 80, or an average of 8. It makes perfect sense to define a group-potency specific to some population like this. Of course, if we changed the weightings so that now more than half of the plants were treated with Blue, then we would reach some new average potency figure for Green. Hence the result for the group-potency is specific to choice of population, exactly analogously to how individual potency was specific to choice of Blue or un-Blue W.

Therefore we may define Green's potency relative to a Blue plant, relative to an un-Blue plant, or relative to a specified mixture of Blue and un-Blue plants. All of these would correspond to different choices of W in our basic formula. What we cannot meaningfully do is try to define some potency for Green independent of any specification

of W at all. For this reason, our table of results cannot be summarised completely by just a single potency score.

A further worry is that the potency of Green in the presence of Blue is 12, and of Blue in the presence of Green 10, yet the total effect is only 14. So the sum of the two individual potencies is more than the total effect. Can this possibly be right? Consider the *joint* potency of C = Green and Blue. By our formula, this is:

E(C&W1) - E(C0&W0) = E(Green&Blue) - E(neither Green or Blue) = 14 - 0 = 14

In other words, the joint potency is just the total effect here. So no potency is ever calculated to be *more* than the total effect, and there is no paradox (just as in the biological example from section 2).

In essence, this last point is no more controversial than the following: if we have a box but no match, then the addition of a match means we shall have a light where before there was none. Equally, if we have a match but no box, then this time the addition of the *box* leads to light where before there was none. Therefore the box and the match *each* individually lead to a light where before there was none. But presumably no one would take this therefore to be paradoxical, on the grounds that now the match and box must together somehow be assumed to lead to *two* lights.

*Causal composition and black boxes*

A recurrent challenge when working with causes is to find out their laws of composition. In order to be useful for prediction, knowledge of the various causes at play is not enough on its own; we also need to know how they interact, or compose, with each other. We cannot simply assume Mill-like additive composition. When defining the strength of interacting causes, the same issue crops up. As we have just seen, knowing the potencies of the Green and Blue fertilisers acting on their own was not sufficient for us to know their potency once they were interacting with each other. On their own, Green had a potency of 4 and Blue of 2, but acting together they had a joint potency of 14.

Our emphasis on context-specificity now delivers another dividend - we no longer need to know any general *laws* of causal composition. Rather, we need know only how causes actually composed in the particular context we are concerned with. For example, suppose the Green and Blue fertilisers combine to give a score of 14 only when it is normal weather. When the weather is hot they combine for a score of 18, except if the weather is also unusually wet in which case they compose additively and yield a score of only 6. Any ambition to define the causal strengths of the two fertilisers in some kind of general, context-independent way would among other things need to be aware of this full pattern of causal composition. By contrast, defining it only case by case as we have done, means that there is no such necessity for all these details. All we needed to know in our example was how Green and Blue actually composed in the specific case we were interested in. Here (let us assume) there was normal weather and so they composed to produce a joint effect of 14. This is all the information we need. The way the two causes might compose in counterfactual hot or wet conditions is irrelevant to defining their causal strengths in normal conditions. In this sense, it is therefore much easier to define causal strengths only context-specifically, since much less information is required. Once given the background conditions, for any given cause C all we need to know is the value

of the effect E in its presence compared with the value of E when it is absent.

There is a further sense in which our definition of causal strength makes life easier - all we need consider is the final value of the effect term. We have taken no notice of the underlying causal mechanism that is producing this effect, for instance whatever may be happening at the molecular level as Green and Blue both impact on the plant at the same time. We leave the causal mechanism a black box, as it were, and take note only of the final result. This may indeed make life easier, but does it come at an unacceptable cost? Consider now an example that might be taken to suggest that it does, in order to see why in fact it does not.

Suppose that in place of fertilisers and plants, we talk instead of workers, managers and production of widgets. Imagine that the worker on his (or her) own can only achieve an output of 2 widgets, as without the manager to provide the necessary final authorisation most of the worker's widget-building labour is left unexploited or incomplete. Imagine next that the manager on his (or her) own achieves an output of 4, since now the necessary authorisations can be made and even without the worker the manager is able to do a little labour himself. Imagine finally that both the worker and manager are present: now the total output will be 14 widgets, since the worker can produce much more labour than the manager was able to, and the manager can dispense all the necessary authorisations so that none of the worker's efforts go to waste. There is thus a positive interactive effect.

The payoff structure here is of course deliberately identical to that of the fertilisers: namely, 2, 4 and 14. It follows that, by our definition, so are the potencies identical too. Thus, the manager has a potency of 4 on his own, but 12 if the worker is there too; similarly, the worker, like the Blue fertiliser earlier, is awarded potencies of 2 and 10. But, the objection runs, are these values really intuitively satisfactory? Reasons why they might not be can, I think, be crystallised into two slightly different objections. First, it might be argued that, as it were, it is the worker who is really doing most of the work and our potency scores do not reflect this. Second, our formula incorporates the full interactive effects into the potency scores for individual inputs - is this acceptable?

