Causation:

An Opinionated Introduction[*]

Julian Reiss

Department of Philosophy
Erasmus University
P.O. Box 1738
3000 DR Rotterdam
The Netherlands

reiss@fwb.eur.nl

draft, November 2007

[*] Work on this manuscript was conducted under the AHRB project *Causality: Metaphysics and Methods*. I am very grateful to the Arts and Humanities Research Board for funding. Thanks to Roman Frigg and Nancy Cartwright for comments on an earlier draft.

Part I: Background and History

1 Introduction

The aim of this introductory text is to provide the reader with the necessary background and conceptual tools to follow and develop the contemporary debate on the topic of causality and causal inference. Contemporary research in this area is highly technical. But taking into account that it has been influenced by developments in sciences such as physics, statistics, epidemiology, artificial intelligence and econometrics as well as philosophy, this is hardly surprising. The degree of technical sophistication is not the only aspect that makes the debate hard to follow. Since many of the issues involved in the work on causality have been debated for many years (some of them, depending on how narrow one wants to define the issues, for centuries and even millennia), the positions defended in the exchange are very refined, and arguments often concern minute details of these positions. This, too, provides a good reason to write a text that is aimed at surveying the various positions in the debate.

So what is this debate about? Causal claims are ubiquitous in science as well as everyday language. To give a number of examples that will be discussed in detail later in the book: smoking causes lung cancer, seatbelts save lives, non-borrowed reserves cause interest rates, asteroids caused the extinction of dinosaurs, the collision with the branch of the tree caused the golf ball to hit the hole in one. There is a great number of issues involved in trying to understand these claims. I want to divide them in two groups. On the one hand, there are a number of primarily philosophical problems concerning the metaphysics and epistemology of causal relations and the semantics of causal claims. Here questions arise such as: what is it in the world that relates cause and effect and thus makes the word "cause" applicable to these situations? Are they, in principle, knowable by us? Do we call one thing cause and the other effect because things of their kind follow one another regularly? Or does the effect depend on the cause in a particular way? Does the causal relation instantiate a universal, that is, a general and abstract property? What kinds of things are related as cause and effect? Are causings observable? Is the concept of causation innate?

On the other hand, there is a more practical, methodological bundle of issues: how do we learn about causal relationships? Are experiments the "gold standard" to establish causal claims? Do all experiments work in the same way? How should experiments be designed? Are interventions the only way to test causal claims? Or are there non-experimental, observational alternatives?

Some of the issues from each group are related to one another in more or less obvious ways. If there is no such thing as causality in the world (as some believe), then any attempt to find out about them must remain futile. On the other hand (as others believe), if causal relations are not knowable by us even in principle, any attempt to speculate about their nature is futile. But there are less fundamental connections as well. For example, it matters a great deal for methodology whether or not determinism holds. If, for instance, all phenomena ultimately are brought about by necessary and sufficient causal conditions, and the probabilistic form of our claims is a mere reflection of our uncertainty and not of an underlying indeterminism, it makes sense to invest in the search for additional causes, which eventually will invariably be followed by their effect. But if, by contrast, the probabilistic form of our claims is a reflection of the underlying indeterminism of the causal relations itself, such a search is bound to be fruitless. Thus metaphysics matters for methodology. But methodology matters for metaphysics, too. At least as philosophers inclined towards empiricism, claims about the nature of things should be made on the basis of our experience of them. Successful empirical methods thus often give us hints about what the underlying structure of things could look like.

The aim of this book is to trace the development of the two topical strands—the philosophical and the methodological—from Francis Bacon to this day. Bacon may sound slightly unconventional as a starting point since most discussions of causality begin with David Hume (though sometimes a short mention of Aristotle's four types of causes precedes the discussion of Hume). The reason is that in these histories the (in some sense) deeper metaphysical and epistemological problem set is given priority to the (in that sense) shallower methodological problem set. And the deeper problem begins with Hume. But the shallower problem begins with Bacon.

Let me briefly explain what I mean by these remarks. Hume defended an empiricism of a certain kind. For empiricists in general, experience plays a prominent role, in particular with respect to how terms receive their meaning and what kind of knowledge is justifiable. A semantic empiricist demands that the meanings of terms are traceable to certain aspects of our experience while an epistemic empiricist demands that all knowledge be grounded in experience. Thus far, these positions are fairly innocuous because of the vagueness of the mentioned terms "experience", "grounded in" *etc.* In Hume's specific version of empiricism, a so-called associationist theory of concepts and knowledge is added. All (meaningful) words are associated with an idea, which, in turn, is a copy of a sense impression. The meaning of a word is the idea associated with it. Observing a red object, for example, I have a direct sense impression of something red. This is then stored in my memory as a copy of the impression or an idea. Whenever I hear or say the word "red" the memory recalls that idea and so the word becomes meaningful.

When we now observe a cause-effect relationship, for instance, the famous billiard ball hitting another one and thus causing it to move, Hume asks where there is the impression of the power or agency that makes the first ball move the second. He argues that we observe nothing but the second ball moving after the first ball and similar patterns of events in many other cases. Hence, when we ask what we mean when we say the first ball *causes* the second ball to move, the answer is nothing but the second event occurring after the first one and events similar to the first being regularly followed by events similar to the second. Thus the meaning of the word "cause" is exhausted by "regular succession".

Only things about which we can be certain merited the label "knowledge" in Hume's system. But they were few and far between: only about what Hume called "relations of ideas" (essentially logic and mathematics) and our current sense impressions could we be certain. Propositions such as "A bachelor is an unmarried man" are thus genuine knowledge. It is part of the idea that a bachelor is unmarried and that he is a man. No conceivable state of affairs could falsify this. By contrast, there is nothing in the idea of a cause that implies the existence of the effect. Consider again our first billiard ball. Nothing in the idea of a ball moving towards another one implies that second will move too. For all Hume is concerned, the second ball might remain at rest or it might disappear with a flash. We believe in a necessary connection between cause and effect, that the first ball must move the second one, only because we are habituated to seeing the regular association between these two kinds of events. But it is not inconceivable that nature might take a different course than the one we have been observing thus far. Beliefs about causal relations are therefore not justified.

For Hume, this was a semantic and epistemological problem. At least according to some interpretations, the fact that we are unable to observe anything in causal relationships but regular succession does not imply that in reality there is no power or necessity that ties together cause and effect. In the 20th century, the epistemic and semantic problem has become a metaphysical one too. Philosophers have then asked what there is in the world that makes it appear regular—at least in part—and because of which we can apply the word "cause".

