

## Section I.3

### Causal Claims: Warranting Them and Using Them<sup>1</sup>

#### *1. The problem: evidence for use*

Vico reminds us that it is we who have created society so its functioning should be transparent to us. It is natural science, not social science, that should be difficult, perhaps impossible. Why then is social planning and prediction so tricky? We can build and commercially reproduce lasers so precise that complex eye surgeries are routine. But we cannot build a precisely operating secondary school system. What is wrong with our knowledge in the social sciences?

Nothing is wrong with our knowledge in social science, nor with how we ascertain it, I answer. We have a panoply of methods for warranting conclusions in social science that are well tried, well developed and well understood. My hypothesis is that our problems with social policy arise primarily from the fact that we do not know how to use the knowledge we can legitimately claim to have. For good policy we need to know how to predict the consequences of very specific measures as and where they will in fact be implemented. Knowledge, whether in natural or in social science, rarely comes directly in

---

<sup>1</sup> This paper was prepared for the National Research Council's conference on evidence in the social sciences and for social policy, March 2005; and for the Nordic Social Science Conference on the effects of public policy interventions, August 2005. My thanks to Damien Fennell for his help and to the National Science Foundation, the British Academy, the Latsis Foundation and the Center for Health and Wellbeing for support for the research. (The material is based upon work supported by the National Science Foundation under Grant Number 0322579. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the view of the National Science Foundation.)

that form; and the kinds of settings, like the auctions for the airwaves, where perhaps it does are, contra Vico, seldom ones we can (or would wish to) create. In general what we know, different pieces of knowledge of different kinds, often from a vast variety of different sources, must be *brought to bear* on the questions at hand. And here our methodology runs out. We are good at methods for warranting conclusions, but not for using them.

I take it that my primary job at this conference [the Nordic Social Science conference on public policy interventions] is to defend the first part of this claim, and that is what I shall spend the bulk of my time doing, turning to use only at the end. And in keeping with the concentration on knowledge that is likely to be most immediately of use in policy, I shall talk principally about methods for warranting *causal* claims.

## **2. Warrant**

### **Two kinds of methods**

Methods for warranting causal claims fall into two broad categories. There are those that *clinch* the conclusion but are *narrow* in their range of application; and those that merely *vouch for* the conclusion but are *broad* in their range of application.

Derivation from theory falls into the first category, as do randomized clinical trials (*RCTs*), econometric methods and others. What is characteristic of methods in this category is that they are

deductive: *if* they are correctly applied, then if the evidence claims are true, so too will the conclusions be true. That is a huge benefit. But there is an equally huge cost. These methods are concomitantly narrow in scope. The assumptions necessary for their successful application tend to be extremely restrictive and they can only take a very specialized type of evidence as input and special forms of conclusion as output.

Those in the second category – like QCA (qualitative comparative analysis) or methods that stress the importance of the mass and variety of evidence – are more wide-ranging but it cannot be proven that the conclusion is assured by the evidence, either because the method cannot be laid out in a way that lends itself to such a proof or because, by lights of the method itself, the evidence is symptomatic of the conclusion but not sufficient for it. What then is it to *vouch for*? That is hard to say since the relation between evidence and conclusion in these cases is not deductive and I do not think there are any good ‘logics’ of non-deductive confirmation, especially ones that make sense for the great variety of methods we use to provide warrant. I will say a little more about this when I catalogue a number of these methods below.

Interestingly, the method that is by far and away the most favoured by philosophers of science – the hypothetico-deductive method – straddles between these two categories.

### **The straddler: the hypothetico-deductive method**

Since Karl Popper and the Positivists onwards, philosophers of science have taken this to be the method that warrants our most reliable scientific knowledge – the method by which our physics is

tested. From the hypothesis under consideration in conjunction with a number of auxiliary hypotheses we *deduce* some more readily observable consequences. If the predicted consequences do not obtain, the hypothesis – or one of the auxiliaries – must be mistaken. This is a paradigm of a method that clinches the conclusion. If our premises are correct (premise 1:  $h \rightarrow o$ ; premise 2:  $\neg o$ ) our conclusion ( $\neg h$ ) must be correct.

But what if the predicted consequences do obtain? That is the heart of the quarrel between Popper and the Positivists. Popper said we can infer nothing; to infer that the hypothesis is true is to commit the fallacy of affirming the consequent. There is no way for a piece of evidence to distinguish among the indefinitely many hypotheses that entail it. The Positivists – and the bulk of scientific practice – do not agree. They take positive results to confirm the hypothesis to some extent, then look for conditions under which the degree of confirmation would be high; for instance, if the prediction is very surprising, or very precise, or there are a great many such predictions, or the hypothesis itself is very simple, or very unifying, or... But none of this can turn an invalid argument into a valid one and thus provide a method that clinches the conclusion from the evidence.

I stress this because of a peculiar asymmetry. We seem to demand more of social science than of physics. We all admit that physics does pretty well. If my colleagues in philosophy of science are right, physics uses a method that cannot clinch conclusions but only vouch for them. Yet many social scientists want clinchers. I think for instance of econometricians who long for identifiability. That means that, assuming a certain abstract functional form, the probabilities inferred from the data should *entail* the equations of interest. We also see it frequently in discussions backing the demand for RCTs, which, as I discuss below, would be clinchers – if carried out ideally.

Of course in physics there is a rich network of knowledge and a great deal of connectedness so that any one hypothesis will have a large number of different consequences by different routes to which it is answerable. This is generally not true of hypotheses in the social sciences. My worry is that we want to use clinchers so that we can get a result from a small body of evidence rather than tackling the problems of how to handle a large amorphous body of evidence loosely connected with the hypothesis. This would be okay if only it weren't for the down-side of these deductive methods – the conditions under which they can give conclusions at all are very strict.