Begin with the first point. Our worker is doing all the hard labour while the manager is merely signing some forms; surely it would be more just, therefore, if the worker received the lion's share of credit for the final output? This objection is only made possible by our knowledge of the details of the causal mechanism underlying the final results. If the objection held up, therefore, it would also be a strong argument against our strategy above of leaving - for the purposes of defining causal strength - these causal mechanisms as black boxes.

But I think the point in fact boils down merely to confusion about the explanandum (a danger also emphasised in [Sober et al. 1992]). Our effect E here was the final output of widgets. For that effect, the worker on his own was indeed unable to produce many widgets, whereas - once the worker was in place - the introduction of the form-signing manager did indeed have a dramatic impact on final widget output. Accordingly, it is desirable that our formula captures this dramatic impact. I suspect the objection here is really more a moral one, and is motivated by the sense that the worker is putting in a lot more labour and physical effort than is the manager, and that this should be recognised. Maybe so, but in that case we should re-specify our effect E to be something like hours of effort or litres of sweat, rather than final output of widgets. Or maybe instead widgets

that are built but unauthorised should be awarded a score of 0.8 of a unit or some such. All these modifications of E would indeed tend to yield higher potency scores for the worker and lower ones for the manager. The point here is that our controversy turns out only to lie in the specification of E, and none of the foregoing constitutes a criticism of our definition of causal strength itself. (Perhaps our contrast here between different specifications of E is an example of the classic one in economics between the labour and exchange theories of value, or more generally is an example of the divergence between moral and economic accounting.)

Of course, none of this is meant to deny that knowledge of underlying causal mechanisms may often be extremely useful, perhaps essential, methodologically. A necessary condition for applying our definition of causal strength in the first place was that we agree on the causal ontology, and presumably knowledge of causal mechanisms is likely to be more than useful for achieving that. So we are not making any claims here about instrumentalism in general. Rather, our claim is much narrower - merely that, once we are agreed on our ontology, then *for the specific purpose of defining causal strength* we need not worry about the underlying mechanism.

*Why we should include interactive effects*

Turn now to the second objection. The worker on his own produced 2 but in combination with the manager produced 14, an increase of 12. When defining the potency of the manager in the presence of the worker, this entire increase of 12 is, as it were, credited to the manager's account. Yet should not the credit for the extra 12 instead be shared between the two? Or perhaps fenced off separately, described as an 'interaction effect', and sharply distinguished from either the worker's or manager's *individual* potencies? An analogous point could be made with respect to the fertiliser example. It might seem that our approach here is counterintuitive. After all, why should the causal strength of the Green fertiliser include Green's interactive effects with Blue? Ought not a cause's potency to include only the 'pure' impact of it and it alone - independent, so to speak, of any help accruing from the co-operation of other causes?

But I think that this sentiment is mistaken. First, on a purely intuitive level, it might equally well be argued back that a definition of causal strength *should* include interactive effects. After all, were it not for the addition of the Green input there would not have been any of these interactive effects in the first place, so surely Green has some responsibility for them. As it were, adding Green fertiliser not only leads to a new quantity of Green effect, it also leads to a new quantity of interactive effect as well. So maybe, with respect to Green's potency, it is no more obvious that interactive effects should be excluded than that they should be included.

Perhaps the real lesson here is that arguing purely at the intuitive level can only get us so far. Instead, we should return to more systematic analysis. In particular, our main argument will be, as it were, the larger picture: namely, that the only coherent logic for defining causal strength as a whole implies that interactive effects *should* be included, just as they are by our definition. We shall argue for this claim by discussing at length a new example.

Consider this question: 'how much difference did Hitler make to the course of history?' This is surely unanswerable without some consideration of how his personality interacted with the political and social conditions that happened to be in place in 1930s

Germany and Europe. More carefully, if comparing the impact of Hitler's personality H with that of the bad social conditions S in Germany, say, we can imagine counterfactual neutral levels of a 'normal' politician H0 of non-genocidal tendencies and of 'normal' social conditions S0 of greater social harmony and economic stability. In the normal social conditions S0, Hitler would presumably have likely had far less success and hence impact. Likewise, with normal politicians H0 even bad social conditions fortunately only rarely lead to world wars and holocausts. So it was only the tragic *interaction* of Hitler with the abnormal social conditions - of H with S - that led to such a huge effect in actuality. In this context, it seems bizarre to rule out interactive effects when assessing causal strength, on pain of saying that Hitler made no difference to the course of history. At any rate, we shall begin by assuming that they are indeed incorporated into causal strengths in the manner our definitions prescribe, and see how we fare.

Suppose first a historian objected that Hitler was not that personally important, and that most of the causal responsibility should instead be allotted to the social conditions. Perhaps so, perhaps not; but the important thing from our point of view is that such a dispute would be over historical facts, not over the deeper issue of our definition of causal strength. In particular, the dispute would centre over the values to award to the following two expressions. Thus, the potency of Hitler's personality is given by our usual formula for PM (with our usual notation for W):

E(H&W3&S) - E(H0&W2&S)
= E(Hitler & bad conditions) - E(normal politician & bad conditions)

And similarly, the potency of the bad social conditions is:

E(S&W3&H) - E(S0&W1&H)
= E(Hitler & bad conditions) - E(Hitler & normal conditions)

So would the historian be right to say that bad social conditions were more important than Hitler? Comparing the two expressions for potency, clearly the critical thing is the relative value of the two counterfactuals E(normal politician & bad conditions) and E(Hitler & normal conditions). That is, the dispute is over the relative sizes of E(H0&W2&S) and E(S0&W1&H). And this is a substantive historical - not definitional - dispute about causal strengths.

