

A Theory of Evidence for Evidence-Based Policy

NANCY CARTWRIGHT WITH JACOB STEGENGA

I. Preliminaries

The project

WE AIM HERE to outline a theory of evidence for use. More specifically we lay foundations for a guide for the use of evidence in predicting policy effectiveness *in situ*, a more comprehensive guide than current standard offerings, such as the Maryland rules in criminology, the weight of evidence scheme of the International Agency for Research on Cancer (IARC), or the US ‘What Works Clearinghouse’. The guide itself is meant to be well-grounded but at the same time to give practicable advice, that is, advice that can be used by policy-makers not expert in the natural and social sciences, assuming they are well-intentioned and have a reasonable but limited amount of time and resources available for searching out evidence and deliberating.

We go into the project with some assumptions. The first is a delimitation of the topic. The guide for which we aim to lay a theoretical base is to be concerned with the use of evidence to estimate, if only roughly, whether, were a proposed policy to be actually implemented, a specific, identified outcome would be produced. We thus do not discuss the broader issue of how to settle on goals. Nor do we discuss how to recognise when a result of a scientific study, formulated using concepts that can be tackled with the procedures of the study, is relevant to the more abstractly and vaguely set out goals that are often the real aims of policy.¹ Nor do we present ideas here on how to come up with a set of candidate policies for achieving a given goal. It is also important to keep in mind that whether a policy will achieve its stated goals is only one of

¹For instance we may want an educational program that makes children better adapted to live full, independent lives and to become contributing citizens but proper scientific method requires the study of precisely defined, measurable outcomes, like reading scores on an Iowa Test of Basic Skills.

many considerations that should go into policy decisions.² We treat here only the far simpler but already difficult problem of judging whether a particular proposed policy is likely to achieve a particular already well articulated goal.

Our second starting assumption is that the project needs to be approached from the point of view of the evidence user, not the evidence producer.

Third, we assume that rigour is a good thing, so that the advice should be firmly rooted in sound principles; but we must not be pseudo-rationalistic. A rigorous argument with nine well-grounded premises and one weak one does not make for a rigorously established conclusion. For the most part, estimates of whether a policy will be successful made in real time will be both rough and uncertain. That is important to keep in mind as policy decisions are made. But it is also important to keep it in mind as advice guides are devised. If advice is to be practicable, it may not be hugely reliable, even if it is ultimately well-grounded. We should aim for advice that improves decisions even if we cannot do the job perfectly. The best should not be the enemy of the good.

Fourth, and closely connected with the third, is that we should not expect policy effectiveness judgements to be very reliable. There are a variety of different reasons conspiring to make these judgements especially difficult, including the obvious difficulties of doing what we propose here as necessary for reasonably reliable judgements. We shall not rehearse these reasons but just offer one remark to make vivid how difficult the task is. Asking if a policy of a specific design will achieve a targeted result is structurally just like asking whether a laser of a specific design will produce a coherent beam when we plug it in. It is difficult to answer that question reliably before actually plugging the laser in—it is similarly complicated to produce advice about what counts as evidence for or against an answer and about how to marshal that evidence to settle on a prediction. Social effectiveness will be even harder since the systems under study are more open, our theories and knowledge of the materials are less secure, and the choice of targeted outcomes is generally dictated by social need, not by an assessment of how achievable they are.

How to think about the problem

Viewpoint

When it comes to evidence-based policy, viewpoint matters. Whether wittingly or not, typical advice guides focus on the *production side* of scientific evidence

²For examples of the many other types of issues that need consideration see section on 'Effectiveness' at pp. ??? below.

and not on the *use* side. They tell us what counts as good science, not how to use that science to arrive at good policy.

Most available guides, like the Maryland rules, the IARC scheme and What Works, provide ranking schemes for the ‘quality’ of evidence. These schemes police the credibility of results that can be counted as evidence. Evidence claims are ranked according to the methods by which they are tested. High quality means that the tests are stringent: Results that pass the tests are very likely to be true. *Randomized controlled trials (RCTs)* are necessary for strong evidence according to the dominant guides. Many object on the grounds that this can mean throwing out a lot of good evidence that we ought to be attending to. This issue is not our concern here. The central concern we raise is that these rankings focus on too narrow a range of *claims that need evidencing*, not that the kinds of evidence admitted are too narrow. Why?

Truth is a good thing. But it doesn’t take one very far. Suppose we have at our disposal the entire encyclopaedia of unified science containing all the true claims there are. Which facts from the encyclopaedia do we bring to the table for policy deliberation? Among all the true facts, we want on the table as evidence only those that are *relevant* to the policy. And given a collection of relevant true facts we want to know how to assess whether the policy will be effective in light of them. How are we supposed to make these decisions? That is the problem from the *user’s* point of view and that is the problem of focus here.

Here is how Dr Sean Tunis, director of the Center for Medical Technology Policy, a US organisation concerned with ways to get better medical evidence, puts the problem: ‘There’s this gulf between what questions researchers have found interesting to study and what questions industry and the N.I.H. have chosen to fund and what users of information most want to know.’ In our terms, the focus has been on the side of evidence *production*, rather than evidence *use*: ‘One starts from the head and the other starts from the tail and they don’t meet in the middle’ (cited in Kolata, 2008).

Effectiveness

There are a great many things we need to evaluate in considering whether to adopt a policy or not. Will the policy work? Does it have unpleasant side effects? Does it have beneficial side effects? How much does it cost? Have we made the correct choice of target outcomes? Is the policy morally, politically and culturally acceptable? Can we get the necessary agreement to get it enacted? Do we have the resources to implement it? Will enemies of the project sabotage it in various ways? Every one of these questions needs answering and in each case evidence will help get the right answer.

We shall confine our discussion, however, to the *question of effectiveness*:

Question of Effectiveness. Will the proposed policy produce the targeted outcomes were it to be implemented in the targeted setting and implemented in the way it would in fact be implemented there?³

A structure for the problem

Start then from the point of view of the policy deliberator trying to estimate whether a proposed policy will be effective. *For a reliable decision one wants credible evidence that, all told, speaks for (or against) the policy.* This simple observation suggests that from the point of view of the user three different issues need addressing:

- 1 *Quality:* When are evidence claims credible?
- 2 *Relevance:* When does an established result bear on a policy prediction and how does it do so?⁴
- 3 *Evaluation:* How should predictions about policy effectiveness be evaluated in the light of all the evidence?

The first is an issue about the production of knowledge by the social and natural sciences; it is the meat of evidence-ranking systems. The latter two are the more neglected questions we focus on.

The fact that the three questions are distinct should not suggest that their answers are unrelated. Despite the common emphasis on question 1, it seems *prima facie* as if the natural starting point is with question 2. First establish what kinds of evidence are relevant to effectiveness. Then, for question 1, provide guidelines that police the quality of evidence of those kinds; and for question 3, propose some scheme for amalgamating or combining evidence.

In aid of this approach one could adopt one or another of the characterisations of relevance on offer from philosophy and methodology of science, where the topic has been explored and debated for years; then follow on with one or another of the schemes available for combining evidence or adapt weighing schemes with known characteristics from other areas, like those for amalgamating preferences or expert testimony.

³Of course there will seldom be a highly certain yes or no answer. So at some point an assessment of the probabilities will have to be made in light of the evidence, even if only roughly. But reasonable probability assessments depend first on understanding the structure of the problem, which is the topic to be tackled first.

⁴But, as mentioned above in fn. 1, there are many important aspects of this issue that we will not discuss here, including how to relate the concepts of scientific studies to those in which goals are often framed.

We adopt a different strategy. We propose to start with an account of how to evaluate claims of effectiveness and work backwards to figure out what kinds of evidence would be relevant for the evaluation, finally returning to the first issue of how to assure that the kinds of evidence claims needed are sufficiently credible to enter into deliberation.

Before beginning with this account, we want to stress the importance for the success of evidence-based policy of covering all three questions. Question 1 is a question for knowledge producers: What is necessary in order to ensure that a claim entered as evidence is likely to be true? Users have in addition to face questions 2 and 3.⁵ Yet most of the rigour and most of the attention is to question 1. We are urged to extreme rigour at one stage, then left to wing it for the rest.