*An example of the h-d method at work.* We find a nice example of the hypothetico-deductive method for a causal hypothesis in the work of economist Angus Deaton.<sup>2</sup> Deaton (like myself) does not believe in 'off-the-shelf' methodology for causal inference. Nevertheless the following example does fall under the hypothetico-deductive method.

There is a widespread correlation, revealed by different kinds of data from different populations, between low economic status and poor health. Deaton maintains that a primary source of this correlation is a causal arrow from health to income via loss of work. Unhealthy people are unable to work; this lowers their income, which is often used as a marker for status. To confirm this, Deaton looks in the National Longitudinal Mortality Study data, where there is a correlation between both low income and low education on the one hand and mortality on the other. He reasons: If the income-mortality correlation is due primarily to loss of income from poor health, then it should weaken dramatically in the retired population where health will not affect income. It should also be weaker

---

<sup>2</sup> Conversation, November 2004, Center for Health and Wellbeing, Princeton, New Jersey

among women than men, because the former have weaker attachment to the labor force over this period. In both cases it looks as if these predictions are borne out by the data.

Even more importantly, when the data is split between diseases that something can be done about and those that cannot, then income is correlated with mortality from both – just as it would be if causality runs from health to income. Also education is weaker or uncorrelated for the ones that nothing can be done about. It is, he argues, hard to see how this would follow if income and education are both markers for a single concept of socioeconomic status that is causal for health.

Thus the hypothesis that there is a significant causal arrow from health to income-based measures of status implies a number of specific results that seem to be borne out and that would not be expected on dominant alternative hypotheses. So the hypothesis seems to receive some confirmation – though it is very hard to say how much confirmation to award it nor how far beyond the National Longitudinal Mortality Study data set to suppose it will hold. (See Part 3 on problems of exporting causal conclusions from where they are confirmed to where they will be used.)

### **Narrow methods that clinch conclusions**

*Derivation from theory.* This is the second in rank of the philosopher's favorites. We can trust a causal conclusion that is deduced from already well-confirmed theories. This is generally supposed to be a far less useful method in the social sciences than in the natural sciences because we have no really good theories of any kind to begin with. But there are a number of factors that ameliorate this lack.

1) We need not look just to ‘high’ theory, abstractly expressed and systematically organized. For instance, as Naomi Oreskes argues,<sup>3</sup> it would be a mistake to think that we do not know the harmful effects of greenhouse gases just because the results may not be derivable from this or that cutting-edge model. The basic account of radiative transfer involving CO<sub>2</sub> was already established in the 19<sup>th</sup> century, by John Tyndall, and reconfirmed by Plass and others in the 20<sup>th</sup> century.<sup>4</sup> This is not ‘high’ theory– this is no cutting-edge climate model – but it is good science, science that has been known and accepted for a long time, based on physics theory, confirmed by laboratory experiments, etc. No one questions it, not even the climate-change deniers. So, now we go to complex climate models, ‘high theory’ in the sense that it is state-of-the-art, the cutting edge of the discipline. And yes, here we get a case where the details of the outcomes of increased CO<sub>2</sub> are uncertain because of uncertainties about the effects of other forcing functions – aerosols and clouds in particular.

On Oreskes’ account what is going on in this case is a lot of fussing about the details of the predictions, and, especially, about the forecasts for the future, as if one had to forecast the future to a high degree of accuracy to make a policy decision. But the fact is, one often does not need a high degree of accuracy to make policy plans. One simply has to know that the basic science is well established, it has made predictions, and those predictions are indeed coming true – a beautiful example of the hypothetico-deductive method.

2) Then there is ‘common knowledge’. There is a lot that we know as well as we know anything and it is not to be disdained because it does not have the character of a ‘scientific theory’ or is ‘merely’, as Aristotle put it, knowledge of what happens ‘for the most part’. ‘Acorns grow into oak trees’. That is

---

<sup>3</sup> Oreskes, N. (in prep), “The Scientific Consensus on Climate Change: How Do We Know We’re Not Wrong?” to appear in *Climate Change*, edited by Joseph DiMento, under contract to MIT Press.

<sup>4</sup> Fleming, J. (1998), *Historical Perspectives on Climate Change*, Oxford: Oxford University Press.

as certain as any of the surest claims of physics. Common knowledge should not be dismissed just because it is common nor because we know that some of the things taken as common knowledge have turned out to be mistaken. That is characteristic of even our best scientific accounts. Just look to physics' journal articles of the past. You will find a huge number of accounts of physical processes that we no longer hold with, and not just because of big theory changes like Newton to Einstein, but rather because that particular detailed use of the theory for that case has been superseded. The correct strategy is surely to assess the uncertainties of common knowledge, not to lose information by dismissing it; nor to assume that we can duck the problem of assessment by restricting ourselves to more 'scientific' claims since we face equal problems of assessing certainty there.<sup>5</sup>

3) We are often very clever at figuring out how to get a lot out of very little theory. Game theory methods provide one such device. The general theory supposed is exceedingly thin. Agents act so as to maximize their expected utility. Then there are auxiliary hypotheses that may or may not be met in given situations, primarily to the effect that agents can reason well and that they are informed about the structure of the 'options' and the 'pay-offs'. Then we devise specific models to consider specific causal hypotheses.

Does loss of skill among workers during periods of unemployment perpetuate periods of unemployment? One model<sup>6</sup> to test this hypothesis supposes that workers gain utility only from wages and leisure and entrepreneurs only from profit, that job/worker matching occurs in a specific way, that there are just two generations in the labour market, that everyone is hired/rehired at once, etc. In this

---

<sup>5</sup> Though perhaps not equal political problems, since these may often be less associated with differing ideologies.