Consider next a second historian, who accepts that Hitler and social conditions were both relevant but who this time wishes to stress that what was most important was not either cause individually but rather the two factors' interaction. Can we represent this claim within our system? Again, I think 'yes'. For clarity, assume that the following familiar table of effects holds (in some arbitrary units):

*Table: bad effects of Hitler and social conditions*

| | Bad social conditions S | Normal social conditions S0 |
|---|---|---|
| Hitler H | 14 | 2 |
| Normal politician H0 | 4 | 0 |

The interactive effect of two causes is the total effect not explained by the sum of those causes' individual effects. Starting from the bottom-right-hand case H0&S0 where both causes are at 'normal' levels, we can see that: alone, Hitler makes a difference of 2 points; and alone, bad social conditions make a difference of 4 points. It follows that, if additive causal composition held, Hitler and bad social conditions should together make a joint difference of 2 + 4 = 6 points. However, the table shows that in fact they make a joint difference of 14 points, not 6. Hence there is a positive interactive effect of an extra 8 points. The claim now is that our definition can capture this reasoning, by which I mean that we can represent the interactive effect separately even though in our system interactive effects are subsumed into the definition of a cause's individual strength.

In essence, the interactive effect is given by comparing the sum of how much difference each cause made individually, with how much difference they made jointly. But of course these quantities are just what we were representing above with our causal potencies. Thus we can rephrase our definition of the interactive effect to read: the two causes' joint potency minus the sum of their two individual potencies. Running this through our numerical example:

*interactive effect between H and S*

= potency of H&S - (potency of H in the absence of S + potency of S in the absence of H)

= [E(H&S&W3) - E(H0&S0&W0)] - { [E(H&W1&S0) - E(H0&W0&S0)] + [E(S&W2&H0) - E(S0&W0&H0)] }

= E(H&S&W3) - E(H&W1&S0) - E(S&W2&H0) + E(S0&W0&H0)

= 14 - 2 - 4 + 0

= 8, as desired.

Therefore, using just our definition, the size of the interactive effect can be represented straightforwardly. It follows that the historian's claim as to the relative importance of interactive versus individual effects can be also be represented in terms of our definition straightforwardly. And at that point, such a claim can again be said to turn only on substantive historical facts and not on philosophical definitions.

Note that we needed to be careful during the calculation since the 'potency of Hitler' could have been referring to either of two different quantities, depending on our specification of W. In particular, the potency of Hitler is 2 points in normal social conditions but 10 points in bad social conditions. That is, the causal strength of Hitler varies with context - which should come as no surprise by now. But from the point of view of calculating the size of an interactive effect between two specified causes, this ambiguity was not a problem since it was clear which choices of W were appropriate here and we merely had to be careful to select the right one.

So much for our strategy of including interactive effects. Now consider the alternative - is it possible to analyse these issues satisfactorily while *excluding* interactive effects from a definition of causal strength? On this view, presumably we should, when defining Hitler's potency, exclude all interaction between Hitler's personality and the bad social conditions. It is not clear to me exactly what in that case we should take Hitler's potency to be instead. One candidate might be the 'pure' effect of 2 points (down the right-hand column of our table):

E(Hitler & normal conditions) - E(normal politician & normal conditions)
= E(H&W1&S0) - E(H0&W0&S0)

But this is just the causal strength of Hitler in normal social conditions, and surely we are interested really in his causal strength in the *ab*normal social conditions that actually obtained in 1930s Germany. So this suggestion seems clearly irrelevant to the problem at hand.

Perhaps an alternative would be to assign to Hitler's personal potency some partial *share* of the interactive effect of 8 points. But unfortunately this leads to several undesirable consequences. First, we risk intractable counting controversies, since how do we decide how many causes should get a share of this interactive surplus, so to speak? As well as Hitler and the social conditions, there are many other factors in the background W which will also have been causally (and interactively) relevant. Since the designation of foreground rather than background is arbitrary here, so presumably these background causes should be equally as entitled to the interactive surplus as our two foregrounded ones. But is it really possible to give a canonical categorisation of every single background cause? (See below for more discussion of the foreground/background distinction.) Moreover, this approach would also give unsatisfactory results in everyday examples. For instance, we could not say that the strength of my kicking a ball is given by the ball's resultant acceleration; rather, we could say only that the strength of my kick was some *share* of that acceleration, the rest of the credit being distributed around the many various interacting background causes like air pressure and temperature, the rules of football, the fact that I played that day, and so on. In other words, we would lose our basic intuitive notion of the strength of a cause being the size of effect it leads to. Finally, in all these situations only one cause ever actually *changes*, namely the foregrounded cause of interest. So even if achieved only via interaction with other causes, a change in effect can non-arbitrarily be credited just to that one changing cause. Together, these points seem to me decisively to tell against the strategy of sharing interactive surpluses among several causes.