But: a chain of defense for the effectiveness of a policy, like a towing chain, is only as strong as its weakest link. So the investment in rigour for one link while the others are left to chance is apt to be a waste. To build the entire chain one may have to ignore some issues or make heroic assumptions about them. But that should dramatically weaken the degree of confidence in the final assessment. Rigour isn't contagious from link to link. If you want a reasonably secure conclusion coming out, you'd better be careful that each premise is secure enough going in.

II. Evaluating effectiveness

How philosophy can help

We propose to borrow our three central principles of the theory of evidence for use from philosophy. The first two provide the basis of the theory and the third, some practical help in implementing it.

- Truth values for causal counterfactuals are fixed by causal models.
- Causes, as J. L. Mackie explains, are INUS conditions.

⁵Is relevance really, as we say, a question for the user rather than the knowledge producer? Many think not. Indeed it is a common criticism of studies in the social sciences that they do not say what they show, what the results bear on, at a practical level. We don't think they can. Perhaps they can do better, but there will always be a great number of relevance judgements that must be left to the user. Whether a given fact is relevant as evidence for a given claim depends on a host of other assumptions, both theoretical and local to the situation. (This is the lesson of the famous 'Duhem-Quine' problem in the philosophy of science.) For causal counterfactuals of the kind we assess in effectiveness evaluations, relevance will depend in addition on *how* the cause is supposed to produce the effect. (See Part V below.)

- In understanding how causes operate and how they operate together, mechanisms matter.

Causes and counterfactuals

For sound policy we need to evaluate whether, if the proposed policy were implemented as it would in fact be implemented, the targeted outcome would occur in consequence. We are looking for the probability of what in causal decision theory is called *a causal counterfactual*.⁶

There is good reason to expect an intimate connection between causes and these special kinds of counterfactuals. Nature forges it. Consider: How does nature decide what effects to produce in a particular situation? First she surveys the causes that will be operating. Next she consults her rules of combination to calculate what should happen when they all act at once. Then she produces the prescribed effects. We can't lose by imitating nature.

That is our proposal. To predict what will result if we introduce some new policy or programme, follow Nature's lead: Reconstruct Nature's list of causes and mimic Nature's calculation.⁷ This provides us with a good way to predict the effects of our policy implementations and we can't go wrong if we succeed. Moreover, any method that does not directly mimic Nature's processes will only get predictions about causal counterfactuals right (or 'right enough') if it has some way of achieving just the same results. Later (in Part IV) we consider 'cheap heuristics' that might get the same conclusions enough of the time in specific kinds of circumstances. These are great when they are available. But their conclusions are only warranted to the extent that we have good reason to believe that they will produce near enough the same results as would a causal model that mimics Nature's procedures.

Since it is often not possible to make life easier and getting the causal model 'right enough' is usually very difficult, any reasonably comprehensive guide will also need to remind policy analysts to expect a great deal of uncertainty and

⁶These are commonly called 'counterfactuals' despite the fact that it is generally possible for the antecedent to obtain, and were it to obtain, the consequent would obtain as well, if the 'counterfactual' is true. Some find 'subjunctive conditional' a more apt label, but the term 'counterfactual' is what is generally used throughout philosophy and we will follow that usage here.

⁷A referee expresses concern over the concepts 'Nature's causes' and 'Nature's calculations'. Perhaps this is an expression of a David Hume-inspired scepticism about causes. There is, however, a large, articulate and compelling body of literature arguing that contrary to this sceptical position causal notions make good sense and are essential for a useful and accurate description of the natural and social world and especially for understanding and evaluating claims about the effects of intervening.

to adopt strategies for dealing with it—strategies like not introducing big policy changes that are difficult to reverse and adopting a muddling through rather than grand planning approach.⁸

Causal models

We propose then to adopt standard philosophic advice as the first principle of the theory of use: To evaluate causal counterfactuals, build a causal model (see Reiss, 2007). But the term ‘causal model’ should not carry a lot of baggage with it, either from philosophy or from the sciences, where various different kinds of specialised causal models are on offer.

What is a causal model?

For our purposes a *causal model* has two essential ingredients, where we separate the first into two parts to highlight issues about implementation that we know policy makers need to take into consideration.

- 1 A list of the causes relevant to the targeted effect that will operate in the target situation. This includes
 - 1.a the causes present in the situation independent of the policy action
 - 1.b any changes in this set of causes introduced in implementing the policy.⁹
- 2 A rule of combination that calculates what should happen vis-à-vis the targeted effect when those causes operate together.

Consider a simple case. Later we shall look at both some real and some pastiche social policy cases. But for now we illustrate using everyday physics. We do so because the reasoning is simple, well-understood, and we are not likely to get involved in subject-specific debates in education or criminology or health policy. More importantly, we choose this kind of case to start with because it is one where our knowledge of the principles and of the aptness of the concepts is secure so that we can focus on the structure of the reasoning needed.

⁸Thanks to a referee for encouraging us to mention this. The referee also suggests consulting William Dunn, 2003 and Charles Lindblom, 1979.

⁹Remembering, as a referee stresses, to include recipient reactions that can affect the outcome. For instance as the referee points out, ‘Whether something is effective in a public policy system depends on whether people like the policy outcome, or even the policy mechanism in its own right (e.g. in the case of some ‘effective’ or ‘coercive’ labour market and welfare policies)’.

The case of the desk magnet versus the industrial magnet. We have access to a desk magnet and to a large industrial magnet. We know the exact strengths of these with a very high degree of certainty—claims about their efficacy for lifting objects have passed far more than two good RCTs; they have centuries of study behind them. Shall we use one of them to lift an object in my driveway? That depends on the other features of the target situation.

First, magnets need helping factors to be effective at all. A desk magnet is useless for lifting a matchstick; it is only the *combination* of a magnet and a ferrous object that produces a magnetic force. Then the acceleration caused by the magnet is still only one part of the story, often one very small part. To know what happens when we apply the magnet we need to know the other forces as well. Here, especially gravity. The desk magnet may lift a pin but it is hopeless for a car, where we need the industrial magnet. We also need to attend to what other forces we introduce in the course of getting the magnet in place. Perhaps the industrial magnet would have lifted the car if only we hadn't thrown the heavy packing case for the magnet into the boot.

Finally we need to know how all these factors combine to produce a result. Often in social contexts additivity is assumed: Add a good thing and the results can only get better. But that doesn't work in even this simple physical case. We get so used to vector addition that we forget that it isn't simple (scalar) addition of effect sizes. Add a magnetic acceleration of 42 ft/sec/sec to that of gravity's 32 ft/sec/sec and you won't usually get an acceleration of 74 ft/sec/sec.

The point is that whether the magnet will be effective at all in the target situation and to what extent depends on nature's causal model of the situation. So the most direct way of predicting its effects is to construct our own causal model in imitation of nature.

We know no one wants to hear this since it seems difficult. But consider: Industrial magnets would pass any number of RCTs, of any degree of stringency. But that's not anywhere near enough to know. None of us would rent an industrial magnet to remove a load of rubbish without looking at the rubbish. Knowledge that magnets just like this *can* lift is only a small part of what one would consider in evaluating whether renting the industrial magnet will be effective in removing our rubbish. If this is so in everyday calculations and in applied science and engineering, why should we expect it to be substantially different—and substantially easier—in social engineering?

Of course constructing causal models is hard, even if the models are rough and we have figured out ways to tolerate uncertainties. Sometimes there are shortcuts, 'cheap heuristics' that get us, more-or-less, well-enough, the same conclusions that the causal model generates. As decision makers we can opt for a heuristic if we want. But there is no avoiding the fact that the choice of the right heuristic depends on the right causal model. We may not wish to build a causal model; we may not know how to; we may think it takes too

much time or money, intelligence or attention. That does not alter the fact that when we buy a policy we are betting on a causal model, willy-nilly, whether we wish to think about it or not.