<sup>6</sup> Pissarides, C. (1992), "Loss of Skill During Unemployment and the Persistence of Unemployment Shocks", *Quarterly Journal of Economics*, 107, 1371-1391.

model, the hypothesis can be *proven* true. So, assuming the theory is correct, we know that the causal claim will hold in any setting ‘sufficiently’ like the one described in the model. We know this with certainty since we can deduce it. The problem is to know what real situations are sufficiently like that in the model. For this we need a different kind of assessment.

Here our rigorous methodology gives out. We have rigid standards for how to ‘test’ results in the model but very little guidance about how to assess where the model results will apply. This is in line with my concerns about ‘evidence for use’ that I stress here.

*Tracing the causal process* (or the ‘mechanisms’) that connects the cause and the effect. This method is not so common in social science as it is in more engineering-related areas, so I will not discuss it here (though it has proven important in various biological and medical studies<sup>7</sup>).

*Probabilistic causality* (Suppes- or Granger-causality) and the concomitant method of *RCTs*. I want to discuss the logic of this method explicitly to underline my dual points: The logic is deductive and the argument structure is exceedingly simple; but the premises are concomitantly exceedingly strong. For both probabilistic Granger- or Suppes- causality and for *RCTs* every possible source of variation of every kind must be controlled if a valid conclusion is to be drawn.

Following the philosopher Patrick Suppes<sup>8</sup> and econometrician Clive Granger<sup>9</sup>, we suppose that for populations picked out by the *right* descriptions  $K_i$ , if  $X$  and  $Y$  are probabilistically dependent and  $X$

---

<sup>7</sup> Cf. Bechtel, W. and Abrahamsen, A. (forthcoming) “Phenomena and mechanisms: Putting the symbolic, connectionist, and dynamical systems debate in broader perspective” in R. Stainton (ed.), *Contemporary Debates in Cognitive Science*, Oxford: Basil Blackwell.

<sup>8</sup> Suppes, P. (1970), *The Probabilistic Theory of Causality*, Atlantic Highlands, NJ: Humanities Press.

precedes Y then X causes Y. If any population P contains such a  $K_i$  as a subpopulation, then X causes Y in P in the sense that for some individuals in P, X will cause Y (in the ‘long run’). This is a standard procedure in the social sciences where we use all ‘other’ known causal factors to stratify a population before looking for correlations in each of the substrata as a sign of causal connections there.

The argument is deductive because of the way the  $K_i$  are supposed to be characterized. Begin from the assumption that if X and Y are probabilistically dependent in a population that must be because of the causal principles operating in that population. (Without this kind of assumption it will never be possible to establish any connection between probabilities and causality.) The trick then is to characterize the  $K_i$  in just the right way to eliminate all possible accounts of a dependency between X and Y other than that X causes Y (there is no correlation in  $K_i$  between X and any ‘other’ causes of Y, there is no ‘selection bias’, etc.). Given that  $K_i$  is specified in this way, if X and Y are probabilistically dependent in population  $K_i$ , there is no possibility left other than the hypothesis that X causes Y.

Of course the epistemic problems are enormous. How are we to know what to include in  $K_i$ ? Sometimes we do know (or think we do). For instance in the Stanford Gravity Probe B experiment to test the general theory of relativity,<sup>10</sup> the environment is so tightly controlled that if we see the predicted precession in the gyroscopes that are now in space we can be fairly confident that nothing else could have caused them than the predicted coupling to relativistic space-time curvature.

---

<sup>9</sup> Cf. Granger, C. W. J. (1980), “Testing for Causality: A Personal View”, *Journal of Economics, Dynamics and Control*, 2, 329-352.

<sup>10</sup> See Cartwright, N. (1989), *Nature’s Capacities and Their Measurement*, Oxford: Clarendon Press, 66-71.

Knowledge of just the right kind is thought to be rare in the social sciences – though we should keep in mind that it is in econometrics where we see this method in use, under the title ‘Granger causality’. Granger causality solves the problem of our ignorance about just what to put in the descriptions  $K_i$  by putting in everything that happens previous to  $X$ . That of course is literally impossible so in the end very specific decisions about the nature of the  $K$ ’s must be made for any application.

One last thing to note about probabilistic/Granger causality is that the deductions are from *probabilities* to causes, not from statistics – i.e., not from summaries of data. So here is yet another source of uncertainty about the premises of the deductions. Not only might we be mistaken about nature of the  $K_i$  for our particular system and about whether or not there can be probabilistic dependencies that have no causal source, we may also be mistaken in inferring probabilities from the data. This is a source of uncertainty that will plague any method that takes population probabilities in the premises. These include not only RCT’s, Bayes-nets methods, invariance methods and methods from econometrics, but any method that looks for necessary or sufficient conditions in a population since these are just a limiting case where conditional probabilities have value 1.

*Randomized controlled trials* are designed to finesse our lack of knowledge about what other reasons might be responsible for probabilistic dependency between a treatment and an outcome. We are all familiar with this methodology so I review it exceedingly briefly. Randomization is supposed to ensure that the ‘other’ causal factors for  $Y$  are distributed equally in the treatment and control groups. Various blindings aim to eliminate other sources of dependency (like selection bias) and to control for factors that randomization misses.