So the program of excluding interactive effects from the definition of a cause's potency seems in practice to be impossible to get off the ground. But turn now to perhaps the intuition that might be motivating such a program in the first place, as arguably it is more appealing in the abstract than when any attempt is made actually to operationalise it. It seems to me that that intuition may be that we want somehow to establish a *general* result for Hitler's personal potency, independent of the specific historical circumstances in which he actually operated. What is his 'pure' personal effect, independent of interaction with 1930s social conditions? How dangerous would a Hitler be in general? Was it just a fluke that he had such a large impact on the course of history? It is important to be clear immediately that this is a *different* question from the one about to what extent Hitler personally was causally responsible for the actual historical events of the Nazi period. The latter question is enquiring about specific causal strengths that actually obtained in real life. The former, new, question concerns instead a *counterfactual* musing - what *would* some causal strengths have been?

In response, first, there is therefore no argument here against our definitions as applied to the actual historical case since it turns out that the particular objection in fact

concerns other, counterfactual, cases. In other words, there is no argument here against including interactive effects in the definition of Hitler's potency for the actual historical case of 1930s Germany. The interest in Hitler's 'general' potency is distinct from this specific historical question.

Second, even when we do go through what Hitler's 'general' potency might be, I think that our same basic point still stands there too. For this general case, we would of course have to specify a particular population of possibilities in order to apply our definition at all. One such definition of Hitler's personal potency could be:

E(Hitler & average conditions) - E(normal politician & average conditions)
= E(H&W4&A) - E(H0&W5&A)

(Here, A = average social conditions, and we re-notate W appropriately.) Since the potency of interest is now itself a counterfactual one, it becomes legitimate to take previously irrelevant circumstances - such as a normal politician in normal conditions - into consideration. The idea is that we are taking Hitler's 'general' potency to be his potency in average social conditions. Of course, the result yielded would be critically dependent on just how we specify this 'average'. My own view is that this formulation is in any case a mistake, since while Hitler might be ineffective in average conditions, this would not rule him out still being all too effective in certain unusual conditions. A better formulation would follow instead our definition of group-PM from section 2. That way, we would end up considering:

[E(Hitler & conditions X) - E(normal politician & conditions X)],

where we can imagine many different social conditions X, specify weightings in order to calculate an average across them, and then present the result as the general potency. Of course, this strategy would still be critically dependent on our choice of which set of counterfactual conditions X is included.

Anyway, the whole 'general' question might or might not seem worthy of any serious scholarly attention; nevertheless, if we were to investigate it, this would seem to be the appropriate way to do so. For our purposes, the real point is that even here we would still be including interactive effects in our definition of causal strength. Thus, for each state of social conditions X, we must calculate E(Hitler & conditions X), and in every case the value of the effect E will of course include the results of the interaction between Hitler and X. Therefore even a general, context-independent interest in how much difference Hitler makes would not argue against including interactive effects in the definition of causal strength. In fact, it would argue in favour of the practice.

Finally, note that including interactive effects in this way can, so to speak, also be validated *empirically*. Return to the fertiliser example, and imagine that we are shopping at market and wondering whether or not to buy some Blue fertiliser. To judge whether this purchase was worth it, we would need to know its potency on our plants. This potency though would in turn depend on whether or not we had Green fertiliser already, in exactly the context-specific way already prescribed, and moreover the potency would incorporate interactive effects also in exactly the way prescribed. That is, the advice furnished by our definition of potency is instrumentally normative. Specifically, if we

already have the Green fertiliser then we should only buy the Blue one too if its cost is under 10 units. And if, by contrast, we do *not* already have Green, then Blue would only be a good buy at a cost of under 2 units. Following the prescriptions of anything other than our formulas here would risk us going bust.

*Foreground and background causes*

Throughout, our definitions have distinguished between those causes left among the background assumptions W, and those one or two brought into the foreground for explicit attention. Of course, the decision about which causes to highlight in this way in effect serves to specify the particular causal potencies we are concerned with. In general, this decision will presumably be motivated pragmatically. But it does highlight another sense in which we have surely always been including interactive effects in our intuitive notion of what a cause's strength is *already* - namely, interaction with background causes.

For example, suppose that I work a pump hard while you work it only softly, and that between us we create the effect of a certain degree of air pressure inside the pump. It seems clear that my hard pumping carries a greater causal strength than your softer efforts. Let causal strength here be, as usual:

$$E(C\&W1) - E(C0\&W0),$$

where now E is the air pressure produced, C is my work on the pump, W is the background assumptions and the notation is as before. Suppose that my and your pressings on the pump did not interact, i.e. that they composed additively. Nevertheless, even then there is one further way in which we are incorporating interaction. What I have in mind is that my work on the pump only produced greater air pressure thanks to interaction with several other causes subsumed in our W - such as the rigidity and airtightness of the pump's lining, the ambient air temperature being what it was, and so on. The latter two causes on their own do not create any air pressure (beyond the ambient level represented by E(C0&W0)), but then neither on their own would my efforts without them; it is only the joint action of all three that yields the air pressure effect E(C&W1). This is therefore an interactive effect which is getting included in our definition of the potency of my work on the pump.