Had we world enough and time

A great deal more can be said about causal models. But it is subject- and discipline-specific and almost always requires expertise and training to do at all properly. Moreover, many scientific models do less than what we demand of a causal model, though they provide more detail and zero in, usually very precisely, on specific features of interest.

Consider a joint effort to explore the causes of delays in emergency rooms (Lane, Monfeldt and Rosenhead, 2000). The modelling expertise was provided by the Department of Operational Research at LSE, while orientation to the problem area, judgements on design choices, and introductions to stakeholders were supplied by Casualty Watch, a project organised as a response to public concern that cuts in the NHS were producing an inadequate emergency service and harming patients. System dynamics was selected as the appropriate modelling medium and the model was calibrated with information from an inner London teaching hospital.

Here's what the model looks like:

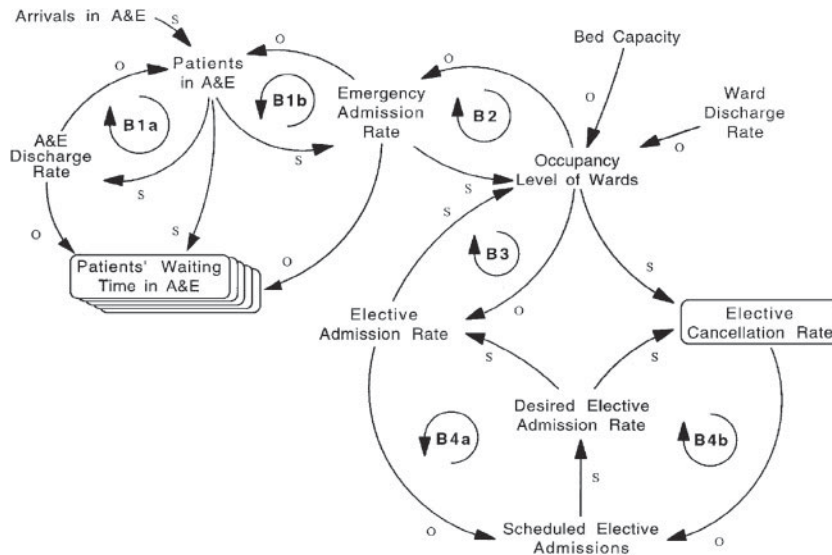


Figure 11.1. Model of delays in emergency rooms.

What's important about this model is its ability to detect and represent feedback loops and its dynamic structure. For instance it makes clear that the number of beds available in the wards both affects and is affected by the number of admissions from A&E and that the number of patients being tended in A&E affects and is affected by the number of patients being admitted to the wards from A&E. It also shows a number of pathways by which an initial cause, say arrivals at the Accident and Emergency Department, influences the final effect, patient waiting time at A&E.

As we shall explain in Part V, tracing through the dynamics like this, step-by-step, can be a big help in constructing a significant part of the second component we demand in a causal model—an account of how causes act together to produce the targeted effect—because it focuses on what auxiliary causes are needed at each step if the salient cause is to produce the next step in the process. Notice, however, that this information is not explicitly represented in the model since the model treats causes singly. At the head of the arrow—at the causes end—is a single variable; e.g. bed capacity, ward discharge rate, and emergency admission rate are all pictured as separate causes of the ward occupancy rate. There is no information encoded about how these different causes combine, in particular which causes must act together before they can contribute to the effect at all. Thus this model, like most professional models, does less than we require, though what it does, it does more precisely and in more detail.

Another example (Figure 11.2 opposite), this from Judea Pearl (1995: 669–70).

In this model, as in the last, causes are at the top of the arrow, effects at the tip. By calling it a causal 'Bayes net' special assumptions are made about the relations among the variables that may not hold in every causal model; for instance causes and effects pictured in the graph are all supposed to be probabilistically dependent. Generally this kind of model comes with numbers as well, ideally the conditional probability for each effect conditional on all the immediately prior causes leading into it. So these models contain more information than is required by our two conditions for a causal model, information of special use¹⁰ in the particular kinds of causal systems that satisfy the

¹⁰This information plus the graph, assuming the graph is causally correct and the Bayes-nets axioms are satisfied, is tantamount to having the full probability measure over all the variables in the graph. It is thus possible to predict the probability of any outcome conditional on values of antecedent variables, which naturally can be very useful. But this raises an important point about modelling to predict singular counterfactuals. A full probability over the relevant variables will allow us to predict how probable a desired effect, is given that the policy variables take the

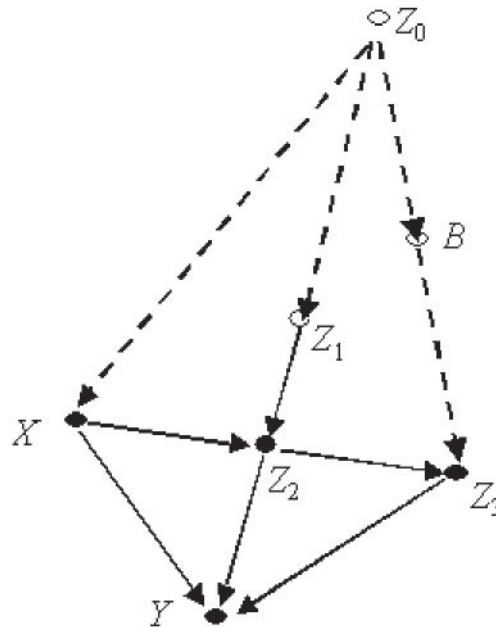


Figure 11.2. A causal Bayes net.
Variables: X : fumigants; Y : yields; B : the population of birds and other predators; Z_0 : last year's eelworm population; Z_1 : eelworm population before treatment; Z_2 : eelworm population after treatment; Z_3 : eelworm population at the end of the season.

special axioms that relate causes and probabilities in a Bayes net.¹¹ But like the dynamic-systems model for emergency room admissions and hospital beds, it also contains less since the model does not show how the causes interact among themselves to affect yields. We know from the graph that Z_2 can influence Y but even if we add to that knowledge of the conditional probability of

proposed values. But only if the probability is over the individual events that will be implemented in the specific way they will be at the specific place and time under consideration. It is just this probability that is so difficult to find—if it exists at all, which many of us doubt.

¹¹For instance, one axiom requires that immediately prior causes on the graph and their effects are always probabilistically dependent, which means that no causes act both positively and negatively by different paths that cancel each other. A second requires that a full set of prior causes screens off a factor from anything except its causal descendants. This implies, among other things, that no causes produce their effects probabilistically in tandem. For instance, no purely probabilistic causes produce a particular effect just in case they produce a particular side effect. Rather all effects are produced independently of all others (cf. Pearl, 2000).

Y on Z_2 we don't know from the graph whether the presence or absence of X is essential to the ability of Z_2 to influence Y.¹²

This kind of missing information is readily supplied by models presented in the form of equations, if they can be constructed. Here for instance is the final equation from a causal model we shall discuss in *Part V*:

$$(*) y_t = \theta\beta[p_t - p_{t-1}] - \theta\beta\pi + y_{pt} + \varepsilon_t$$

Here y_t is output at t and p_t is price at t so $[p_t - p_{t-1}]$ is a measure of inflation; ε_t is a random 'error' variable.¹³ This equation yields as a next step the classic Philips curve representing a trade-off in which rising inflation causes decreasing unemployment. Once the parameters, θ , β , and π , are filled in the equation shows how the two causes represented—inflation, $[p_t - p_{t-1}]$, and earlier output, y_{pt} —combine to produce later output, y_t ; in this case, simple linear addition.¹⁴

We will present a simple physics example (see *Illustrations of INUS conditions*, below, p. ???) where a complete set of causes is also laid out in an equation, but the rules of combination for the causes are more complicated involving not simple (scalar) addition but also multiplication and vector addition.