What is important here is that this problem is a philosophical one. Hume asks whether *in principle* causal relationships are observable and whether *in principle* we are justified in making causal claims. A little over 100 years earlier, Francis Bacon was also troubled by causal relations. The aim of his philosophy was to help improve the human condition by gaining control over nature. With the right kind of knowledge—knowledge we today would call causal—we could control phenomena by controlling their causes. This requires methods of causal inference—methods Bacon sought to systematise in his writings (in particular his *Novum Organum*). Bacon knew very well that knowledge that would give us power of that kind is not infallible. But for him, this was not a philosophical problem. Rather, he understood it as the *practical* problem of devising methods that minimise error in inferring from causes to effects. To take our billiard ball example once more, it is certainly true that many things can happen when one ball approaches the other. The Baconian strategy would be to examine the conditions under which it is most likely that the expected effect—the second ball moving away—actually occurs. Thus we would find out that if we want the second ball to move, we had rather not glue, nail or other wise stick it to the table; we had better make sure it is made of marble rather than some explosive with a shell that looks like marble; we had better control for other forces that might act on it such as a strong wind or a magnetic field if the ball was charged *etc*. To be sure, this does not solve Hume's philosophical problem. Even after all precautions have been taken, it is still not inconceivable that the second ball does not move away upon being struck by the first. But for all *practical* purposes we can be sure the ball will move.

These two kinds of considerations have troubled philosophers and scientists ever since Hume, sometimes as parallel concerns within the same thinker. Although Hume is famous for his scepticism about causality, already in his *Treatise*, there is a section entitled "Rules by which to judge of causes and effect", which provide an attempt to lay down tools of causal inference. John Stuart Mill, who lived a century after Hume, is well-known both for deep philosophical thoughts about causality, laws and related topics as well as his experimentalist canon of methods of causal inference (which in fact are a development of some of Bacon's ideas). In the greater part of the twentieth century these two strands have been separated due to a division of labour between philosophers and methodologically inclined scientists. However, in the past twenty or so years, philosophers again have made contact with science in attempts to learn from the methodologists and systematise their ideas. This text retells this story in two parts. Part I provides the background and history by examining both Bacon's and Hume's problems and their solutions. It will also discuss the reasons why their solutions are faulty or incomplete. Hume's regularity account of causality will then be traced through Mill to John Mackie, whose 1970s version of the account provides probably the most tenable but still faulty form.

Part II discusses the various contemporary alternatives to the regularity account. David Lewis's (and his followers') counterfactual account is thought to provide a genuine alternative to the regularity account while the probabilistic theory is regarded as more of a development that takes into account that causal relations are hardly every universally true (though smoking is thought to cause lung cancer, it is neither true that all smokers develop lung cancer nor that all cases of lung cancer are due to smoking). Both share a commitment to reductivism: like Hume's, these accounts attempt to analyse causal claims in non-causal terms. The process theory was originally designed to amend a version of the probabilistic account and overcome its difficulties. It has since been developed into an independent theory of physical causation. Finally, manipulation and natural experiment accounts go back to Bacon and regard experiments as essential to or at least important for causality. Unlike most other accounts, they are non-reductive and understand the problem of causality to be how to extract new causal knowledge from observations and previous causal knowledge. They are thus more methodological in nature than most of the twentieth century alternatives. But before we get to ending with methodology, let us start with methodology by examining Bacon's problem.

2 Francis Bacon: The Glory of Science

Francis Bacon (1561 – 1626) is—along with Descartes—sometimes regarded as the most original and influential philosophical thinker of the scientific revolution. He was deeply dissatisfied with the state of the art of science of his time and sought to reform it by way of a new, revolutionary method. With respect to their *ambition* and the *dissatisfaction* with current state of science, one can thus see remarkable parallels between him and Descartes. But whereas Descartes wanted to provide secure foundations for the sciences by initially sweeping away all beliefs and readmitting only what has passed his method of doubt, Bacon started with what was available and wanted to improve matters piece by piece by his novel method. And whereas Descartes aimed at what one might call moral certainty, Bacon, probably prompted by his almost life-long engagement in political matters, aimed at a practical exploitation of the knowledge thus gained, at "power over nature", having the improvement of human welfare in view.

Thus, Descartes and Bacon shared a disregard for the way science proceeded at their time. What precisely was it that bothered Bacon? Simplifying, Bacon found that two methods which were dominant in the sciences of his time were at fault: the syllogism and simple induction. The syllogism is a valid deductive argument with two premisses, a major and a minor, and a conclusion. Premisses and conclusion consist of a quantifyer (such as "All", "Some" or "No"), a subject term (such as "Animals", "Humans", or "Socrates") and a predicate term (such as "mortal" or "stub-nosed"). A valid syllogistic argument form is for instance: "All A's are B's. All B's are C's. Therefore, all A's are C's". To give an example:

All whales are mammals.
All mammals are vertebrates.
All whales are vertebrates.

Bacon of course recognised that there is nothing wrong with a valid argument as such. His criticism concerns rather the way in which its premisses are established. He writes (NO I 14):[2]

> The syllogism consists of propositions, propositions of words, and words are tokens of notions. Therefore—and this is the heart of the matter—if the notions themselves are muddled and carelessly derived from things, the whole superstructure is shaky.

[2] All references are to the Urbach and Gibson 1994 edition of the *Novum Organum* (NO). The Roman numeral denotes the book, the Arabic numeral the number of the aphorism.

A syllogism such as the above is logically valid as an argument form.[3] That is, the truth of the premisses guarantees the truth of the conclusion. But validity does not guarantee soundness: one or more of the premisses may be false, and therefore, the conclusion may be false, too. For Bacon, the premisses may be false (and, in Aristotelian science, are likely to be false) in at least two ways. There can be either something wrong with how the premiss is quantified (*e.g.* "All birds can fly" is false while "Some birds can fly" is true) or there can be something wrong with the notions that express subject and predicate themselves. The adequate formation of notions (or scientific *concepts*) is important for the truth of the premisses because they classify phenomena. Popper often cites Captain Cook's discovery of black swans in Australia as an instance of falsification of the generalisation "All swans are white". But whether or not the observation of a black swan falsifies the generalisation depends on our concept of "swan". If one aim of our classifications is to establish generalisations with as broad scope as possible, we might change definitions and reclassify the black birds found in Western Australia as something other than swans. In that way, we would change the concept of swan rather than the generalisation.

In both cases, *i.e.* in forming concepts and in establishing generalisations, the reason for the likely error lies in the mistaken way in which they are derived from experience. For example, with respect to scientific generalisations or, in Baconian parlour, "axioms" he says (NO I 25):

> The axioms now in use have been derived from a meagre and narrow experience and from a few particulars of most common occurrence…

One school of natural philosophy—according to Bacon the "empirical school"—does conduct experiments but too few in number and not sufficiently systematically (he thought that the alchemists fall under this category). From this "narrow and obscure foundation" they jump to "conclusions of the highest generality".

The other school—according to Bacon the "rationalist school"—derives its principles from the notions of ordinary language. The problem with this approach is not so much that common notions do not represent anything in nature. To the contrary, they do reflect "familiar occurrences". The problem is rather that this school, in particular Aristotle, subjects all further findings to the prior conceptual system. Here is a statement about how Bacon regards Aristotle's use of experiments (NO I 63):

> Nor should it count for much that in [Aristotle's] books On Animals, and in his Problems and other treatises, he often cites experiments. For he had come to his decision beforehand, without taking proper account of experience in setting up his decisions and axioms; but after laying down the law according to his own judgement, he then brings in experience, twisted to fit in with his own ideas, and leads it about like a captive.