The logic is derivative from that of probabilistic causality: If  $\text{Prob}(Y/X)$  is different in the control group from in the treatment group, it must be different in one of the  $K_i$  subpopulations;<sup>11</sup> and if a probabilistic dependency occurs between  $X$  and  $Y$  in a  $K_i$  subpopulation, then  $X$  must cause  $Y$  in that subpopulation and hence in any larger population of which it is a part. (This does not of course mean that it cannot also be true that  $X$  prevents  $Y$  in some other  $K_j$ ,  $j \neq i$ , and hence prevents  $Y$  in the total population, in the same sense in which it causes  $Y$  in the total population. So, for instance, a drug that tests well in a perfectly conducted *RCT* will definitely be curing some group of the test population but it may simultaneously be killing those in some smaller group.)

As with any deductive method, the conclusion can only be as certain as the premises. The important one here is that by randomizing, blinding and controlling in various ways, other sources of probabilistic dependence have been eliminated or their effects calculated away. We do know some typical problems to watch out for – the placebo effect, experimenter bias, and the like. But what might actually confound results in a given case requires a close and intelligent look. Confidence in the results requires that somebody knows a lot about the specific populations involved and the procedures throughout. I notice that people sometimes talk as if there is a formula for how to proceed and if we just follow it the results will be reliable. But, as with all methods, there is no avoiding the need for a great deal of good judgment, sound detailed knowledge and good sense in conducting an *RCT*.

*Controlled experiments; natural experiments.* The logic here is familiar. In principle we control so tightly that when the predicted outcome obtains, nothing but the hypothesized cause could have brought it about. It is commonplace to remark on how hard it is to do experiments in social science.

---

<sup>11</sup> This is guaranteed by the fact that  $X$  and  $K_i$  will not be dependent in an ideal *RCT*.

But sometimes we have the good luck to find a situation in which the controls occur naturally, without our contrivance. There has been a recent push to look hard for these in order to draw causal conclusions in economics.<sup>12</sup> As with any deductive method, the results for either natural or contrived controlled experiments can only be as sure as our assumptions.

*Bayes-nets methods.* Bayes-nets are graphs representing probabilistic independencies. Add some assumptions about the relations between probabilistic dependence and causality and we can use them to infer new causal relations from known causal relations and facts about probabilities – probabilities as they occur in the population under study, not experimental probabilities. The methods will produce every set of causal relations among the variables under consideration that is compatible with the input information and the background assumptions.<sup>13</sup>

As with the probabilistic theory of causality, Bayes-nets methods suppose that two factors will be probabilistically dependent once the ‘right’ background factors are held fixed if and only if they are related as cause and effect when those background factors obtain. This immediately restricts applicability; for instance the methods cannot be relied on in populations where there is ‘selection bias’ for joint effects. They also suppose that causes and effects will be dependent simpliciter, thus ruling out that the positive and negative influences of a given factor via different routes can cancel. There is in addition a kind of minimality or simplicity assumption. Importantly, as with most econometric methods for causal inference, these methods will only apply to variable sets for which the input variables (those not caused by any of the variables in the set under consideration) are all independent,

---

<sup>12</sup> Cf. Card, D. and Krueger, A. (1997) *Myth and Measurement: The New Economics of the Minimum Wage*, Princeton: Princeton University Press. See also Hamilton, J. (1997) “Measuring the Liquidity Effect”, *American Economic Review*, 71, 1, 80-97.

<sup>13</sup> Cf. Spirtes, P., Glymour, C. and Scheines, R. (1993), *Causation, Prediction and Search*, New York: Springer-Verlag.

which is a considerable restriction. Finally, they tell whether or not a factor is causally relevant but nothing about the strength of relevance or the functional form. (This matter is addressed in the two following methods.)

*Econometric methods.* Econometrics has well-developed ‘structural’ methods that allow the deduction of the strength of causal connections between factors in a preselected variable set, provided stringent conditions are met. These methods begin by assuming that a particular set of functional forms correctly represents the causal structure generating the observed data. What is to be discovered are the parameters that turn these functional forms into real functions – roughly, one function for each effect, where any factor that appears with a non-zero parameter on the right-hand-side is judged to be a cause of that effect, with the parameter giving the strength of causal influence.<sup>14</sup>

In addition to assuming that the abstract functional forms are the right forms – the ones the causal principles at work actually have, the causal principles we aim to discover are also taken to meet what are called ‘identification conditions’. These require that there not be too many causal connections between the factors of interest, which is necessary for disentangling the different causal connections from the observed data. Another important condition is that the factors taken as inputs (not caused by any factor in the preselected variable set) be probabilistically independent and also that they not restrict each other’s values. This is required to guarantee that the observed data does not result from hidden common causal relationships between factors not modelled explicitly by the functional relations. Finally, statistical conditions must also be met so that the observed data sample does not

---

<sup>14</sup> Just how these relations represent causal structure is set out by Herbert Simon and is further described in Section III.3 in this book and by Damien Fennell. See Simon, H. (1953), “Causal Ordering and Identifiability”, *Models of Man*, New York: Wiley and chapter 1 in Fennell, D. (2005), *A Philosophical Analysis of Causality in Econometrics*, PhD dissertation, University of London.

‘accidentally’ misrepresent the underlying data generating processes. If all of these conditions are met, then one can deduce the strength of the causal connections between factors in the variable set of interest.

Here, as with the other narrow-clinching methods, secure conclusions are bought at the price of stringent conditions that are difficult to meet. In these structural methods, one must know the functional form of the causal structure and that structure must not be too dense. Such conditions alone are very demanding and without them it is not clear what follows from the observed data.

*Invariance methods.* There is a correlation between a fall in a barometer and a storm coming. But if we manipulate the barometer in arbitrary ways (ways that vary independently from the ‘other’ causes of a storm), for example by smashing it, the correlation will break down. For some nice kinds of systems,<sup>15</sup> given a sufficiently careful formulation of what we mean by *invariant*, we can prove that a functional relation – one we suppose we have observed to be true, say – will represent a true causal relation just in case it is invariant under all arbitrary variations of the dependent variables.