Equations for calculating the exact result of a given set of causes are wonderful when one can get them. But they may not be possible even in principle for many cases. Even a complete set of causes may act only probabilistically, not fixing a value for the effect at all but only a probability. In fact we hazard that that is more often the case than not. And even that may be wishful thinking. Nature herself may proceed with less quantitative precision, not fixing even a final probability, perhaps only a direction of change. Whether she does so or not, this level of precision is generally well beyond the ability of normal policy deliberators. Also, as our colleagues at a recent conference on causality

¹² Many of those developing the theory of causal Bayes nets describe them as a method for 'causal discovery'. We think that's right. They are tools on the knowledge production side; a way to sidestep the need for RCTs by establishing efficacy with the same degree of rigour as an RCT but using population, not experimental, data. They may even be of more immediate relevance to policy than an RCT if the data comes from a population reasonably deemed similar in the right respects to the target population. Still, without further additions, they are not enough to evaluate causal counterfactuals. (Though see Judea Pearl's (2000) beautiful work on how to use them to evaluate the probability of causal counterfactuals, given input probabilities for exogenous factors and given that the special Bayes-nets axioms hold in the system under study.)

¹³ Hence ε_t has a probability measure over it. This variable does not refer to any 'known quantity' but serves at one and the same time to stand for omitted causes and measurement errors and as a representational device to allow a deterministic-looking equation to represent a purely probabilistic connection between the designated causes and the designated effect.

¹⁴ When ε_t is included, the causes will not fix a value for the effect but merely its probability.

urged: Our list of causes will almost always be incomplete; the very best we can hope for is a probabilistic assessment of the outcomes and even that should generally not be too precise. So don't get hung up trying to produce equations.

But that is not advice to ignore the need to get a grip on the dominant causes that will affect the outcome or the need to bet on what they do in combination. It is just advice not to expect a degree of precision or a degree of confidence that neither the subject nor our capabilities can support.

INUS conditions

Introduction

To evaluate a causal counterfactual one needs to consider the major causes at work and how they combine. One characteristic of causes widely accepted in philosophy can help with both enterprises. As J. L. Mackie (1965) argued, causes are INUS conditions.¹⁵

An *INUS condition* is an **I**nsufficient but **N**ecessary part of an **U**nnecessary but **S**ufficient condition.

The factors we normally call causes are, according to Mackie, INUS conditions. Causes—in our usual sense of the word—are not enough on their own to produce an effect. Causes work in cooperation; they need helping factors. It takes both a lighted match and a good stack of logs and brush to produce a bonfire. Together a set of factors that are **SUFFICIENT** to produce an effect make up what we shall call 'a complete causal complex'. Each factor in the complete causal complex—for example, the brush or the logs or the lighted match—is **INSUFFICIENT** by itself to produce the effect. Still, each has got to be there or that complex won't produce the effect. That's why the separate factors in the complex are **NECESSARY**.

The whole complex itself however, while sufficient for the effect to occur, is generally **UNNECESSARY**. That's because there are almost always other ways—other 'complete causal complexes'—to produce the same result: One can also make a bonfire with a stack of dry straw and packing cases and a cigarette lighter, or with dry straw, packing cases and a well-aimed bolt of lightning. Each of these different complete causal complexes is sufficient to make the effect occur but none is necessary since each of the other complete complexes will do as well. And each complex contains a number of cooperating factors,

¹⁵That is, all causes are INUS conditions. But not all INUS conditions are causes.

like the lighter or the brush, each one of which is insufficient by itself for the effect but is necessary if the complex of which it is part is to do the job.

INUS conditions are not just a topic for philosophers. Epidemiologists have developed a compelling way to understand INUS conditions with the use of pie graphs to represent sufficient and component causes. Each slice in a given pie represents a component cause and a whole pie represents a sufficient cause, a 'complete causal complex'. A single pie slice on its own is insufficient to cause a disease; the whole pie is needed. So in the philosopher's vocabulary a single pie slice is an INUS condition.

Below are two complete causal complexes for a disease with the component causes shown as pie slices. There are some shared component causes (C1 and C2) and some unique component causes (C4 and C8, for example). Also, we indicate the unknown component causes as CN in the left pie and CM in the right pie.

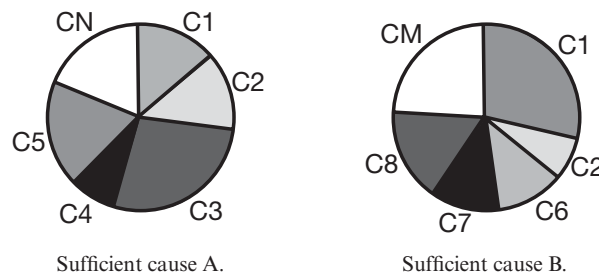


Figure 11.3. Two sufficient causes and their component causes.

Here is an example. Smoking causes lung cancer but not all smokers develop lung cancer. There are other factors, perhaps genetic factors and other life-style and environmental factors, that contribute to developing lung cancer. So in Figure 11.3, Sufficient Cause A would be the constellation of factors, including smoking, that together cause lung cancer; smoking could be C3. But people also develop lung cancer without ever smoking. So in Figure 11.3, Sufficient Cause B would be a constellation of factors not including smoking (C3 is not present) that together cause lung cancer. Working in a coal mine for example could be C8.

Illustrations of INUS conditions

In this section we provide examples from different subjects to illustrate what INUS conditions are and how they work together to produce an effect. The

first is an example about the effectiveness of laws requiring bicycle helmets in reducing head injuries among cyclists.

BICYCLE HELMET EXAMPLE

Vigorous debate regarding the efficacy of bicycle helmets to reduce head injury has been published in the pages of the *British Medical Journal* (See especially Robinson, 2006: 722–5 and numerous letters in response). Case–control studies suggest that cyclists wearing helmets have fewer head injuries than cyclists not wearing helmets, whereas time-series studies in jurisdictions that have passed helmet laws do not show a clear decrease in the rate of head injuries after helmet laws have been implemented and in some cases these studies suggest an *increase* in head injuries after the law is implemented.

At first glance this is paradoxical. Our intuitions, supported with evidence from case–control studies, say that helmets should reduce head injuries whereas helmet compulsion laws fail to show much benefit and in some cases possibly show an *increase* in head injuries.

There are methodological reasons that could partly explain the differences between these studies. A worry about confounders in the case–control studies could exaggerate the estimated efficacy of helmets: There is some evidence suggesting that helmet wearers are overall safer bicycle riders, are involved in less severe accidents, are richer, and more likely to be white. A worry about confounders in the time-series studies could dampen the result of introducing helmet laws: Over the periods of these studies there have been more cars on roads and these cars have increased in size and speed.

Leaving aside a discussion of the methodological quality of case–control studies versus time-series analyses, this paradox can be understood by thinking about INUS conditions. The case–control studies give one piece of a causal pie: Helmets can cause a reduction in head injuries. But those studies don't tell about the other pieces of the pie, that is, other factors that are causally relevant to a cyclist's head injury, things like driver behaviour, cyclist behaviour, and road conditions. Now, there is evidence to suggest that at least some of these things change with helmet wearing.¹⁶ Drivers give less space to cyclists who are wearing a helmet and cyclists take more risks (a 'false sense of security' phenomenon). So helmet compulsion laws don't just change one

¹⁶This naturally suggests that a feedback model as with the A&E study above would be a good one to try if one wants to lay out the steps in the causal process in aid of producing what is called a causal model here.

piece of a causal pie, they change several pieces. And that could partly explain the discordance between the two kinds of studies.

The nice thing about this bicycle example is that it illustrates two lessons at once. First, the importance of identifying the other INUS conditions that go into a complete causal complex, i.e. the other slices in the same pie—which one can think of as ‘helping factors’ necessary in order for the policy lever to work: Helmet wearing in combination with usual driver behaviour will decrease head injuries from bicycle accidents; helmet wearing with more dangerous driving may increase head injuries.

Second, it reminds us that in thinking about INUS conditions we need to pay attention to the unintended consequences of our actions. In implementing a policy we may not only produce unwanted side effects; we can, as in this case and in the Lucas example to be discussed, introduce factors that undermine the effectiveness of the very policy lever we employ. Of course we will always be plagued by uncertainty. We are in no position to predict many of the unintended outcomes of our policies. But some we can predict if only we think about them in the right way.