In both cases, then, the problem is that notions and axioms, *i.e.* our scientific concepts and the generalisations in which they figure, are not grounded enough in experience. The school of philosophy engaged in experimentation lacks perseverance and systematicity in their endeavour. The other school starts from the concepts of ordinary language rather than experience, and twists the meanings of concepts in order to fit a deductive system.

The second method that earned Bacon's scorn is that of simple, enumerative induction. Although Bacon does not explicitly define what he means by that, John Gibson and Peter Urbach found a nice example from a contemporary textbook of logic: "Rheynshe wine heateth, Malmesey heateth, Frenchewine heateth, neither is there any

[3] Not all syllogisms are valid. See Glymour 1992, Ch. 2.

wyne that doth the contrary: Ergo all wine heateth (Gibson and Urbach 1994, p. 47).

There are two related problems with simple induction. First, because it merely extrapolates from particular observations, it can never reach beyond the surface phenomena. Since, however, one aim of Baconian science was to explain phenomena in terms of their underlying causes (see *e.g.* NO II 1; also Urbach 1987, pp. 28ff.), simple induction is ill-suited to further an important aim of science. Second, Bacon thought that one of the problems of both the syllogism and simple induction was that their use does not lead (or has not led) to the discovery of new phenomena. Since they merely summarise what is known, they are not able to create new phenomena and thereby further confirm scientific hypotheses.

Bacon's alternative vision of the "true induction" or "interpretation of nature" is to subject the phenomena systematic empirical study, and gradually ascend from particular experiments via intermediate principles to laws of great generality. His system comprises three interrelated stages:

1. Observation and Experiment
2. Classification and Concept Formation
3. Eliminative Induction and Causal Inference.

In a nutshell, the method proceeds by systematically making and recording observations of the phenomenon of interest (natural and controlled); classifying the observations according to a conceptual scheme; and finally, by means of eliminating of false causal hypotheses, infer the true causal law that governs that phenomenon. To take Bacon's own example, let us suppose we try to find the causes of heat. We collect observations of hot phenomena (*e.g.* sunlight, boiling water, agitated animals, vinegar on skin), arrange them in tables ("The Table of Presence": list instances of the phenomenon wherever it is present; "The Table of Absence in Proximity": conjoin the phenomena of the first list with instances that are as similar as possible but where the phenomenon is absent—conjoin hot sunlight with cool moonlight *etc.*; "Table of Degrees": order similar instances according to the strength with which the phenomenon occurs—animals get hotter the more they exercise *etc.*); finally, eliminate false hypotheses as to what could be the causes of the phenomenon (*e.g.* reject the hypothesis that a celestial cause is the cause of heat because there are fires on earth whose heat does not depend on any celestial body and so on).

The product of this scheme Bacon calls the "First Vintage", and he defines (NO II 20, emphasis original):

> *Heat is an expansive motion, checked, and exerting itself through the smaller parts of bodies.* […]

The form of a causal law is that of necessary and sufficient causal conditions for the phenomenon. Thus, a phenomenon and its cause are interchangeable: whenever the cause is present, the phenomenon will also be present and whenever the phenomenon is present, its cause is present, too.

The term "First Vintage" suggests that the process has not come to an end. This careful use of words illustrates that Bacon did not regard his method as infallible. One possible source of error stems from the fact that the three stages of the inductive scheme are interrelated. We can observe, record and classify only in relation to a conceptual scheme. Many of our concepts or notions are, however, as Bacon says, "muddled and carelessly derived from things". The aphorism in which Bacon makes this claim continues: "The one hope, therefore, lies in true *induction*" (NO I 14, emphasis original). But if a good conceptual scheme is both presupposition as well as result of the inductive process, we seem to have a problem. Bacon does not explicitly acknowledge the existence of this circle. A possible solution could be the following. Run the three stages of the inductive process using the entrenched conceptual scheme; if the First Vintage yields a result in the form of necessary and sufficient causal conditions for the phenomenon of interest, stop here; if not, adjust the conceptual scheme accordingly and start at stage one; if the Second Vintage yields a satisfactory result, stop here; and so on.[4]

So how does the new method of interpreting nature solve the problems that beset the old science? The first improvement concerns how experiments and observations are made. Bacon criticised that hitherto experiments have been conducted on a more or less random basis, without system and meticulousness. His scheme subjects experimentation and observation to a rigorous technique where, for example, for positive instances negative counterparts are deliberately sought, or where experiments and observations that vary the degree to which a phenomenon is present are looked for. The second development concerns the fact that concept formation is part of the inductive process. Bacon pointed out that many concepts of our everyday language are muddled and need to be improved upon. In his scheme our classifications are as much part of the empirical investigation as our generalisations about the classified items. Third, in Baconian induction, generalisations are made tentatively and gradually, slowly rising up from low-level experimental laws to principles of higher generality. Fourth, he systematised aids of the inductive process in a list of "Prerogatives of Instances". These "Prerogatives" include descriptions of situations in which causal inference is particularly easy (for example, when a phenomenon occurs all by itself or when we find two circumstances that differ solely with respect to the presence or absence of the phenomenon), descriptions of situations where an experiment can decide between two competing causal hypotheses (the "crucial" experiment) and descriptions of tools that help experimentation and observation such as the telescope and the microscope. With these aids, Bacon can hope to transcend the boundary of the observable and proceed to explain phenomena in terms of their underlying causal structures and processes.[5] Fifth, given that causal hypotheses are now framed in terms of the underlying causal structures and processes of phenomena, it can be expected that new phenomena are suggested, whose existence can be experimentally verified. If that is the case, further credibility is lent to the original hypothesis. A final advantage of the new method is that if the process yields a positive result, we have moved a bit closer to Bacon's overall aim of gaining control over Nature. This is because if we know the law, and if we have the technological means to bring about the cause, we will bring about the effect—or the phenomenon of interest. If, to stay with above example, we know that heat is motion of a certain kind, and we can manipulate the particles (the "smaller parts") of bodies such that we control this motion, we can control the phenomenon of heat.

So this is where the circle closes. The overarching aim of Baconian science is to establish principles or "laws" that can be exploited to gain control over nature. These principles are causal in character: knowing about a phenomenon's causes, and being able to manipulate these causes gives the scientist or engineer control over the phenomenon. The established techniques of the syllogism and its related methods as well as simple,

[4] For details, see Reiss forthcoming d.

[5] One of these tools is analogical reasoning. If one observes that animals get hotter the more they exercise and move, one can reason by analogy that motion of the unobservable particles of matter also induces heat. Bacon realises that this form of reasoning is less certain than others, but it is a form of going beyond observable phenomena nonetheless.

enumerative induction fail at this task as they get us only to the surface correlations between phenomena. Hence, a method of reliable causal inference is required: a method that, on average, makes it more likely than not to find out about the causes of a phenomenon. Bacon's *interpretation of nature* is supposed to achieve just that.