### **Broad methods that ‘vouch for’ conclusions**

The advantage to the deductive methods that clinch their conclusions is that we know exactly what we would have to do using those methods to become more certain about the conclusions – get more

---

<sup>15</sup> For example, for linear equations where the dependent variables can take any combination of values together in systems where any true functional dependencies must result from underlying causal laws. This last is analogous to the assumption required for probabilistic causality, that all probabilistic dependencies arise from underlying causal laws. See Section II.6 in this book for details.

certain about the premises. Often we don't know how to do that; worse, frequently we know the premises are false, or probably false. These are of course problems for any kinds of methods, but they are especially severe for the deductive methods because the requisite premises are so demanding that we cannot expect them to obtain generally. How do we know, for an RCT or Granger causality for instance, that *all* other sources of probabilistic dependence have been randomized over or controlled for and how do we know that we are studying a systems where all dependencies are due to causal connections?<sup>16</sup>

Here we must be careful to avoid a logical mistake. If the premises of a deductive argument are true, the conclusion must be true. What if we do not know they are true but are only willing to assign a probability to them? If we assign a probability of say 90% to the premises taken jointly and we do not know anything else relevant, then it is reasonable to assign a probability of 90% to the conclusion. That however is very different from the case where we are fairly certain, may even take ourselves to know, 9 out of 10 of the premises, but have strong reason to deny the 10<sup>th</sup>. In that case the method can make us no more certain of the conclusion than we are of that doubtful premise. Deductions can take us from truths to truths but once there is one false premise, they can't do anything at all. That is why we need to take seriously non-deductive methods. I'll review a few of these that I have worked with and try to look at what the relation of evidence to conclusion might be in each case. I will spend a little more time on the first case to exhibit the difficult in laying out what the relation really is.

---

<sup>16</sup> Indeed we know this is frequently not the case since many factors are temporally correlated with no causal connection, so at least we had better 'detrend' data before we begin to apply the methods.

*Qualitative comparative methods (QCA)*.<sup>17</sup> This method starts from what philosophers, following J.L. Mackie,<sup>18</sup> call the *INUS* account of causation, which acknowledges both that what we usually call a cause (like  $C_{ij}$  in the formula below) is usually only a part of a total cause sufficient for the effect and that most effects have multiple separate causes. On this account causes are *insufficient* but *necessary* parts of *unnecessary* but *sufficient* conditions:

INUS condition:  $E \equiv C_{11}C_{12} \dots C_{1n} \vee C_{21}C_{22} \dots C_{2m} \vee \dots \vee C_{k1}C_{k2} \dots C_{kr}$ .<sup>19</sup>

So to discover the causes of an effect  $E$  in a given population, sample the population, then look for factors that make a formula of INUS form true in that sample. (These methods are sometimes called ‘Boolean algebra methods’ because they employ huge truth tables to determine the INUS formula.) This raises the problem of statistical inference that I noted with respect to methods that move from probabilities to causes. Results in the sample may not be true of what would occur in the population as the population gets increasingly bigger.

Even in the most ideal uses, however, the method cannot clinch the results because INUS conditions are not causes. The INUS formula represents an association of features, a correlation, and we know that correlations may well be spurious. Consider for example a situation in which the following are the correct causal principles:

---

<sup>17</sup> See Ragin, C. (1998), “The Logic of Qualitative Comparative Analysis”, *International Review of Social History*, 43, supplement 6, Dec, 105-124. See also Lieberman, Stanley. (1992) "Small N's and Big Conclusions: An Examination of the Reasoning in Comparative Studies Based on a Small Number of Cases", 105-118, in Ragin, C. and Becker, H. (eds.) *What is a Case? Exploring the Foundations of Social Inquiry*, New York: Cambridge University Press.

<sup>18</sup> Mackie, J.L., (1974), *The Cement of the Universe*, Oxford: Oxford University Press.

<sup>19</sup> Here  $\vee$  means ‘or’.

$$X_2 \equiv AX_1 \vee W$$

$$X_3 \equiv BX_1 \vee V.$$

If these are true, so too will be

$$X_3 \equiv BX_2 \neg W \vee BX_1 \neg A \vee BX_1 AW \vee V.$$

Thus  $X_2$  is an INUS condition for  $X_3$  though not a cause of it.

Suppose that we know that a given factor is an INUS condition for another, and that is all we know. Does that provide warrant for the conclusion that the first is a cause of the second; if so, how much warrant? It is not unreasonable to suppose that if a factor is a cause of another it will be an INUS condition for it; but there are many other reasons as well why it might be an INUS condition. This is just the quandary I described with the hypothetico-deductive method and it has no straightforward resolution.

Though comparative qualitative analyses cannot clinch a result, they have many advantages over various deductive methods. By contrast with RCTs and Bayes-nets methods, a QCA result is not just a yes-no verdict – ‘yes, the factor is a cause’, ‘no, it is not’. Rather we learn the functional form of the causes. With this method we can learn that the cause is a cause for some individuals and not for others and the method is geared to determine which. Concomitantly, it is difficult to apply because it needs a

complete set of causes – there is no way within the method to deal with ‘omitted’/‘unknown’ factors as there is with econometric methods.<sup>20</sup>

Also, although it is not formally part of the method, the fact that we must look in detail at the individuals in the population for factors that will make up an INUS formula has great side advantages. In the first place it can alert us to a better more concrete reformulation of the effect of interest. Very often what we aim for as an effect is something very general – improved educational attainment, better attitude, more ability to function in a job. We must operationalize these one way or another to get any study going. Looking at cases in detail often shows that our operationalization is wrong, too narrow, leaves things out, misses the mark. In the second place the choice of possible causes is more readily adjusted to the specifics of the cases at hand. The variables in the study are less likely to be standardized and hence have more flexibility to replicate the correct details of the causal stories for the individuals in the population.