The failure of the California class-size reduction programme may well be a case in point. The reduction in class-size was rolled out state-wide over a very short time. That necessitated the hurried hire of a large number of new teachers and in consequence teaching quality went down (Bohrnstedt and Stecher, 2002). But teaching quality is a slice of the same pie as small class size: Reducing class size cannot be expected to increase reading scores without the cooperation of good teaching. The point is that this unintended consequence of the policy implementation is the kind that might well be foretold if careful thought is put towards it. So in producing a practicable guide based on the principles here, one will have to figure out ways to remind users to think about the unintended consequences of their policies and implementations and to help them do so.

PHYSICS EXAMPLE

An object of charge q_1 with centre of mass at a distance \mathbf{r}' from the earth's centre is accelerating at a distance \mathbf{r} from a second object of charge q_2 . It is also, of course, subject to the earth's pull. Letting M represent the mass of the earth, the object's acceleration is given by:¹⁷

$$\text{Acc} = \epsilon q_1 q_2 / \mathbf{r}^2 \oplus GM / \mathbf{r}'^2$$

¹⁷ Assuming there are no other forces at work and ignoring the generally negligible gravitational attraction between the two objects themselves.

The first term (the ‘Coulomb acceleration’, $\epsilon q_1 q_2 / r^2$) is **sufficient**—it is enough—to obtain a contribution to the acceleration. But it is **unnecessary**. There are a lot of other causes that can contribute to the acceleration even if the Coulomb force isn’t there. So too with the second term (the ‘acceleration due to gravity’, GM/r'^2): The presence of the earth’s mass a distance r' away is sufficient for producing a contribution to the acceleration but it is not necessary.

Consider next q_1 . Without it there is no Coulomb force. So it is a **necessary** part of the first term. But it is **insufficient** since it cannot produce a contribution to the acceleration on its own but only in consort with another charge (q_2) and some separation (r). The same is true of each of the other factors appearing in the first term, as well as of the factors M and r' in the second term.

The factors q_1 , q_2 , r , M , and r' are all *causes* of the acceleration in anybody’s books. And they are each, as Mackie claims, INUS conditions; each is insufficient but necessary to a causal complex that is sufficient for obtaining a contribution to the acceleration, but no one of these sufficient causal complexes is necessary for a contribution. Moreover, we know the functional form of the relation between the factors.

Functional Form

Merely knowing what the INUS conditions are is less helpful than knowing the formal relationship among the factors—but we need to know what the factors are before we can investigate their relations. In the physics example we know the *full functional form* for the production of acceleration: We know all the possible causes; we know which ones combine together to make a single complex sufficient for producing a contribution; we know the functional form for their mutual relations within the complexes—e.g. the distance in the Coulomb term appears in the denominator and is squared; and we know how the contributions combine to produce an overall effect—by vector addition.

There are standard methods used in the social sciences, and especially in econometrics, for teasing out aspects of the full functional form of the relations between causes and effects, and clearly physics has been very successful at this. That’s ideal for predicting causal counterfactuals. Most often for real policy cases in real time, however, there is little hope for much headway on the full functional form. That is why we have opted to focus on INUS contributions—at least with a reasonable understanding of these one will know what auxiliaries are necessary if the policy variable is to have a hope of being effective. But it is worth having the ideal in mind since it is structurally like the less ideal cases that must be dealt with in social policy.

Some philosophical niceties

DICHOTOMOUS VARIABLES VERSUS CONTRIBUTIONS

Mackie introduced INUS conditions in the context of dichotomous variables, that is, a variable that takes yes/no values. Does the patient survive; does the magnet lift the pin; does the bicyclist sustain a head injury? For Mackie a complete cause is a sufficient condition for an effect in the logician's sense of 'sufficient': The presence of the complete cause implies the presence of the effect. In this case there is no question of how different complete causes combine. If C implies E then C and C' implies E, no matter what C' is. So if any one sufficient condition for an effect is present, the effect is present; adding more makes no difference.

Many of the effects of interest in social policy are not dichotomous however but can take a variety of values, like acceleration in our second example. That is, the variables of interest are multi-valued rather than dichotomous. In these cases each complete causal complex operating on its own will produce some value for the effect. But when they act together the effect will be different from that produced by any one alone. Each affects the value of the outcome but does not determine it. When this happens, one can talk about the *contribution* the complete causal complex makes to the effect, as we did in the physics example. Then the possibilities for the rules about how causes combine multiply. The most obvious are simple addition and subtraction. But there are many other possibilities, as in the vector addition of mechanics or log linear combination prevalent in economics. These are the rules needed for the second component of an ideal causal model of the kind we urge above (*What's a causal model?*, pp. ???-??).

Now that we have made explicit the difference between how causes work together in the case of dichotomous versus multi-valued effects, it is time to tidy up an earlier formulation. We urge 'Causes are INUS conditions'. But we did not say *for what* they are INUS conditions though our language in discussing the examples reveals that there are two different answers. For dichotomous effects a cause is an INUS condition for the existence of the effect. For multi-valued effects causes are INUS conditions for the existence of *a contribution to the effect*. In the physics example for instance where both the Coulomb force and the force of gravity contribute to the acceleration, q_1 and q_2 are both insufficient but necessary parts of a causal complex, $\epsilon q_1 q_2 / r^2$, that is itself sufficient for a contribution to acceleration but not necessary since a contribution to acceleration can come from other sources, like gravity (For more on 'contributions' see *inter alia* Cartwright, 2007 and 2009). Throughout we will continue to use the expression 'INUS condition for effect X' ambiguously to refer to

INUS conditions for the presence of X when X is dichotomous and to refer to INUS conditions for a contribution to X when X is multi-valued.

NOT ALL INUS CONDITIONS ARE CAUSES

Causes, we say, are INUS conditions. Beware. We do not say, 'INUS conditions are causes'. The reason is the well-known problem of spurious correlation. Two factors can be correlated without either causing the other; similarly two factors can be sufficient for each other without either causing the other. Consider a simple case of dichotomous variables, where one factor C causes both E and E', neither of which has any other causes. Then both E and E' occur if and only if C occurs, which implies that E occurs if and only if E' occurs. So E and E' are each sufficient for each other. So we don't claim that all INUS conditions are causes. But we agree with Mackie and other philosophers that causes are INUS conditions, either for their effects or for contributions to the effects.

Why fuss about INUS conditions?

Usually when discussing policy one focuses on a single cause, that is, a single INUS condition. But it is not possible to predict the effect of that cause without considering *all the other INUS conditions and the relations among them*.¹⁸ Thinking in terms of INUS conditions then serves several purposes:

- It focuses attention on the fact that there are usually a number of distinct causal complexes that contribute to the effect. (So one doesn't expect the match to light the logs without the dry brush.)
- It focuses attention on the other factors that are necessary along with the policy variable if the policy is to have any effect at all. (So we don't bother to rent the magnet if the rubbish isn't ferrous.)
- It focuses attention on the functional form of the relations of the variables within a single causal complex. (So we expect that increasing the separation between charges does not increase but decreases the Coulomb acceleration because the separation is in the denominator.)
- It focuses attention on the overall functional form: How do the separate causal complexes combine? (Recall our earlier remark. Often in social contexts one assumes additivity. But that doesn't work in even

¹⁸Sometimes we are only interested in estimating what difference the policy will make and even then sometimes only the direction of change so that we can get by without an estimate of size. For that we clearly need somewhat less information. To be discussed in Part IV.

simple physical cases. The vector addition of classical mechanics is after all a long way from the simple linear (scalar) addition of effect sizes.)

And the notion of INUS *contribution* is useful because it more adequately accounts for the facts that most effect variables, or outcomes of interest, are not dichotomous and that most causal factors themselves contribute to the effect to varying degrees rather than dichotomously.