There are surely many things wrong with the details of Bacon's system: his concept of a "form" or law of nature as necessary and sufficient causal condition for a phenomenon is naïve and faulty; he does not provide criteria for when to accept an instrument as reliable; a number of concepts in his system are obscure to say the least. But the point of discussing his ideas in detail is to introduce what one might call "Bacon's problem". That problem is to find or design reliable methods of causal inference—methods which maximise the chance of establishing true causal laws. Unlike Hume's problem, which we shall discuss in the next section, this is an entirely practical, scientific affair. Importantly, it is experience itself rather than any *a priori* consideration which assesses the reliability of particular methods. Solving Bacon's problem does not *guarantee* that inferences made on the basis of reliable methods lead to correct results. To the contrary, it is explicitly acknowledged that all methods of causal inference are fallible. All the Baconian can do is to devise methods that are good enough for all practical purposes. An important distinction for the Baconian is, thus, that between the correlations between phenomena or "accidental regularities" as they have become to be known and the true causes of things. Manipulating a mere correlate of a given phenomenon allows implies controlling the phenomenon only accidentally. But manipulating the causes of a phenomenon implies controlling it (we will see caveats to this claim further below). Hume's problem was a much less practical and much more principled problem.

3 David Hume: The Scandal of Philosophy

Bacon sought to establish a new foundation for science through a revolutionary method. David Hume (1711-1776), Scottish philosopher and historian, aimed higher than even that: he wanted to extend the experimental method to philosophy itself. His first book, the *Treatise of Human Nature*, which he wrote when he was only in his early twenties, is subtitled "An Attempt to introduce the experimental Method of Reasoning into Moral Subjects".[6]

Like his contemporary Newton, Hume tried to find principles that unify a large number of diverse phenomena. Newton was successful at accounting for physical phenomena as diverse as the trajectories of projectiles near the surface of the earth, the orbits of planets or the periods of pendulums by means of a small set of general laws. In the same manner, Hume sought to unify various phenomena of the mind under a set of general principles that are established—as in Newton—on the basis of experience only.

The analogue of material bodies in Newton's theory are ideas or *perceptions* in Hume's (perceptions being the more general term that comprises impressions and ideas); and forces between material bodies in Newton are analogous to principles of association between perceptions in Hume. This is why his theory has also become to be known as the *associationist theory* of knowledge and meaning.

The more general category of perceptions—beliefs, desires, sensations *etc.*—falls into the two narrower categories of impressions and ideas. Impressions "enter [the mind]

[6] At the time, "moral" had a broader meaning than it has today. The "moral sciences" comprised not only ethics but all sciences that deal with human affairs—including philosophy itself.

with most force and violence" while ideas are "the faint images of these in thinking and reasoning" (T Bk I, Pt 1, Sect 1). Although Hume marks the distinction between the two concepts only in terms of the force or vivacity with which the perception is felt, we can think of impressions as all those sensations or inner feelings that are currently excited by the momentary presence of its original cause (*e.g.* the sensation of a tree when one looks at a tree or the sensation of anger when one is angry), whereas ideas are whatever is present in the mind when one remembers an earlier impression (*e.g.* the memory of the tree I saw yesterday or the memory of my anger last week). All perceptions can be simple and complex and they can be of sensation and of reflection. Complex ideas can be analysed into simple ones such as the complex idea of a red house resolves into the simple ideas of "red" and "house-shaped" *etc.* Further, all perceptions divide into perceptions of sensation, *i.e.* those which can be thought of as caused by external objects, and perceptions of reflection, which concern the inner states of the mind such as feelings, pains *etc.*

Hume thinks that all ideas that we have correspond to an impression made beforehand. The reason that we have the idea "red" is that we have observed red things. The reason that we have the idea "angry" is that we have felt anger. This is true also for complex ideas. The reason that we have the idea of a "golden mountain" is that we have made impressions of things that are gold and of mountains. To have the idea of the golden mountain we conjoin them. We have an idea of God because we have impressions (of reflection) of intelligence, wisdom and goodness, we augment them to the extreme and conjoin them.

At least in the *First Enquiry*, which Hume wrote some eight years later as a more popular and lively account of his ideas, this theory of how our concepts come about doubles as a *normative* theory to determine whether or not a concept is meaningful. He writes (Enquiries I, II, 17, emphasis original):

> When we entertain, therefore, any suspicion that a philosophical term is employed without any meaning or idea (as is but too frequent), we need but enquire, *from what impression is that supposed idea derived*? And if it be impossible to assign any, this will serve to confirm our suspicion.

All knowledge, according to Hume, divides into relations of ideas and matters of fact (this has come to be known as "Hume's Fork"; see *e.g.* Cohen 1977). Relations of ideas, that is, the truths of logic and mathematics, can be ascertained by thinking alone. For example, in order to find out that the square of the hypotenuse is equal to the square of the sides of the two sides of a right-angled triangle, we do not need to measure the sides of any real triangle, we do not need to go beyond our ideas. The conclusive test for whether something is or is not a relation of ideas is whether its contrary is conceivable. It is inconceivable that 2 + 2 should not be equal to 4 or that the squares of the sides of a (Euclidean) right-angled triangle should not add up to the square of the hypotenuse. But it is not inconceivable that strawberries are blue or loganberries red. Matters of fact, thus, are ascertained by either direct observation or memories of direct observations or, when unobserved things are concerned by the relation of cause and effect. If, for example, I claim that my friend is in France, and I give as reason or evidence for my belief a letter I have received, that belief is founded by a causal relation between the letter and my friend. Similarly, if I predict that the sun will rise tomorrow, again that belief is grounded in a causal relation between the planets and the sun (among other things). This division of knowledge into these two kinds is an important empiricist strand in Hume. Only the truths of logic and mathematics are known *a priori* by demonstration; all knowledge about contingent matters of fact or, as Immanuel Kant later called it, all synthetic knowledge, is based on experience. He thus rejects claims to knowledge that concern a matter of fact (say, God's existence) but which are supposedly based on *a priori* reasoning (say, the ontological proof).

But what is the foundation for our belief in the causal relation itself? Why do we believe that we can retrodict from an effect to its cause (as in the letter from France example) and predict from a cause to its effect (as in the sunrise example)?

That relation can itself be founded on a relation between ideas or on experience. Hume quickly dismisses the former possibility. Attending to the cause only implies nothing about the effect. We can use our test for whether a relation is a relation of ideas here: is it conceivable that my friend is not in France though I have received a latter? Is it conceivable that the sun will not rise tomorrow though it has done so every day in the past? Yes, it is. Hence, causation is itself a product of experience.

What, then, are the impressions people have when they observe a causal relation? Hume describes it as follows in the *Abstract* to his *Treatise*:

> Here is a billiard-ball lying on the table, and another ball moving towards it with rapidity. They strike; and the ball, which was formerly at rest, now acquires a motion… There was no interval betwixt the shock and the motion. *Contiguity* in time and place is therefore a requisite circumstance to the operation of all causes. 'Tis evident likewise, that the motion, which was the cause, is prior to the motion, which was the effect. *Priority* in time, is therefore another requisite circumstance in every cause. But this is not all. Let us try any other balls of the same kind in a like situation, and we shall always find, that the impulse of one produces motion in the other. Here, therefore is a third circumstance, *viz.* that of a *constant conjunction* betwixt the cause and effect. Every object like the cause, produces always some object like the effect. Beyond these three circumstances of contiguity, priority, and constant conjunction, I can discover nothing in this cause…

Contiguity, temporal priority of the cause to the effect and constant conjunction are all there is in the objects that are related causally. But now we may ask further why we believe that whatever has been related in this way will continue to do so. Why do we believe that cause and effect are related by *necessity*? What is it that guarantees that the sun will rise tomorrow when all we know is that hitherto every day the sun has risen?