*Reasoning from models and model systems.* Another broad method for providing support for causal conclusions is by establishing results in a model, then reasoning from the model results to claims about the target situations. The kind of reasoning employed is as widespread and diverse as the different kinds of models used. These vary from highly concrete models, such as actual physical systems – rats or toy airplanes or prototypes – to computer simulations to extremely abstract models, such as thought experiments.

---

<sup>20</sup> Though in econometric methods, as I noted, we often have to make very strong exogeneity assumptions about the omitted factors.

This method is used widely throughout the social and political sciences. Evolutionary models, for instance, are used to account for higher murder rates among young men<sup>21</sup> or for the (currently topical) divergence in mathematical achievement between women and men.<sup>22</sup> Economics and political theory are rife with game theoretical models, where relatively simple premises provide persuasive hypotheses about the factors that may be driving complex phenomena. For instance,<sup>23</sup> Schelling's model shows how segregated neighbourhoods can arise even if the individuals in those neighbourhoods individually prefer integrated neighbourhoods and Akerlof's influential 'lemons' model from microeconomics shows how asymmetric information can lead to overpayment for used cars. In social psychology, ethological models are used to generate plausible hypotheses about causes of human behaviour. In medicine we use real concrete model systems, like rats. And computer simulations are gaining popularity everywhere.

In reasoning from models and model systems, two distinct questions about warrant must be answered. First, how warranted is the causal conclusion in the model? Second, how does the model conclusion provide warrant for causal claims outside the model? The first is a question of *internal validity* of the kind I have been considering throughout. It gets a huge variety of answers. In game theory models, the results in the model should be certain – they follow deductively. Not so in the evolutionary models where the theory is not tight enough to entail conclusions. For real model systems we have available the whole panoply of methods that we have already reviewed. The second is a question of *external*

---

<sup>21</sup> Daly M., Wilson M. (1999) "An Evolutionary Psychological Perspective On Homicide" in D. Smith & M. Zahn, eds. *Homicide Studies: A Sourcebook of Social Research*, 58-71. Available at: <http://psych.mcmaster.ca/dalywilson/chapter5.pdf>.

<sup>22</sup> Geary, D. C. (1996) "Sexual Selection And Sex Differences In Mathematical Abilities", *Behavioral and Brain Sciences*, 19, 229-284. Available at: <http://www.missouri.edu/~psycorie/GearyBBS96.htm>.

<sup>23</sup> These are discussed in Sugden, R. (2000) "Credible Worlds: The Status Of Theoretical Models In Economics", *Journal of Economic Methodology*, 7, 1-31.

*validity*, which faces all methods since we seldom establish results in the very population and in the very situation in which we want to apply them. I turn to it when I take up issues of use.<sup>24</sup>

*Ethnographic methods.* I shall not review these since there was a separate review of them made for the National Research Council conference.

*Mixed indirect support.* Consider Jean Perrin's influential arguments for the existence of atoms.<sup>25</sup> Atoms were indicated by a large number of different kinds of studies involving different methods, in different places, with different materials, etc. Assumptions about exactly what an atom is or exactly how it behaves were not univocal across these studies.<sup>26</sup> Nor were any of the studies entirely satisfactory in themselves; they were almost all flawed in one way or another. Nevertheless, Perrin reasoned, atoms must exist. It would be too improbable that the different flaws in all these different kinds of studies worked out in just the right way to give the same mistaken conclusion.

Consider an analogous case in the human sciences. Recall the discussion of health and status above. (See *An example of the h-d method at work.*) Michael Marmot<sup>27</sup> argues that the stress induced by low status, particularly by social isolation and high demand/low autonomy work, causes poor health. He marshals a great amount of different kinds of evidence to support the conclusion, like long-term

---

<sup>24</sup> For a more detailed discussion of internal validity in economic models see Section III.5 in this book.

<sup>25</sup> Discussed in Salmon, W. C. (1984), *Scientific Explanation and the Causal Structure of the World*, Princeton: Princeton University Press.

<sup>26</sup> Peter Galison argues that this is characteristic of contemporary physics theories. Different groups, especially theory versus experimental groups, seldom have the same interpretation for what on the face of it looks to be the same claim. See Galison, P. (1997), *Image and Logic: a Material Culture of Microphysics*, Chicago: Chicago University Press.

<sup>27</sup> Marmot M. (2004), *Status Syndrome: How Your Social Standing Directly Affects Your Health and Life Expectancy*, London: Bloomsbury.

longitudinal studies of Whitehall civil servants, experiments on primates, statistical studies of the correlation between income and health in various places, facts and statistics about the health failure in Russia, medical studies of the relations between physiological stress markers and various health problems, and psychological studies of the relationship between ‘stressful’ tasks and physiological stress markers.

Should this body of evidence be judged convincing? Recall that breadth, variety, precision and novelty of evidence are at the core of warrant on our most standard philosophic account of scientific method. On the hypothetico-deductive account, we look for evidence that should obtain if the hypothesis were true – then we demand that there be a lot of it, sufficiently varied, novel, and not easily accounted for by other hypotheses. Figuring out if this is the case for Marmot’s hypothesis – or for any hypothesis – is not easy, and it cannot be done by formula. But doing so is far more realistic than looking for some single study that could clinch a causal hypothesis like this.