Two central principles for a theory of use

We now have two assumptions that form the core of a theory of evidence for policy effectiveness:

Principle 1: A good way to evaluate whether a policy will be effective for a targeted outcome is to employ a ‘causal model’ comprising

- A list of causes of the targeted outcome that will be at work when the policy is implemented
- A rule for calculating the resultant effect when these causes operate together.

Principle 2: Causes are INUS conditions.

III. The neglected questions

With these two theoretical principles in place we return to the three issues of quality, relevance and evaluation. If one is to evaluate policy counterfactuals via causal models, as we propose, this imposes criteria of relevance and via that also affects the standards of quality. A causal model, even if rough and approximate, requires a great deal more information than we are in the habit of looking for.

Requisite information for evaluating policy effectiveness: Information is needed about:

- The causal factors that will operate:
 - What factors causally relevant to the targeted outcome are in the situation? This breaks naturally into two questions:
 - What’s there?
 - Is it causally relevant?
 - What factors that are introduced during implementation will be causally relevant? Again this breaks into two questions:
 - What will we do?
 - What factors among those we introduce will be causally relevant?

- How these combine in producing the effect. Here one should pay particular attention to
 - What auxiliary factors are necessary along with the policy variable to produce the targeted effect?
 - How do different factors within a single complex (different segments of the same pie) combine?
 - How do different causal complexes (different pies) combine?

These are empirical questions and any answers that are proposed should have evidence to support them. This sets our criterion of relevance:

An empirical claim is *evidentially relevant* to a policy effectiveness estimate just in case it helps to establish:

- i. What's there in the target situation
- ii. What will be introduced in implementing the policy
- iii. The causal relevance of any of the above factors for the targeted effect
- iv. The method of calculating joint effects.

This formulation does not eliminate questions of relevance; it only pushes them back a level. One still needs to know what kinds of evidence are relevant for establishing what's there, what factors are causally relevant, and for claims of how they combine. The point at the moment is that relevance is a far broader church than the one we are used to practising in. In principle one should have evidence for all the components that need to be used in supporting an effectiveness claim. In practice some facts will be fairly obvious and not need much evidencing; and one will necessarily take a good many shortcuts. But the task for this paper is not to jump into shortcuts but rather to lay a principled foundation for judging policy effectiveness, including evaluating shortcuts and deciding how much to bet on them.

The broad-church relevance criteria in turn affect issues of quality. Most current guides focus on the quality of *efficacy* claims. Depending on context and philosophical leanings, these can be read as claims that the policy can work, or that it does work under specific conditions, or about its average effect in a particular population under special implementations across some range of conditions. Efficacy claims help support the causal relevance of the policy variable, which is part of category iii. The usual ranking schemes police the quality of efficacy claims. But how should the quality of the other kinds of claims needed as evidence for the remainder be policed?

This issue needs to be faced and dealt with, however fallibly, in designing a well-grounded comprehensive advice guide, convenient as it would be to ignore it. Recall our cautions about chains of argument. It is no use having one or two highly certain premises in arguing for or against policy effectiveness. The conclusion can be no more certain than the weakest premise. In adopting a policy,

one is betting, willy-nilly, that all the requisite questions have the right kinds of answers. One can do that on a wing and a prayer. But that is not an evidence-based decision. So it is important to figure out reasonable and usable sets of advice about how to manage the need for evidence and not to institutionalise ignoring the need.

Here is probably where Nancy Cartwright first got into trouble with those who maintain that RCT-backed policies are the only ones with a reasonable evidence base. We are happy to take RCTs as a gold standard—for something. In our view they are provably good at establishing efficacy conclusions, as are a number of other methods, such as deduction from sound theory and certain econometric methods (See ‘Causal Claims: Warranting Them and Using Them’, in Cartwright, 2007). But that is from the point of view of the evidence producer.

Evidence users want to know if a policy will work for them. That, as everyone has really known all along and as we have been stressing here, requires a lot more information than the information supplied by an RCT or a good econometric model that establishes the efficacy of the policy variable; and that information needs evidence, including evidence about what can often be a really tough question—how the causes combine.

Things look very different when one surveys the problem from the user’s point of view from how they do when looked at from the point of view of the scientist charged with producing sound results to offer up as evidence.¹⁹ Imagine we are offered two policies. One has very good RCT evidence in favour of its efficacy but we have very weak ideas and information about what the requisite helping factors and major inhibitors for it are. The second is a policy that comes with a theory that suggests what helping factors are needed—and these are ones that are either in place for us or cheap to put in place. Suppose the theory has some reasonable evidence in its favour and the associated policy has some evidence for efficacy, but not gold standard? Which has stronger evidential support in favour of its claim to be effective if we implement it?

¹⁹It is because we are concerned with evidence users rather than evidence producers that we do not talk of ‘external validity’. External validity starts with a result and asks where outside the experimental context it will obtain. The answer is generally ‘not many places’, especially for RCT results where there is good reason to expect the same result only in situations where the effect has the same set of causal factors and the probabilities over these are the same. The problem for users is not how to use some special nugget of well-established result but rather how to assemble and treat all the evidence that can help with all the issues involved in estimating what will happen in their specific case.

This is a question that depends on the actual details and in many cases there won't be any very good answer. But sometimes normal educated judgement will—and should—reasonably go for the second policy though the evidence for its efficacy is clearly less compelling. That's why we made such an issue at the start of this paper about chains of support, which are only as strong as their weakest link. Adding more rigour at one point can raise the overall probability that the policy will be effective but that increase in probability can be offset by too much guessing later on. We do not have guides that provide enough of the right kind of advice considering all that is required.

It would be wrong of course to suggest that these other issues have not been tackled at all. A lot of hard work and serious thought has been put into what is already available. But much of it is piecemeal, directed at specific problems, starting from specific places in midflow. We need a foundation that considers the problem of evaluating effectiveness of counterfactuals as a whole. It is only on the basis of such a foundation that one will be in a position to judge how reasonable it is to leave out specific considerations, to take specific shortcuts and to make specific heroic assumptions. The theoretical foundation proposed here is meant to do that job. It is not the only one possible but it is a foundation laid specifically with a view that practicable advice needs to be built up from it.

IV. Making life somewhat easier

Perhaps suggesting that we want to provide an advice guide based on the idea of constructing a causal model sounds like a tall order. Sometimes it is, particularly when there is a demand for very precise predictions or predictions that one can be very sure of. But we should not be too frightened of the project. For it is one we are well used to. We regularly build causal models in making decisions in our daily lives as we think through the possible effects of our actions and policies. Consequently the schema should not be seen as too exotic or impractical. It, or something like it, is used all the time.

For example, recently Nancy's favourite red-and-white-striped tee shirt was soiled looking. Should she wash it in hot water? Well: Hot water only works if the shirt has a reasonable amount of cotton in it and it won't work against coffee or ink stains. Even with cotton it can be counterproductive if the hot water makes the stripes run. And she knows that she has to be especially careful in loading a hot wash since the shirt will go grey if some dark socks are inadvertently included. All told, given her cotton shirt with garden

dirt and the determination to be careful in loading the machine, she reckoned (correctly) that the shirt would come out clean in a hot water wash.

This is a homely example but it illustrates our claim that people build what we call ‘causal models’ all the time when making policy decisions. The problem for evidence-based policy is how to use evidence to build them better and to estimate the degree of confidence policy analysts should have in the results of their efforts.

Perhaps you do not find this familiar kind of example comforting. The idea of insisting on causal models stills sounds too daunting. Nevertheless, Nature will use a causal model to decide what outcomes to produce when we implement our policies whether we wish to follow her lead or not. The right answers to the questions of quality and relevance will depend on the models she chooses. So, daunting or not, advice on these questions should reflect that.

We can, however, sometimes make the job less daunting. Consider: We would in general like to be able to predict the actual value of the effect that would follow the implementation of a proposed policy. By *just how much* will household burglaries drop if a community-wide property marking program is adopted? But often that will be difficult because we do not know how to predict what else will be going on. What other causes of burglaries will be in place at the time? Often we cannot assume that the causes will be the same then as they are now. (This is the reason JS Mill said economics cannot be an inductive science.) So we can’t estimate what other ‘sufficient’ causal complexes will be at work contributing to the outcome, let alone what their combined effect will be. In these cases we may be satisfied with reasonable assurance that the policy will produce an improvement in the effect over what would be the case without it, whatever that is. If so, life is somewhat easier.