One suggestion would be to use an additional premiss in an argument that carries us from causes to effects of the form "(by necessity) the future resembles the past". But that premiss again must either be demonstratively certain or based on experience. Again, it is conceivable that the future does not resemble the past, so the premiss is not demonstratively certain. On the other hand, experience will not carry us into the future. On the basis of experience, we could only contend that past futures have always resembled past pasts but we cannot infer from that that future futures will resemble future pasts or that they must do so.

Hume's solution to this puzzle consists in a naturalistic *explanation* of what people do when they infer from observed causes to unobserved effects or from observed effects back to unobserved causes rather than a *justification* that people are entitled to causal inference. He says that "custom and habit" makes us expect the effect upon observing the cause or vice versa. Because we have seen a billiard ball move when struck by the cue ball in many cases we have formed the expectation of this relation habitually. The impression, thus, of the necessity of the causal relation, *i.e.* of the fact that the second ball *must* move when struck by the first, has nothing to do with anything in the objects. It is rather our anticipation of that happening, and it is located in the mind rather than in the objects.

The problem of induction, the problem of whether past experience gives us any justification for claims made about the future (or broader, whether observed matters of fact give us justification for claims made about unobserved matters of fact), is often referred to as "Hume's problem" (see *e.g.* Howson 2000). But there are in fact two issues. The first concerns the *experienceability* of causal relations. If we say that one billiard ball caused another to move, what is it that we can really experience about that fact? For Hume, we can observe nothing in the causally related objects except their contiguity, the temporal priority of the cause and the regular association of similar objects. The second one is the problem of induction. It concerns the *projectibility* of causal relations. If we say that one billiard ball caused another to move, can we infer from it that the next billiard ball (in the same sort of arrangement) will also cause the other ball to move?

The two issues are not the same. Suppose that, contrary to (the first set of) Hume's worries, each time a cause produces an effect a little speech bubble appears in our field of vision, saying "I am causing!". Causal relations would thus be observable. But Hume's sceptical argument would still go through: observed past causings do not imply anything about future matters of fact. The second billiard ball might still just remain in its place or disappear or change into a beautiful princess. On the other hand suppose, contrary to Hume's second set of worries, that a new project to build another Tower of Babel has been successful, we have spoken to God directly, and he has assured us that the future will resemble the past. This would, however, not imply anything about whether or not we can experience causal relations. If we follow Hume's associationist theory of knowledge, there is still nothing to be known about similar objects but their contiguity, priority and constant conjunction (with the difference that now we can project past regularities into the future with certainty).

The reason that these two problems collapse into one in Hume's theory is that he regards the causality of the causal relation as a form of necessity, and that necessity as *logical* necessity. Thus, in this theory, if causal relations were observable, we could predict with certainty an effect upon observation of the cause—just as we can "predict" with certainty that there is an unmarried man when we see a bachelor. Supposedly even God could not make the facts of logic untrue, and therefore causal relations will have to continue to hold.

But neither does necessity have to mean logical necessity nor does causality have to imply necessity. There are forms of non-logical necessity such as nomological (= in accordance with the laws of nature) or metaphysical necessity. A useful way to understand these concepts is to imagine that in addition to the world we live in there are a (possibly infinite) number of *possible* worlds. A possible world is a fictional entity constructed as a prop for thinking about how the might have been. For example, in this world I've been sitting at my desk all morning, typing the text you are reading. But I could have gone downtown to buy a new stereo instead. Thus we can say that there is a possible world in which I get up in the morning, get dressed, take the subway downtown and buy a stereo. Or, in this actual world the sun has been shining all morning. In another possible world it has been raining. And so on. We can now define as "possible" whatever is the case in at least one possible world. "Metaphysically necessary", on the other hand, is whatever is the case in all possible worlds. Other forms of necessity and contingency lie in between.

Now imagine a picture in which the occurrence of a cause necessitates the occurrence of its characteristic effect *in this world*, but that that is so is itself a contingent (non-necessary) fact *about this world*. In other possible worlds the same cause may necessitate other effects or no effect that all (David Armstrong holds such a position, see his 1978, ch. 24). A cause necessitates its characteristic effect but that this relation holds is itself not metaphysically necessary—in other possible worlds other causal relations obtain. According to such a point of view, it is conceivable that causal relations are observable (and therefore knowable) and causes necessitate their effects, but still one fine day God decides to jumble up all causal laws that have held up to then and to make new laws true. According to this point of view, then, a solution to Hume's first problem, the problem of causal experienceability, does not imply a solution to his second problem, the problem of causal projectibility.

On the other hand, causality does not have to be equated with some kind of necessity, logical or otherwise. Nancy Cartwright (1999, p. 72), for example, thinks of causes in terms of "enablers" rather than "necessitators". She thinks that a concept along the lines of Max Weber's "objective possibility" would help to understand causal relations, not that of necessity. Thus in her account, the relation between causality and induction is even less tight. Even if God did not change the causal laws, an effect could fail to follow its cause since all the cause does is to make the effect possible or, in Cartwright's words, "allows the effect to occur".

In the twentieth century, the early empiricists' associationist theory of knowledge and concept formation has been abandoned almost completely. Few people require these days that concepts are linked to prior impressions in order to be meaningful or that there must be an impression of something if it is to count as knowable. Thus, sense impressions do not seem to be necessary for knowledge. Further, it appears that we are able to err about our current sense impressions as much as about other things. Thus, impressions seem not to be sufficient for knowledge either. But if that is so, Hume's theory faces a number of different challenges. We may ask, for example, whether it is really the case that non-causal facts (such as "there is a billiard ball rolling in front of me") are much easier verifiable than causal facts (such as "the billiard ball in front of me pushed another one and set it to move"). We may ask further whether it is really the case that no causal facts are required to ascertain non-causal facts.

Curiously, many philosophers have clung on to Hume's theory of causation for a long time—as if they still believed in the associationist picture of knowledge. They have also used it to answer different kinds of questions than Hume asked. Hume, notoriously, rejected to ask questions about domains that lie beyond what is knowable by sense impressions. He thus could not answer the metaphysical question of what causality consists in—in the objects. His account is an epistemic one, and to a lesser extent, a semantic one. Hume asks what we can know about the causal relation, and according to his theory of meaning, what we can know about causality is that what doubles as meaning of causality.

As we will see, Hume's theory runs into a number of difficulties if understood as a theory of causality as it is in the objects. It will turn out that none of Hume's conditions is individually necessary, nor are they jointly sufficient. To recap, here is a summary of Hume's theory:

Hume's Theory of Causation

C causes *E* if and only if

(i) *C* is (spatio-temporally) contiguous to *E*,

(ii) *C* occurs before *E* and

(iii) All *C*-type events (*i.e.* all events that are similar to *C*) are followed by *E*-type events (*i.e.* events that are similar to *E*)

(iv)* Upon the observation of *C* we experience a feeling of anticipating *E*.