### **3. Use**

If warranting causal claims is a difficult matter, judging how we can put them to use is even more difficult. For it is a different enterprise altogether, requiring a different set of considerations, different kinds of background knowledge and different procedures – and these are generally far less well understood, less well articulated and less rigorous. I shall point to some of the central problems.

### **What claim has been established?**

When we want to put our claims to use, it is essential to know exactly what claim it is that has been warranted. It is useful to think in terms of two different problems: the claim itself – what actually is established by a given method; and its scope – for what populations and in what situations is the result established.

*What is the claim?* First, different methods will warrant causal claims of different forms. For instance, RCTs tell about the overall effect of a cause, averaged across subpopulations in which it may behave differently – indeed oppositely. (A drug that cures one part of the population may kill another.) Other methods require more information to apply but give more specific information. For instance, Granger causality tells what happens in each of the relevant subpopulations. What it says is that in those subpopulations (in the ‘long run’) the cause will produce the hypothesized effects in at least some individuals and it should produce opposite effects in none. There are also well-known variations where we learn not about increased numbers of outcomes but increased levels or perhaps increases in the mean. Econometric methods give the full functional form for the relation between a set of causes and their effect; QCA also gives the full functional form, but only for yes-no variables. These are matters that we need to be alert to when we think of using the results to support policy.

A second – and age old – problem is in deciding on the concepts to use to describe both a policy and its putative evidence. Consider an example from the natural sciences. Bodies that are not acted on by forces travel on geodesics – straight lines. But what is a straight line depends on the geometry of the surface. So suppose an experiment is performed on sphere. The body moves in a great circle. If we

take this result to be good evidence for the claim: 'Bodies that are not acted on by forces move in circles', we can go far astray if the application in mind is for a flat table top.

This example illustrates that it is not always a good idea to express the conclusion that a piece of evidence is taken to warrant in too narrow or too concrete a way. On the other hand it is equally dangerous to follow the opposite strategy. The sometimes disastrous effects of overgeneralizing are well-known. But also, expressing results in too abstract a vocabulary can render them almost useless, especially in social sciences where bridge laws that provide concrete interpretations of abstract concepts (like 'unemployment', 'abuse', 'incentives' etc.) are scarce. 'Love thy neighbor as thyself.' Perhaps that is good advice but what does loving one's neighbor amount to in this or that specific situation?

*What is the scope of the claim?* Evidence is always collected in some population in some circumstances. With most methods the inferences that are licensed from that method are tied to the populations and situations in which the evidence is obtained and license to go beyond those must come from somewhere outside that method.

Consider an ideal controlled experiment. It can tell us with certainty what the effect of a given cause is – in the circumstances of the experiment. But in order to do so, the circumstances of the experiment must be extremely unusual ones. What follows with 'certainty' from an ideally carried out experiment is what the cause does there, in those very unusual circumstances. The method itself tells us nothing about what the cause does elsewhere. Often the point of a controlled experiment is to establish what J.

S. Mill called a *tendency law*.<sup>28</sup> These do not tell us what effect occurs when the cause is present but rather what the cause contributes to an effect in more realistic circumstances where other causally relevant factors have not been eliminated. (An example is Coulomb's law for the force exerted on one charge by another. This is never really the force a charged particle experiences in the neighbourhood of another charge since gravitational attraction will always contribute as well.)

We need three different kinds of considerations then before a causal claim from a controlled experiment can be put to use. 1) Is the experiment set up in such a way that we can conclude what we are supposed to be able to – that in the experimental situation the causal hypothesis is true? 2) Is this the kind of causal relation for which we are entitled to think there is a tendency law? On what grounds? 3) Supposing we do have a tendency law. How do we reckon what will happen in any real situation where the cause operates? For the tendency laws governing forces, we have vector addition to calculate what happens when tendencies operate together. What do we have to do the job for us in particular cases in the social sciences?

The point I want to stress is that the method of the controlled experiment, which can clinch an answer about a causal hypothesis in an experimental setting, goes no way to answering questions of the second and third type. For the most part, we have no serious methodology for answering those kinds of questions, and certainly not methodologies that can be articulated and defended with the rigor with which we can treat the methods for warranting causal claims. When it comes to putting scientific claims to use, we quickly fall back on loose argument and intuition.

---

<sup>28</sup> I discuss tendency laws in economics further in Section III.5. For cautions about drawing tendency conclusions from controlled experiments and thought experiments see Reiss, J. (2002) "Causal Inference in the Abstract or Seven Myths About Thought Experiments", *Causality: Metaphysics and Methods Technical Reports* CTR 03/02, CPNSS, LSE. See also Alexandrova, A. (2005) "Applying Economic Theory: From Models to Institutions", unpublished manuscript, University of California, San Diego.

The issue of external validity is no less problematic for other methods. In an RCT if the population under study is ‘representative’ of the target population, then the results of the experiment can be extrapolated from the experimental to the target population. Here at least if we ‘sample’ from the target, we have good statistical guidelines for how to get a representative population. Of course we often are not able to sample from the target population, or if we can, not able to do so in the correct way. The same holds for QCA and econometric methods. Reasoning from sample models and sample systems is even more difficult. What lessons exactly are we to take away from Schelling’s model about any real case of segregation? As in the case of controlled experiments, with all of these methods, rigor gives out when we try to justify exporting results from the populations and situations in which they are established. But if we cannot export results, they are of little use.

### **Are results stable under interventions?**

Knowing the scope across which a result is true tells us where we can use that result for prediction.<sup>29</sup> But policy is more complicated. Policy involves changes: manipulating causes in hopes of producing the concomitant effects, changing them in ways they do not naturally vary as the system works on its own. Change is dangerous since we do not always know exactly what we are doing when we decide to manipulate a cause. In particular, our actions can undermine the very structure that gives rise to the causal principles we rely on to predict the outcomes of our actions.