So far, this paper has focused only on the interaction between foreground causes, deliberately separated away from the background conditions W, and in particular on whether or not this foreground-foreground interaction should be included in a definition of potency. But of course, ultimately it is just a pragmatic matter which causes we put in the foreground to analyse and which instead we leave in the background. For instance, perhaps we could have taken your working on the pump for granted, put it in our W and concentrated instead on the relative impacts of my work and of the ambient air temperature, contrasting the case of a summer's and winter's day. Which has a greater impact on the air pressure - my working, or switching from winter to summer? All along, it has implicitly been acceptable to include interaction with air temperature when it was a background factor, and uncontroversially so. How then could it suddenly become unacceptable to include it now just because we have suddenly designated it a fellow foreground one? So we need to include interactive effects in any definition of causal strength, on pain of leaving that definition not robust with respect to arbitrary

foreground/background designations.

Related to this, there also arises one further undesirable consequence of excluding interactive effects - namely, that our scores for causal strength would now vary when we do not want them to. To see why, suppose that, besides Hitler's personality and German social conditions, there is now a third factor we wish to consider: say, the weakness F of foreign powers at intervening, relative to some neutral level F0 of more assertive foreign powers. Suppose first we want to compare the potency of Hitler's personality with that of social conditions, just as before, to see which is the stronger. As we saw, in this case Hitler's potency is yielded by:

E(Hitler & bad social conditions) - E(normal politician & bad social conditions)
= E(H&W7&S&F) - E(H0&W6&S&F)

(For clarity, here we take W to be the rest of the world excluding all three of foreign powers, social conditions and politician, hence the inclusion in the formula above of all of these factors explicitly, and the new indices on W.) Next, suppose we now want instead to compare Hitler's potency with that of the foreign powers' weakness, again in order to see which is the stronger. So this time F rather than S is foregrounded, and our formula for Hitler's potency yields:

E(Hitler & weak foreign powers) - E(normal politician & weak foreign powers)
= E(H&W7&F&S) - E(H0&W6&F&S).

Fortunately, the net result is that our formula for Hitler's relative potency is exactly the same in both cases. Therefore the score our formula gives does *not* depend simply on which of the other causes we designate to be foreground and which background. The point is that this is surely how we should want it to be, since presumably the choice of *competing* cause - here whether we are comparing Hitler's importance with that of foreign weakness or instead with that of social conditions - is irrelevant to how strong a cause *Hitler* was.

Now the sting is that if, by contrast, we *exclude* interactive effects, then we no longer get this desirable constancy result. Take the first case above, comparing Hitler's personal potency with that of social conditions. The 'pure' Hitler effect, independent of any interaction with bad social conditions, would presumably be:

E(Hitler & normal social conditions) - E(normal politician & normal social conditions)
= E(H&W9&S0&F) - E(H0&W8&S0&F)

But in the second case, comparing the 'pure' potency of Hitler against weak foreign powers rather than against bad social conditions, the formula for Hitler's potency would now be (for appropriately notated W):

E(Hitler & normal foreign powers) - E(normal politician & normal foreign powers)
= E(H&W11&F0&S) - E(H0&W10&F0&S)

Comparing the two formulas for Hitler's personal potency, we see immediately that

they are now *different*. The first includes normal social conditions but weak foreign powers (i.e. S0&F), the second bad social conditions but normal foreign powers (F0&S). In other words, by excluding interactive effects, our score for Hitler's personal potency is no longer robust with respect to arbitrary foreground/background designations.

The only remedy for this I can imagine would be to stipulate that *all* foreground causes other than Hitler always be set to their 'normal' levels. Here, that would mean to assume in all calculations that foreign powers are normal not weak and social conditions normal not bad. That way, our inconstancy problem for Hitler's personal potency would be solved in this case. But of course, given the arbitrariness of what is designated a foreground or background condition, we would now be faced with the obligation somehow to set *all* causes in W to their 'normal' state. For instance, we would need to identify the 'normal' states of contemporary levels of literacy, of population growth, of human psychology, of planetary distance from the sun… and so on. Such a task would seem daunting indeed, yet our score for Hitler's potency would be critically dependent upon getting it right.

The obvious alternative, of course, is simply to set each cause at the level that actually obtained in reality. But to do this while simultaneously avoiding the problem of undesirable inconstancies, it is necessary to include interactive effects in any definition of causal strength.

*Conclusion*

Intuitively, perhaps it is not immediately clear whether or not to include interactive effects in our definition. But once we realise that causal strengths must be understood context-specifically, it quickly becomes apparent that we should. Inclusion seems the natural strategy in the case of an actual example from political history. It is able to evaluate interactive effects separately if desired, and also provides a good account and treatment of the main opposing intuition. By contrast, analysis based on excluding interactive effects runs into difficulties almost immediately. Moreover, it is unable to accommodate interaction with background causes, or to keep results for causal strength independent of arbitrary foreground/background designations.

Therefore there is a clear answer as to how a definition of causal strength should handle interactive effects. And happily, all the desiderata our account satisfied earlier remain satisfied now.