This version of the definition disambiguates Hume's original in two ways. First, following the standard philosophical discussion I interpret Hume as regarding *events* as the causal relata (a relatum is that which stands in a relation with something else). In Hume's own writings he sometimes uses objects, sometimes events as the things related by causation. In many cases it appears to make more sense to relate events rather than objects—it is the striking of the match that causes it to light, not the match or the match box; it is not the red cloth in itself that infuriates the bull but the torero's waiving it in front of his eyes *etc*. Second, while it is clear that Hume held a psychologistic account of necessity, *i.e.* he understood necessity in terms of our attitude towards objects rather than in terms of properties of the objects themselves, it is not always clear whether he held a necessitarian account of causality or a more purely understood regularity account. The definition (i) – (iv) describes the necessitarian account, while leaving out (iv) would turn it into the pure regularity theory (the option of leaving out the last condition is indicated by the asterisk). of I omit his account of the necessity of the causal relation. For the purposes of this book nothing hinges on the difference, so I do not want to take a stance in the debate (for a discussion, see *e.g.* Beauchamp and Rosenberg 1981, ch. 1).

Before discussing counterexamples to Hume's theory, it must be noted that there are a number of possible attitudes to counterexamples in general, two of which are relevant here. In both cases the situation is that either we have a *prima facie* causal relation but it does not come out as causal under the theory at hand or we have a relation which *prima facie* is not causal but under the theory it comes out as such. One attitude is to take the *prima facie* counterexample at face value and demand the theory be amended or abandoned. The other is to regard the theory as definitional of causality, and thus to reject the counterexample as a pseudo case (or in case the theory regards something as causal while *prima facie* it is not, to accept it is causal after all).

In what follows I will give more weight to the counterexamples and thus demonstrate a need to at least improve on Hume's theory. But it is important to point out that this is not the only option we have. In particular, if we follow Hume and believe that all we can experience about causal relations are his three criteria, and in order to know something, we have to experience it, then we must reject the counterexamples because we cannot know about them. Nonetheless, I shall follow the opposite route.

Neither contiguity nor priority nor constant conjunction are essential to causation. The problems with contiguity and priority of the cause are similar. They are not so much that there are hosts of actual counterexamples of causes that act at a distance or simultaneous or backwards in time but rather that it is unwise to exclude the possibility of such cases because that would incapacitate our ability to construct certain scientific or metaphysical theories. For example, macro economic theorising concerns mainly aggregate magnitudes such as inflation (which is an aggregate composed of price changes of individual goods and services) and unemployment (which is an aggregate composed of individual people that are classified as unemployed according to some set of criteria). Many macro theories predict that a sufficiently large increase in government expenditure will raise (or at least, change) the rate of inflation (*cf.* Hoover 2001, p. 125). There is, however, no sense in which government expenditure is contiguous with inflation. Now, to the extent that one wants to allow macro economic causality of that kind, one cannot demand that causes are contiguous with their effects. There are also theories in quantum mechanics where causes operate spatially discontinuously.

The same is true for the temporal priority of the cause. There are cases of simultaneous causation in macro economics (*cf.* Hoover 2001, ch. 6). Furthermore, several philosophers have attempted to construct causal theories of time (*e.g.*

Reichenbach 1956). Thus, if one wants to ground the temporal order of things in their causal order, one cannot built temporal relations into a theory of causality. Today, it is fairly generally accepted that backwards causation is a possibility, and therefore the temporal priority of the cause cannot be part of a theory of causality (see *e.g.* Dowe 1992).

Interestingly, Hume appears to not have regarded either contiguity or the temporal order as really essential to causality (see *e.g.* Noonan 1999, ch. 3). For example, ideas could figure in causal relations for him, and as they are not located in space, they cannot be "contiguous" to one another. But giving up these conditions merely opens up the door for even more counterexamples. If causes neither have to be contiguous to nor precede their effects, how many non-sensical correlations will there be, which we count as cases of causation?

The problem with constant conjunction is twofold. First, there is a strong intuition that causal relations are intrinsic in the sense that whether or not *A* causes *B* depends on nothing but *A*, *B* and relations between them, no matters of fact beyond that. Paul Humphreys illustrates this thought with the following example (Humphreys forthcoming, p. 12):

> I discover my bicycle tire is flat. Curses! How did that happen? I inspect the tire and there it is, the proximate cause, having a hole (in the tire). Now ask: suppose every other instance of a tire having a hole in it and followed by the tire's being flat was absent from the world… What possible difference could removing the other instances of the regularity make to the causal efficacy of having a hole bringing about the flatness in the tire? None, because the entire causal force of having a hole is present at the exact location where that property is instanced…

According to Hume's theory, causality is an extrinsic relation in a twofold sense. First, whether or not *A* causes *B* depends on whether events similar to *A* are also followed by events similar to *B*. Second, the apparent necessity of the causal relation is supplied by the human feeling of expecting *B* upon observing *A*. This, too, has nothing to do with the intrinsic relation between *A* and *B*.

But even if we do not share these metaphysical views, it is clearly the case that "*A* causes *B*" does not imply that "universally, all *A*'s are followed by *B*'s". To begin with, even under determinism a cause will be followed by its characteristic effect only if it operates unimpeded. "Here is a billiard-ball lying on the table, and another ball moving towards it with rapidity." What happens then is an entirely open affair. The steady ball might be glued to the table, and therefore repel the cue ball; a sudden wind might blow and deflect the cue ball such that it misses the other; a meteor might hit the house and bury the table including both balls such that they will never strike; I might take the steady ball of the table to annoy the players *etc. etc.* Thus, causality seems to be a more complex affair than just constant conjunction between event types.

Another possibility is that causes may fail to bring about their effects even under ideal conditions because they act probabilistically. Again, there is no proof that the world is indeterministic (though there is strong evidence in micro physics that at least some phenomena are) but we do not want to exclude that possibility *a priori*.

The intuition that causal relations are intrinsic to the situations they obtain in is also a problem for condition (iv), the necessity of the causal relation. Whether or not *A* causes *B* does not seem to depend on whether or not there is a human observer who judges that *A* does indeed cause *B*. One charge frequently made is that this condition makes the theory anthropocentric—but causes do not care whether or not there are human beings and whether they care to observe causal relations.

That the four conditions are not jointly sufficient for causation can be seen from the fact that the theory cannot distinguish causes from concomitant effects. Suppose again that determinism is true. Suppose further that smoking causes both bronchitis as well as lung cancer, that all smokers develop both bronchitis and later lung cancer, that only smokers contract this kind of bronchitis and that we expect the occurrence of lung cancer upon observing it. Here a concomitant effect, bronchitis, is contiguous with the principal effect, it precedes it in time, it is universally followed by the principal effect, and the relation is (for us) necessary. But bronchitis does not cause lung cancer.

I aimed to show two things in this chapter. First, Hume transformed the Baconian problem of the reliability of methods of causal inference into the much deeper, much more principled problems of whether causal relations can be experienced *at all* and whether projections from observed situations to unobserved ones can be justified *at all*. Second, regarded as a solution to the Baconian problem, Hume's account fails. None of the four criteria individually nor the joint set serves as a good test for whether a relationship is causal or not.