In this case establishing just a couple of facts will allow us to ignore the other sufficient causal complexes (all the other ‘pies’) and concentrate on those that include the policy variable.²⁰ What we need to know is that no alternative complex of causes will be so dominant that it swamps the policy complex, either positively or negatively, making its effects negligible. For instance, there is no point offering a low cholesterol diet to improve longevity to a man who will be executed in the morning. Nor in installing a fancy electronic lock on Nancy’s old Rover sedan since, her daughter assures her, there is no chance that it will be stolen.

²⁰Complex relations between the sufficient causes are possible however, so sometimes even for these kinds of cases it is not a good idea to ignore other causal complexes. Suppose, for example, that adjusting one component cause of a cluster (one slice of a pie) modifies another component cause of the same cluster—the example about bicycle helmets illustrated this—then, if the secondary modified component is also a component of another cluster, the effect of the second sufficient cluster will be modified.

So . . . if we are content to settle for the claim that the policy will make an improvement on what would otherwise have been the case were the policy not implemented and we have good enough reason to think that nothing will swamp the effects of the policy, then we are justified in focusing just on the policy variable and the factors necessary for it to succeed in producing the targeted effect.

A warning reminder is worth making however. We all know that a successful policy—one that did indeed produce an improvement over what would have been—can easily be judged a failure if it does not produce an improvement over what used to be. Policy consumers are apt to be unimpressed by the claim: ‘Yes things have got worse. But they would have been far worse still if we hadn’t acted as we did’ even if it is true. In these cases one needs to have a good account of what other causes operated to counter the policy effects and good evidence that that is really the correct story.

V. Mechanisms: a principle in aid of practical advice

The primary purpose of the ‘theory of evidence for use’ is to provide principled grounds for practical advice. To this end we propose to borrow one more tenet from our colleagues in philosophy to add to the basic principles of the theory, albeit one more informally put.

Principle 3: Mechanisms matter.

Methodologists like RCTs in part because RCTs provide evidence for causal relations without our having to know the mechanisms by which the cause produces its effect. Policy makers generally share this lack of interest in mechanisms. They are concerned only with whether the policy will produce the targeted results and do not care about the mechanisms that will drive the result. Still, when we want to try to put a cause to work, getting a better understanding of the mechanism can make a big difference. The importance of mechanisms for causal discovery, causal understanding, and causal prediction has been heavily stressed in recent philosophical literature. What though is a mechanism?

Causation is all the rage in philosophy now; mechanisms are centre stage in the discussion. Not surprisingly then there are a wide variety of different characterisations on offer.²¹ Here we are not going to rely on any of these

²¹ We shall describe some of these approaches in order to stress by contrast that none of these are what we mean by ‘mechanism’ here. Here we mean an answer to a ‘how’ question that can help in

(including Cartwright's) since they are generally both too narrow and too abstract to be of help to those non-expert in the sciences. Rather, we make use of an informal notion of mechanism common to many of the formal accounts. This is a notion that can provide a help for policy makers—a prod for the imagination—in identifying the auxiliary factors (the other INUS conditions) that are necessary along with the policy variable to produce the targeted effect. For these purposes we take a mechanism to be an answer to the question:

How would the policy variable bring about the desired effect?

Two different ways of answering can help in finding auxiliary factors:

1 Trace out the causal pathway from policy variable to effect. Seeing what should come next at each step helps focus on what would be required in addition to the policy variable to make the next step happen.

2 Many social results are achieved by calling into play general, often familiar, routine phenomena, such as loyalty, mother-love, fear of punishment, desire to conform, desire to be recognised. Different helping factors will be required besides the policy variable to set different general mechanisms into operation. So recognising which general mechanisms will be called on can be a big help in identifying the necessary auxiliaries.

Tracing the causal pathway: an example from economics

Robert Lucas famously argues that it is generally counterproductive for governments to intervene to regulate the economy on the basis of observed regularities (1976). That's because people will figure out what is happening and act differently, thereby undermining the very regularity the government depends on for predicting the effects of its policies. One of his striking examples is that

finding INUS auxiliaries. As to other senses of mechanism: Judea Pearl explores causal models that take the form of linear equations, one equation for each effect variable on the left-hand-side, laying out a complete set of causes for it on the right-hand-side. Many people call these equations 'mechanisms', as in a simple supply and demand model in economics where the equation for the quantity supplied is said to describe 'the supply mechanism'; that for the quantity demanded, 'the demand mechanism'. Nancy Cartwright (cf. Cartwright, 1999 and 2007) talks about a mechanism (or a 'nomological machine') as a fixed (enough) arrangement of parts that when set running can give rise to stable in-put/out-put relations. For our UCSD colleague William Bechtel, 'A mechanism is a structure performing a function in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena' (Bechtel and Abrahamsen, 2005). Alternatively Peter Machamer, Lindley Darden and Carl Craver (2000) define mechanisms as 'entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions'.

of the Phillips curve, the empirically observed trade-off between inflation and unemployment that was used by policy makers in the 1950s and 1960s to control unemployment via inflation. Lucas uses a ‘rational expectations’ model to show that the Phillips curve will break down if people know what the government is doing. His model reflects a story that answers the question, ‘How does rising inflation produce a lowered rate of unemployment?’ In so doing it unearths some crucial auxiliary factors that have to be in place besides inflation if inflation is to reduce unemployment.

We have seen a version of the Phillips curve already, above, pp. ??–??, in ‘Had we world enough and time’.

$$(*) y_t = \theta\beta[p_t - p_{t-1}] - \theta\beta\pi + y_{pt} + \varepsilon_t.$$

According to this equation an increase in p should make for an increase in output. We can suppose that an increase in output will in turn lead to an increase in employment. Hence the equation describes a trade-off between inflation and unemployment. But it is of no use for policy, says Lucas. His story goes like this: How much output suppliers produce depends on the price they expect their good to sell for and on what they expect their expenses to be. In the Lucas model the average price for goods in the economy serves as a proxy for expense. So in the model the amount of a good supplied in a given period depends on the ratio of the price of the good to the expected economy-wide price for goods in that period. Lucas assumes that suppliers will be good guessers about the economy-wide price: The economy-wide price that they expect is the average economy-wide price that actually obtains. In this case overall output of a good will be proportional to the ratio of the price of the good to the mean of economy-wide prices. So the output of a good will be greater when the price of the good exceeds the mean of prices across the economy. That means that there will be a positive relationship between output and price increase. Another causal process that we won’t describe provides Okun’s law, under which increases in output lead to increases in employment. The two processes together thus imply that rising prices will reduce unemployment.

What happens if the government decides to intervene to increase inflation over what it would have been? Assuming that the Phillips curve (along with Okun’s law) still holds, unemployment should go down. Not so, Lucas argues, because suppliers are good estimators of the effects of the government action on average price. If they know about the government’s actions, they will predict the average price rise that will in fact occur. The expression for output of a good has price for the good in the numerator and, assuming suppliers are good estimators, average price rise in the denominator, recall. So the rise in price suppliers see for their product, which appears in the numerator, will

prompt an increase in output only if it is not offset by the increase in the average prices in the denominator that inflation will entail. Indeed, if the denominator goes up proportionately faster than the numerator, the government policy to increase prices in the economy can even create a drop in output and thereby cause an increase in unemployment.

Where in equation * do we see this important factor—the average of economy-wide prices? It is hidden in θ . But rehearsing the causal process step-by-step, as in the Lucas story, brings it out of hiding. The only way that inflation can increase output is if the average price rise this involves does not result in an increase in the overall price rise expected by suppliers big enough to offset the rise in price the suppliers see for their own products. The trade-off between inflation and unemployment holds when it does because suppliers do not expect the overall rise in prices. Thus the requisite helping factor on the Lucas story—the INUS factor necessary to allow inflation to work its lowering effects on unemployment—is the failure of the suppliers to foretell the inflation. That suggests that if the government is going to succeed in the strategy of encouraging inflation in order to reduce unemployment it had better not let people know that that is what it is doing.