# 3 Further discussions

*Relation to general causation*

As already stated, our definition is only applicable once we are *given* a particular causal ontology. In addition to that, it also seems that none of the issues raised by defining causal strength turn on any of the classic debates surrounding causality, such as how to infer causes from statistics, or whether we are Humeans or realists. Nevertheless, causation and causal strength do presumably still have some connection. In particular, if something registers a positive strength then we would take this to imply that it is a contributing cause with respect to that particular effect. On standard definitions of a contributing cause, this is clearly so. Equally, if something registers zero strength then

we would take this to imply that it is *not* a contributing cause with respect to that particular effect. However, this latter statement has been challenged.

[Sober 1988] states: 'I therefore seem to find myself in the paradoxical position of saying that genes can be a cause of height, even if they are judged to have zero magnitude ... causes may make no difference, but they are causes nonetheless' (pp317-18). That is, it is possible that even though something is clearly a cause of an effect, nevertheless (on at least some definitions) we can assign it zero strength. But 'perhaps this air of paradox can be dispelled … it is not hard to fathom how causes can fail to be necessary for their effects' (pp317-18). For example, Holmes's shot is clearly the cause of Moriarty's death. Nevertheless, if Watson was present too, the shot was not necessary for Moriarty's death since that would have come about in any case. Accordingly, in such circumstances we would assign the shot zero causal strength in the DM sense even though it is still clearly a cause.

But I think the 'air of paradox' here just rests on conflating absolute with *relative* causal strength. Thus Holmes's shot is certainly a cause, but if Watson shot too then it has zero *relative* strength. In our terminology, it is clear that something can have a high PM, yet still have a zero DM with respect to another cause of equal PM. In my view, there would only be a real paradox if the *PM*, or *absolute* strength, of Holmes's shot was assigned to be zero - and that is not the case. Once we speak only of absolute strengths, then causal strength again seems to track general causation obediently.

Similar remarks apply to other standard examples of causal pre-emption analogous to the Holmes-Watson case. What of cases of overdetermination, such as if we both poison and stab someone to death? Our definition would award maximum PMs to each cause if W does not include the other, and zero PM to each if W does include the other. The two causes' joint potency will also be maximal. This perhaps adds little to the existing literature on cases of overdetermination, but the point here is merely to confirm that our definition of PM does not diverge from our general intuitions about causation, such as they are in cases like these.

*Locality*

Intuitively, the concept of causal strength seems to be something intrinsic and local. (Note that the topic here has nothing to do with the sense of 'locality' found in quantum mechanics [Sober 1988].) For example, there is a direct causal chain connecting Holmes's shot to Moriarty's death regardless of what Watson might or might not have thought of doing. But [Sober 1988] suggests there is a sense of causal strength that, surprisingly, must be understood non-locally - namely, the second intuition in the Holmes-Watson example, according to which Holmes's shot makes no difference. On this second understanding of causal strength, 'even if causality is local, the magnitude of causality need not be' (p318).

But we have now seen that this second understanding of causal strength is in fact just the *relative* strength of a cause, and furthermore that a relative strength is just the difference of two absolute strengths. It is surely hardly surprising that a relative strength should be 'non-local' in the sense of being dependent on what it is relative *to*. Any assignation of relative strength will be dependent on the choice of comparison: the strength of C1 relative to C2 always depends on the strength of C2 as well as C1. But this no more makes causal strength mysteriously 'non-local' than does the similar

dependence on external factors of relative speed make *speed* mysteriously non-local, or the dependence on external factors of relative mass make *mass* mysteriously non-local.

Next, recall again our actual definition of the PM of a cause C:

E(C&W1) - E(C0&W0).

This has the form of a definition of relative strength, of course - in particular, relative to the neutral level C0. Often C0 has a natural interpretation as a zero level of C, or as its absence. In these circumstances, I think it is clear that (for a given E and W) a cause's absolute strength is indeed independent of external factors and hence 'local', and furthermore that the non-locality of relative causal strength is (as just argued above) no counterargument to this.

The situation does become more complicated when there is no natural zero, for instance with the interpretation of C0 in our biology example when C was environment. In such cases, it may be that we set C0 with reference to something external, so that, strictly speaking, [Sober 1988] is right to insist on the non-locality of causal strengths in biology. But C0 is always neutral with respect to E *by definition*, so it seems strange to insist on the non-locality of absolute causal strength here. It is not as if by varying other inputs we can thereby vary our score for that strength, as we can with relative strengths.

Finally, we return to a concern first raised in section 1 - if PM and DM are each definable in terms of the other, why we should we take PM to be the more fundamental of the two? Our discussion now suggests that there is some reason for preferring PM in this regard. Turn again to the analogy with speed. If we wish to determine the speed of a 70mph car, we need pay no attention to any other car but merely need measure only how fast this particular car is travelling, whereas if we wish to determine the *relative* speed of the 70mph car then of course we must also take into account the speed of some other car or reference object after all. But there seems to be a sense in which something physically important about the car is indeed independent of what any of the other cars are doing - for instance, its absolute speed may be causally independent of that of the others. So it seems desirable for a definition of the relevant physical quantity to capture this locality, and accordingly we should prefer to talk in terms of speed rather than relative speed.