---

<sup>29</sup> We could of course think of the problems described here and in the next section as problems of the scope of a claim. But I think it is useful to divide the issues in this way since the source of the problems of scope is different in the different cases.

Social scientists talk about one aspect of this problem under the heading *reflexivity*: people change in response to the way we study them, the way they conceive themselves, or in reaction to what they suspect will happen to them.<sup>30</sup>

Another aspect does not necessarily rely on the responsiveness of self-conscious agents but can arise whenever the causal principles we trust in depend on some ‘deeper’ ‘underlying’ structure. If a set of causal principles derives from some more fundamental set, then, when we change the way a cause is brought about – we bring it about in some new way by our policies – we cannot but change the underlying structure and we might well do so in a way that undermines the very principle we are trusting to for our predictions. This is a continuing theme in economics. It is the reason J. S. Mill argued that economics cannot be an inductive science;<sup>31</sup> econometricians have worried about it from the start;<sup>32</sup> and it is the basis for the famous ‘Lucas critique’ of macroeconomics and one of the central Chicago School arguments against government policy interventions.<sup>33</sup>

As before, the methods for warranting claims of stability-under-interventions for a causal connection are very different from those that warrant the causal claims themselves; and they are less well articulated, less well understood and less rigorous.

---

<sup>30</sup> For instance, see Finlay, L. and Gough, B., eds. (2003), *Reflexivity: A Practical Guide for Researchers in Health and Social Sciences*, Oxford: Blackwell.

<sup>31</sup> Mill, J.S. (1836), “On the Definition of Political Economy and the Method of Philosophical Investigation Proper to It”, repr. in *Collected Works of John Stuart Mill*, vol. 4, Toronto: Toronto University Press.

<sup>32</sup> Cf. articles by Ragnar Frisch or Trygve Haavelmo in Hendry, D. and Morgan, M. (1995), *The Foundations of Econometric Analysis*, Cambridge: Cambridge University Press.

<sup>33</sup> Lucas, R. (1976), “Econometric Policy Evaluation: A Critique”, *Carnegie-Rochester Conference Series on Public Policy: The Phillips Curve and Labor Markets*, I, 19-46.

### **Where details matter**

There may be good evidence for the effectiveness of a policy conceived, as we usually do, in the abstract, but the actual outcomes may depend crucially on the fine tuning of the method of implementation. Recall the case of laser engineering, mentioned at the beginning, and consider the early stages of development. There was a great deal of evidence, both theoretical and experimental, that ‘inverted’ populations of atoms properly triggered can produce highly coherent light. But we know that the results – what actually happens – depend hugely on exactly on how the laser is engineered.

Or consider poverty measures.<sup>34</sup> Policy may set whether a poverty line should be relative or absolute and if relative, in what way (for instance, two thirds of the median income). But the results – for instance, the poverty ranking among European countries – depend crucially on dozens and dozens of details of implementation (how to deal with individuals versus families, wealth or welfare benefits versus earned income, etc.), details where it seems very different decisions can be equally motivated and the rankings will come out very differently depending on how these decisions are taken.<sup>35</sup>

The more the details matter, the more the problems of evidence multiply. Naturally more evidence is needed to judge the consequences of taking a decision one way rather than another. But also, it is unlikely that there will be much evidence direct evidence to hand

---

<sup>34</sup> See Atkinson, A. B. (1987), “On the Measurement of Poverty”, *Econometrica*, 55, 4, 749-764 and Atkinson, A. B. (1998), *Poverty in Europe*, Oxford: Blackwell.

<sup>35</sup> Though recall, sometimes the opposite is true, as for instance in the case of climate change discussed above.

since it needs to be considered, not in the abstract, but each decision in the context of the overall proposal, where the consequences of any one decision will depend on what details are supposed already to be in place. This can put severe limitations on how many alternatives can be rationally considered since working through the evidence for any one is difficult and costly. In situations like this it is important to have as good a general understanding as possible in order to make an intelligent selection of which alternatives to explore to begin with.

### **Counterfactuals and causal models**

Most of our warranted causal information comes in pieces. But what we need for policy is the whole picture. We want to know what would happen if various proposed policies were implemented, and implemented in the way they would actually get implemented: what will result from the cause and from its method of implementation, where both are subject to the action of the other causes and interferences that will occur; and not just what happens with respect to the effect in question – we need to know about harmful and beneficial side effects as well. So we need more than piecemeal knowledge of what causes what. We need a causal model.<sup>36</sup> Again, our methodologies for how to construct causal models for new target situations from even highly stable well-warranted causal claims are very poor.

---

<sup>36</sup> See Section III.6 in this book as well as Reiss, J. and Cartwright, N., (2004) “Uncertainty in Econometrics: Evaluating Policy Counterfactuals” in Mooslechner, P., Schuberth H. and Schürtz, M. (eds.), *The Role of Truth and Accountability in Policy Advice*, Cheltenham: Edward Elgar.

#### 4. Conclusion

We do have good methods for warranting knowledge claims in the social sciences. The more secure they make the conclusion, though, the more background knowledge we must have in order to apply them. So social science is hard, but not impossible. Nor should that be surprising; natural science is exceedingly hard and it does not confront so many problems as social science – problems of complexity, of reflexivity, of lack of control. Moreover the natural sciences more or less choose the problems they will solve but the social sciences are asked to solve the problems that policy throws up. And here I think we do find special problems for social science. We have very good methods for gathering social science knowledge but considerably less good advice about how to put it to use. So, I urge, what we most need to study, is now how to do social science *but how to use it*.