This case illustrates two points of interest here. Equations are nice because they express precise quantitative relationships. Still, true equations may leave a lot out and especially a lot we need to know for policy success. Even equations that are 100% descriptively accurate can fail to lay out all the INUS factors necessary to enable the cause they picture to produce the expected effect. Second, thinking through the causal process step-by-step—answering a *how* question—can make these helping factors apparent.

Identifying the means of production: a criminology example

We quote an example from Nick Tilley (forthcoming) at length to illustrate how thinking about the general mechanisms called into play by the policy variable in order to produce the effect can also help in identifying auxiliary factors:

Take property marking. What is it about it that is expected to ‘work’ as a crime prevention measure? Property marking might increase the risk to offenders by making it more likely that they will be caught with stolen property, successfully prosecuted and punished. This in turn may mean:

- 1 More offenders are incapacitated,
- 2 Some offenders are deterred from future crime,
- 3 And/or other prospective offenders are deterred as they come to appreciate what will happen to them if they try to commit the crime.

Alternatively (or in addition), the perceived increased risk of apprehension, regardless of the reality:

- 4 May lead (some) prospective offenders not to commit crime in the first place.

For property marking to 'work' in relation to any individual offender in the first way,

- (a) Property that is liable to be stolen has to be marked,
- (b) Offenders have to fail to remove or disguise the marks,
- (c) Authorities have to check that property that might be stolen has property marks on it,
- (d) Police have to link the marked property back to those from whom it has been taken,
- (e) Those found with the stolen property have to be unable to cook up a plausible enough story about why they legitimately have it in their possession,
- (f) The prosecutor has to be persuaded that the case is worth taking to court,
- (g) The judge/jury have to be persuaded by the evidence,
- (h) A custodial sentence has to be passed, and
- (i) There have to be offences that the incarcerated person would otherwise be committing but for the fact that he or she is in prison.

For property marking to work in the second way, (a–i) have to be in place, and

- (j) the penalty has to be sufficiently salient that the offender makes decisions that do not lead to further offences or which lead to fewer offences.

For property marking to work in the third way (a–j) have to be in place, and

- (k) Prospective offenders need to know, appreciate and sufficiently fear the penalties applied that they will make decisions not to commit offences that they would otherwise commit.

For property marking to work in the fourth way (a–k) need not be in place, but,

- (l) Prospective offenders must know that property is (or may very likely) be marked
- (m) Prospective offenders must be persuaded that the marking significantly increases their risks of being caught and penalised if they steal the marked goods, and
- (n) The expected penalties must be sufficient to lead them to decide not to commit the offences they would otherwise commit. . . .

Thus, what might work to bring about a crime drop through property marking depends on contextual contingencies.

Tilley's 'contextual contingencies' are just the auxiliary factors we have been talking about in discussing INUS conditions, factors that must be in place along with property marking in order for property marking to bring about a drop in crime. Focusing, as he recommends, on *how* property marking is supposed to achieve these results directs attention to these essential factors.

VI. In sum

Our aim has been to lay the foundations for constructing a comprehensive advice guide for evaluating policy effectiveness claims, a guide that is practicable and at the same time rests on sound general principles. To this end we propose three principles. First, policy effectiveness claims are really causal counterfactuals and the proper evaluation of a causal counterfactual requires a causal model that (i) lays out the causes that will operate and (ii) tells what they produce in combination. Second, causes are INUS conditions, so it is important to review both the different causal complexes that will affect the result (the different pies) and the different components (slices) that are necessary to act together within each complex (or pie) if the targeted result is to be achieved. Third, a good answer to the question ‘How will the policy variable produce the effect’ can help elicit the set of auxiliary factors that must be in place along with the policy variable if the policy variable is to operate successfully.

A guide based on these principles will have to help users construct their own causal models and use evidence to judge how good they are. It should also provide shortcuts, what Gerd Gigerenzer has called ‘cheap heuristics’, that can achieve near enough the same conclusions with less input (Gigerenzer, Todd, and the ABC Research Group, 2000). Most of these will apply only in special conditions. Part of the job before offering them to users will be to show that these shortcuts are indeed good ones in the right circumstances, then to describe the circumstances for the users in a way that can be understood and applied.

All this is something of a tall order for users. That just makes our job hard. We need to do the best we can to help those who need to evaluate effectiveness do so as well as possible, even if the process will inevitably be flawed. Recognising that it will be flawed means making clear that policy effectiveness judgements will almost never be very secure; and so far as possible, one should hedge one’s bets on them. It does not mean giving up on the attempt to construct a causal model, or alternatively defending that a particular short cut will do almost as well. For, as we have stressed, when one bets on an effectiveness counterfactual, one is betting, willy-nilly, on the causal model that underwrites it. The whole point of evidence-based policy is that bets like this should be taken consciously and be as well informed by evidence as is practicable. It’s no good ducking the problem. We’d better just get on with figuring out how to make this as simple and user friendly as possible.

References

- Bechtel, W. and Abrahamsen, A. (2005), 'Explanation: a mechanistic alternative', *Studies in History and Philosophy of the Biological and Biomedical Sciences*, 36: 421–41.
- Bohrnstedt, G. W. and Stecher, B. M. (eds.) (2002), 'What we have learned about class size reduction in California' (Sacramento, CA, California Department of Education).
- Cartwright, N. (1999), *The Dappled World* (Cambridge, Cambridge University Press).
- Cartwright, N. (2007), *Hunting Causes and Using Them* (Cambridge, Cambridge University Press).
- Cartwright, N. (2009), 'Causal laws, policy predictions and the need for genuine powers', in Toby Handfield (ed.), *Dispositions and Causes* (Oxford, Oxford University Press), pp. 127–58.
- Cooper, H., Robinson, J. C. and Patall, E. A. (2006), 'Does homework improve academic achievement? A synthesis of research 1987–2003', *Review of Educational Research*, 76: 1–62.
- Dunn, W. (2003), *Public Policy Analysis* (New York, Prentice Hall).
- Gigerenzer, G., Todd, P. M. and ABC Research Group (2000), *Simple Heuristics that Make Us Smart* (New York, Oxford University Press).
- Kolata, G. (2008), 'New Arena for Testing of Drugs: Real World', *New York Times*, 4 Nov. 2008. Accessible at: <http://www.nytimes.com/2008/11/25/health/research/25trials.html?_r=1>.
- Lane, D. C., Monefeldt, C. and Rosenhead, J. V. (2000), 'Looking in the wrong place for healthcare improvements: a system dynamics study of an accident and emergency department', *Journal of the Operational Research Society*, 5: 518–31.
- Lindblom, C. (1979), *Usable Knowledge* (New Haven, CT, Yale University Press).
- Lucas, R. (1976), 'Econometric Policy Evaluation: A Critique', *Carnegie-Rochester Conference Series on Public Policy*, 1: 19–46.
- Machamer, P., Darden, L. and Craver, C. (2000), 'Thinking about mechanisms', *Philosophy of Science*, 67: 1–25.
- Mackie, J. L. (1965), 'Causes and conditions', *American Philosophical Quarterly*, 2: 245–64.
- Pearl, J. (2000), *Causality: models, reasoning and inference* (Cambridge, Cambridge University Press).
- Reiss, J. (2007), *Error in Economics: towards a more evidence-based methodology* (London, Routledge).
- Robinson, D. L. (2006), 'No clear evidence from countries that have enforced the wearing of helmets', *British Medical Journal*, 332: 722–5; numerous letters in response, available online at: <<http://www.bmj.com/cgi/eletters/332/7543/722-a>>.
- Tilley, N. (forthcoming) 'What's the "what" in "what works"?' Health, policing and crime prevention', in J. Knutsson and N. Tilley (eds.), *Evaluating Crime Reduction. Crime Prevention Studies*, vol. 24 (Monsey, NY, Criminal Justice Press).