An exactly analogous argument holds with respect to causal strengths. There seems to be something physically significant we are trying to capture with the notion of causal strength, and this something is independent of the strengths of other external causes. Accordingly, we should prefer a definition of causal strength that captures this sense of locality. It follows that we should therefore regard PM as more basic than relative PM, i.e. regard it as more basic than DM.

*Commensurability versus separability*

[Sober 1988] agrees with biologists that making sense of absolute causal strength in the genes-environment case is impossible, and argues that this is due to the two causes' effects being incommensurable. Whereas gravitation and electricity, for instance, each act on a Newtonian particle via the common medium of force, we seem to have no such common currency for comparing the impacts on a plant's height of genes and environment. Perhaps if this were not the case then we could after all judge whether genes or environment had the greater absolute strength. [Lewontin 1974, p402] makes

the same point vividly: 'if two men lay bricks to build a wall, we may quite fairly measure their contributions by counting the number laid by each; but if one mixes the mortar and the other lays the bricks, it would be absurd to measure their relative quantitative contributions by measuring the volumes of bricks and of mortar.' Accordingly, [Sober 1988, p. 312] offers the following conjecture: 'For it to make sense to ask what (or how much) a cause contributes to an effect [i.e. to evaluate what we have been calling PM], the various causes must be commensurable in the way they produce their effects.'

If this really were true, it would represent a serious limitation on the applicability of our definition. All causes are 'commensurable' in so far as they impact on the same effect - indeed our definition exploits just this fact. But the claim here is stronger, namely that they need also to be commensurable in the *way* they produce their effects. Thus although the bricklayer and mortar-mixer each contribute to the same final effect of a built wall, still the strengths of their contributions are not comparable because they each contribute in incommensurable currencies, as it were. There are many causes that are incommensurable in this second sense. Must we declare them all incomparable for causal strength? I shall argue not, and that actually commensurability in this sense is irrelevant to the issue at hand.

There are also two ways to interpret Sober's phrase 'make sense'. The first is just that it refers to whether or not an absolute strength is definable at all. On our definition of PM, this seems straightforwardly to depend on whether or not an acceptable C0 can be found and therefore has nothing to do with commensurability. There is a second interpretation though - that commensurability is necessary for what I shall call *non-trivial* assignations of absolute strength. Thus even on our scheme, the actual calculation of PMs in the gene-environment case is admittedly trivial since both factors were awarded identical strengths, namely exactly the plant's actual height (section 1). But even if taken in this second way, I still disagree with Sober's conjecture.

True enough, there does seem to be something distinguishing cases like genes-environment with trivial causal strengths from cases like gravity-electricity with interesting ones. I believe the important factor is not commensurability though; rather, it is what we might call *separability*. By this I do not mean merely that two causes are distinguishable (although this too is necessary); rather, I mean that their *effects* are (potentially) distinguishable. For example, genes and environment are easily individuated, but nevertheless their effects are still not separable. The key structural feature is really whether or not two causes are individually sufficient to produce any effect. If they are not, as with genes and environment, then it can make no sense awarding them different absolute strengths since their individual contributions to the final effect are inseparable. Without the other, neither input could produce any final effect at all, and this property is symmetric. If, on the other hand, as with gravity, electricity and the motion of a Newtonian particle, the two causes *are* individually sufficient to produce some effect, then (and only then) each may be deemed individually responsible for different particular quantities of that effect, and hence this time their absolute strengths may indeed differ.

Note that this whole issue applies only to absolute, not to relative, causal strengths, that is to PMs not DMs. For example, which is more important for producing speech, my brain for thinking of the words or my vocal chords for generating the physical sound?

Clearly, both are necessary for producing any speech at all and so in our sense are inseparable. Accordingly, each must be awarded the same absolute causal strength. But comparing my vocal chords when healthy to when they are hoarse, it may well be that my power of speech is a little bit greater - thereby yielding a positive but small relative strength for healthy vocal chords. Comparing my powers of speech before and after a major stroke, on the other hand, it may be that the difference is now enormous, indicating a much larger relative strength for the healthy brain. Thus sometimes two causes' relative strengths may be interestingly comparable even when their absolute ones are not. Two inputs' individual insufficiency implies the triviality only of their PMs, not their DMs.

To summarise so far: our interest lies in what determines whether or not, in our terminology, a calculation of PM is trivial. In the Newtonian particle case the calculation is not trivial; in the genes-environment case it is. Is commensurability or separability the key factor? I have been arguing that it is the latter. Now, in the Newtonian particle example we have both commensurability and separability, and we get non-trivial PMs. In the genes-environment example, we have neither commensurability nor separability, and the PMs are trivial. So neither of these is a decisive case, since of course they are both consistent with either of commensurability or separability being the key factor. To illustrate that separability is indeed what really matters, we shall need to create two further examples, this time with commensurability and separability diverging.