Although clearly an epistemological (and, to a lesser extent, semantic) matter for Hume, most later philosophers have started to ask different kinds of questions. In particular, they have asked the semantic and metaphysical questions what causal statements mean, and what the causal relation, as it is in the objects, really consists in. In what follows, I will nonetheless bundle Hume's epistemological problem together with the semantic and metaphysical problems as they are all similarly deep and principled and thus similarly opposed to Bacon's practical problem of finding reliable rules to learn from experience.

As the following chapters will show, in the past two hundred sixty five or so years since Hume published his *Treatise*, science and methodology have made great advances with respect to Bacon's problem but although we believe we understand Hume's problem much better today, there is still no solution in sight nor even a consensus about what a possible solution could look like. It is therefore very unlikely that in the twenty years left till Bacon's fourth centenary C.D. Broad's hope will come true (Broad 1926, p. 67):

> May we venture to hope that when Bacon's next centenary is celebrated the great work which he set going will be completed; and that Inductive Reasoning, which has long been the glory of Science, will have ceased to be the scandal of Philosophy?

4 Immanuel Kant: The Copernican Revolution in Metaphysics

→ yet to be written

5 John Stuart Mill: The Methodologist as Gentleman

John Stuart Mill (1806-1873) can, in many ways, be regarded as a follower of Bacon rather than Hume. He did not understand the problem of induction to be the deep challenge Hume saw in it. The principle of the uniformity of nature for him was simply a second-order generalisation from many cases of causal laws that have been individually established. He rather took Bacon's challenge to develop methods for *causal inference*. But as we have seen in the Introduction, hypotheses about what causality is in the objects and the methodology to find out about causal relations are co-dependent. Understood as a theory of causality as it is in the objects, Mill also improved on Hume's views.

In particular, Mill qualified Hume's theory in at least four ways. For Hume, a cause is an event-type that is invariably followed by another event-type, its effect. Mill pointed out that few causes operate in an otherwise causal vacuum (Mill 1874, Bk III, Ch.5 , Sect. 3):

> It is seldom, if ever, between a consequent and a single antecedent that this invariable sequence subsists. It is usually between a consequent and the sum of several antecedents; the concurrence of all of them being requisite to produce, that is, to be certain of being followed by, the consequent".

A corollary of this idea is a change in the relata of causation. According to Mill, it is not events but states or conditions that are related causally. Suppose someone eats a poisoned dish and dies subsequently. What Mill calls the *real cause* of his death includes not only what one would normally regard as an event—the digestion of the poisoned meal—but also a host of other conditions such as a certain bodily constitution, a particular state of the present health and even negative conditions such as the absence of an antidote in the person's bloodstream *etc*. These conditions are more permanent states and not normally captured if we phrase the causal story in event-language.

Mill also adds a principle he calls "plurality of causes". That is, every effect (an effect-type is meant here) can be brought about by a variety of sets of causal conditions, not just one. If a forest fire is the effect under investigation, it could have been brought about by a short circuit, by an unwarily thrown away cigarette or by deliberate arson, say (each "salient" causal condition mentioned here needs additional conditions to be followed by a fire, all of which include the presence oxygen or another gas capable of sustaining a flame, the presence of flammable material *etc*.).

The fourth qualification is that generalisations that hold only conditionally are not regarded as causal (*cf.* Mill 1874, Bk III, Ch. 5, Sect. 6). Night invariably follows day but it is not its cause. Mill solves that problem by demanding that causal generalisations hold unconditionally, that is, not contingent on an underlying causal structure which may be subject to change. Night follows the day only as long as the current arrangement of planets around the sun continues to hold. If, say, a large comet enters the planetary system such that it orbits the sun in between the sun and the earth and it is sufficiently close to the earth, no day will follow that night any more. The same would be true if the earth would be kicked out of its orbit by another comet *etc*.

With this last requirement Mill seems to have overshot his target. All or at least most causal generalisations hold conditional on some other causal arrangement. We surely want to say that arsenic causes death in humans. But that is of course only true as long as certain facts about the human physiology do not change. It is entirely conceivable that we evolve in such a way that in a number of generations arsenic will not be harmful to humans any more. All causal generalisations in the special sciences such as geology, biology, medicine, meteorology, economics, sociology, anthropol-ogy and so on are of this kind. If we do not want to exclude the existence of causal relations *a priori* in these domains we had better not demand that causal generalisations hold unconditionally.

In the following I want to discuss briefly the "canon of inductive methods", which is a set of standard experimental arrangements that Mill developed on the basis of some of Bacon's ideas, and then see how they fare with respect to some of the qualifications Mill built into his version of the regularity theory.

Mill's Methods

Mill proposed a canon of five methods to infer causes from their effects or effects from their causes. They are supposed to apply to situations where causes have various antecedents and where they can be brought about by a variety of sets of causal conditions. But as we will see, they are fairly limited in these cases.

The Method of Agreement: "If two or more instances of the phenomenon under investingation have only one circumstance in common, the circumstance in which alone all the instances agree is the cause (or effect) of the given phenomenon".

Mill uses capital letters to denote cause-factors and small letters to denote effect-factors. If, then, ABC are followed by abc but ADE by ade, the method of agreement tells us to infer that A is the cause of a, since B and C could not have produced a as they are absent in the second situation. B and C (as well as D and E) have thus been eliminated as necessary conditions to produce a. In fact, the method of agreement is a development of a prerogative that Bacon called "Solitary Instances" (NO II 22). He says: "[Solitary Instances] are those that show the nature in question in subject having nothing else in common with other subjects apart from that nature itself".

The Method of Difference: "If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance save one in common, that one occurring only in the former; the circumstance in which alone the two instances differ, is the effect, or the cause, or a necessary part of the cause, of the phenomenon". If ABC is followed by abc but BC is followed by bc, then A is judged the cause of a (or a necessary part thereof).

The method of difference is the second kind of "Solitary Instance" in Bacon, and it is in fact the method of the controlled experiment. In a controlled experiment the idea is to hold everything (which may be causally relevant to the effect) constant, vary the putative cause and observe whether or not the effect obtains.

The Joint Method: If two or more instance in which the phenomenon occurs have only one circumstance in common, while two or more instances in which it does not occur have nothing in common save the absence of that circumstance, the circumstance in which alone the two sets of instances differ, is the effect, or the cause, or an indispensable part of the cause, of the phenomenon.

The Method of Residues: Subduct from any phenomenon such part as is known by previous inductions to be the effect of certain antecedents, and the residue of the phenomenon is the effect of the remaining antecedents.

The Method of Concomitant Variation: Whatever phenomenon varies in any manner whenever another phenomenon varies in some particular manner, is either a cause or an effect of that phenomenon, or is connected with it though some fact of causation. (*Cf.* Bacon's "Instances of Quantity", NO II 47.)

This is in fact an early version of the Reichenbach principle, which we will discuss below in Part II. It says that correlations arise on account of Nature's causal laws—that correlations are not "brute" so to speak.

As Mill recognised himself (Bk III, Ch. X, Sect. 2), the method of agreement fails when effects can be produced by various causes. His example concerns an application to the comparison of two people, say, two artists or philosophers or two generous or selfish people are compared as to the circumstances of their education and history. If we found they had only one circumstance in common, we could not argue that that is the cause for their character. A character can be brought about by a plenitude of causes, and thus the two people could have agreed in their character although their histories had nothing in common.