Our first will be a case where the causes are commensurable but inseparable. Imagine a primordial soup in the early history of the Earth, in which there are two chemicals which can react to synthesise some complex organic combination but which will only do so given a certain activation energy. Imagine further that there are two thunderclouds passing overhead, a large one and a small one. Suppose that a lightning bolt from the large cloud is more energetic than one from the small cloud, but still cannot on its own provide enough energy to trigger the reaction in the primordial soup. Therefore neither of course can the small bolt. However, if the two lightning bolts strike simultaneously, then the combined energy of the two together does go past the activation threshold and the chemical reaction will be triggered. In other words, the two bolts are individually insufficient but jointly sufficient. Assume finally that the two bolts then do indeed strike simultaneously and that the chemical reaction is indeed triggered; what PM should each bolt be awarded?

The effects of the two lightning bolts are surely commensurable if anything is - they are, after all, two examples of exactly the same phenomenon. But I claim that their impacts, *with respect to this effect*, are nevertheless inseparable. Individually, neither triggers the chemical reaction; together, they do. Therefore, measuring their causal strengths in the usual way by how much of the effect we are interested in they produce, we have to conclude that individually each has zero strength, while together they have full strength. This of course is exactly the same situation as in our genes-environment example: on their own, neither genes nor environment can produce any plant, while together they produce a full plant. The important point is that commensurability plus inseparability has yielded a case of trivial absolute strengths. All the causal strengths here will be either zero or maximal, and there is no way of saying that the strength of one bolt is any different from that of the other.

Note particularly that this analysis holds even though we have specified that one bolt

is bigger than the other. Intuitively, of course, one might assume that the bigger bolt should be assigned more strength than the smaller one. It is this intuition, perhaps, that motivates the emphasis on commensurability - since the energies of the two lightning bolts can be directly compared (i.e. are commensurable), *therefore* differential strengths can be assigned. But *in this context*, I believe such reasoning is incorrect. Remember, the specific effect we are concerned with here is the chemical reaction, and this is dichotomous - it either occurs or it does not. To be sure, when considering how efficacious they are at producing other effects, for instance inducing voltage in a wire, then of course the two lightning bolts may well have different strengths. But when considering our particular effect of triggering the chemical reaction, because of the activation energy threshold I do not see how assigning different strengths could be justified. In our *particular* example, that is, the calculation of absolute strength must surely be trivial, even though our two causes are commensurable, and even though in *other* contexts one of these causes may score a bigger absolute strength than the other.

The bottom line, as it were, is that what matters is separability, not commensurability. [Sober 1988] gives a thought-example of genes and environment each contributing 'height particles' to a plant, and claims that this would enable non-trivial comparisons of absolute strength through enabling commensurability of genes' and environment's effects. But my view is that these height particles could only achieve that in so far as they led to separable impacts on the plant's final height. Their commensurability is irrelevant.

This conclusion is reinforced if we consider finally the last category of example, this time the other way round from before: now, with separability but not commensurability. Suppose I am taking my dog for a walk on a windy heath and he gets interested in a ball lying in the grass a long way from me. I want him to come back to me, so call out to him. Assume that hearing or seeing my call induces him indeed to move back to me. Now suppose that at exactly the same moment as my call, an especially huge gust of wind blows up. Suppose further that he is a relatively small dog and accordingly that this huge gust physically blows him towards me, independently of any voluntary motion he might also have been undertaking. So we now have two independent causes - namely the dog's reaction to my call and the physical gust of wind - each producing the same effect, namely the dog moving closer to me. What is each cause's strength with respect to this effect?

I think we can answer that straightforwardly. The definition of the wind's PM is how much the dog moves given the gust of wind compared to if there had been no wind. And similarly, the PM of my call is given by how much the dog moves in response to it compared to if I had not called. This straightforwardness is a direct result of the easy separability of the two causes' effects. The two causal strengths could perfectly conceivably differ from each other, and the case is clearly analogous to our Newtonian particle example in this respect. But unlike electricity and gravitation in that Newtonian particle example, the two causes here do *not* seem to be commensurable. My call presumably stimulates some reaction in the dog's brain, and thence voluntary movement. The gust of wind, by contrast, bypasses such mechanisms completely and simply physically pushes the dog's body. How could we define one unit of wind gust and equate it to one unit of call? There is no analogue to the common role of force in the Newtonian particle case. But despite this lack of commensurability, non-trivial absolute causal

strengths are still calculable. Once again, we see that what really matters is only that we are able to distinguish two causes' impacts on the final effect, in other words that what really matters is separability.

# References

Humphreys, P. [1990]. *The Chances of Explanation*, Princeton: Princeton University Press.

Lewontin R. [1974]. 'Analysis of variance and analysis of causes', *American Journal of Human Genetics*, 400-411.

Northcott, R. [2003a]. 'Can a cause have a potency independent of context?', presentation to University of London graduate philosophy conference, May 2003.

Northcott, R. [2003b]. 'Causal strength and the analysis of variance', presentation to LSE 'Causality: Metaphysics and Methods' group, July 2003.

Northcott, R. [2003c]. 'Approximate truth and causal strength in science' PhD dissertation, London School of Economics.

Sober, E. [1988]. 'Apportioning causal responsibility', *Journal of Philosophy*, 303-318.

Sober, E., E.O. Wright and A. Levine [1992]. *Reconstructing Marxism*, chapter 7: 'Causal asymmetries'