That Mill included the method in his canon suggests that he ultimately did not believe in the plurality of causes. Alan Ryan, for example, thinks that (Ryan 1987, p. 51)

> the plurality of causes is not a fact about the world, but a fact about our inadequate classification of phenomena. If we reclassify and reclassify, we shall find in the end that there is no such thing as plurality of causes; the universe is composed of a multitude of single facts, which follow and precede one another in an absolutely rigid order.

In order to find such "single facts", we need to decompose the chaos of complex phenomena into small parcels where it is indeed true that one and only one set of antecedents is invariably followed by a particular consequent. This is the analytic part of Mill's methodology of analysis and synthesis (the synthetic part regards the combination of several causal laws to calculate the outcome when many causes act jointly). Note the resemblance with Bacon in this matter too. For Bacon, too, it was part of the inductive process to find a classification of phenomena such that the concepts pick out "simple natures". A simple nature, in turn, is a property such as "heat" or "colour" that is governed by a law. But a law is nothing else but a necessary and sufficient causal condition for its nature. All Mill really adds here is the idea that effects usually do not follow single antecedents but rather a complex of antecedents. But, as for Bacon, the sets of causal antecedents are necessary and sufficient for their consequents. The plurality of causes—and concomitantly the view that causes are mere sufficient conditions for their effects—is a principle that may apply to everyday cases of causation as well as the sciences that concern complex phenomena such as the social sciences but in the realm of "real science" causes are both sufficient and necessary for their effects.

We can infer from this that in order for Mill's methods to work, we need to describe all the relevant facts completely and relevantly (*cf.* Ryan 1987, p. 53). Otherwise we could, as in above example about the two people with similar characters, mistake an accidental correlate as the cause of a phenomenon, or fail to find the cause altogether if it is excluded from the description (if, say, we attempt to locate the cause of the tides in conditions that obtain on earth only). But we may ask, then, what good are methods for which one needs to know already everything that is relevant.

Another precondition is the truth of the law of universal causation: that every event has a cause. Consider the method of difference. Adding a cause A adds an effect a, holding constant all other factors. But if it were possible that the effect a occurs uncaused, the method would not demonstrate that A is really its cause. All it would show is that A is either a's cause or a appeared uncaused—which is not much.

That Mill was happy to use an unproved inductive principle such as this is another fact which shows that he was in Bacon's business rather than Hume's. Mill sought to establish methods that enable us to learn from experience. But he did not attempt to validate our ability to learn from experience. Empirical truths about the world can be known but they remain empirical truths and cannot be converted into demonstrable truths. This judgement is shared by Alan Ryan (1987, p. 57, emphasis original):

> The crucial point is that Mill is *not* "justifying induction" in the usual sense of that phrase. Philosophers should have known better than to be deceived into thinking that he was doing so, since the objection which Mill brings to inductions transformed into deductions is exactly that which be brings to any syllogism, "considered as an argument to prove the conclusion [the argument is that any syllogism is in fact begging the question because the major premiss presupposes the conclusion] (Mill 1874, Bk II, Ch. 3, Sect 2)."

6 John Mackie: State-of-the-Art Regularism

Australian-born Oxford philosopher John Mackie (1917-1981) formalised and developed Mill's sophisticated regularity account in a number of ways. If we denote cause-states or conditions by Latin letters, effect-states or conditions by Greek letters, prefix a “¬” to denote the absence of a condition, let “→” stand for “is invariably followed by” and “←” for “is invariably preceded by”, we can say that Mill’s theory amounts to the following:

Mill’s regularity theory of causation:

$$ABC\neg E \text{ causes } \Phi \text{ iff } ABC\neg E \rightarrow \Phi \ \&\ \Phi \leftarrow ABC\neg E.$$

But Mackie also takes Mill at face value and includes his principle of the plurality of causes (see Mackie 1974, p. 61). The reason for this is probably that one of his aims was to find an analysis of our ordinary concepts. Since the phenomena that we describe with our ordinary concepts do have various possible causes, it makes utter sense to include that principle in an analysis. Therefore, we need to amend the above statement by a series of disjuncts that represent the other possible causes:

$$\begin{aligned} &(ABC\neg E \text{ or } DGH\neg J \text{ or } KLM\neg N) \text{ causes } \Phi \text{ iff} \\ &\qquad (ABC\neg E \text{ or } DGH\neg J \text{ or } KLM\neg N) \rightarrow \Phi\ \& \\ &\qquad \Phi \leftarrow (ABC\neg E \text{ or } DGH\neg J \text{ or } KLM\neg N). \end{aligned}$$

A conjunction of factors such as ABC¬E is a “minimal sufficient condition” in case any of the conjuncts is removed, the effect will not follow. A single factor is thus an INUS condition: an *insufficient* but *non-redundant* part of an *unnecessary* but *sufficient* condition (p. 62). A further proviso made by Mackie concerns the fact that causal relations typically occur against a background of standing conditions that we do not single out as causal factors in our judgements. If Jones lights up a cigarette in his apartment and the apartment house blows up, we would normally say that the gas leak rather than the lighting of the cigarette was the cause of the explosion—the ignition is relegated to the causal field. If, on the other hand, the explosion occurs in a plant that produces chemicals using explosive gases and it is normal that gas leaks occur, we would judge the negligence of the worker who lit up a cigarette to be the cause of the explosion. Here, then, the gas leak is relegated to the causal field (*cf.* p. 35). With this amendment, Mill’s theory reads:

$$\begin{aligned} &\text{In F, } (ABC\neg E \text{ or } DGH\neg J \text{ or } KLM\neg N) \text{ causes } \Phi \text{ iff} \\ &\qquad \text{in F, } (ABC\neg E \text{ or } DGH\neg J \text{ or } KLM\neg N) \rightarrow \Phi\ \& \\ &\qquad \text{in F, } \Phi \leftarrow (ABC\neg E \text{ or } DGH\neg J \text{ or } KLM\neg N). \end{aligned}$$

Another qualification concerns the use of words. What Mill calls “the real cause” is the whole antecedent. Mackie calls this the “full cause” but points out that in ordinary usage what we mean by cause is seldom the full cause but a single INUS condition or a single occurrence of an INUS condition (*cf.* p. 64).

This is the best shot at a regularity account of causation Mackie can think of but he notices that there are still problems involved, in particular with the direction of causation and with what one might call “causal connection”. One of the difficulties is that, like Hume’s original account, it cannot distinguish real cases of causation from accidental correlations. Mackie’s, now classic, counterexample shows that the sounding of the Manchester factory hooters is an INUS condition—and thus is judged a cause—of the Londoner workers stopping work. Suppose the Manchester hooters sound at 5 pm. Suppose, too, that anything else which could make them sound if it was not 5 pm (*e.g.* a faulty mechanism that triggers the hooters too early) is absent, and anything that is additionally needed to make them sound is present (the working mechanism). Now, the sounding of the hooters is an INUS condition for it actually being 5 pm. Here one can see what the problem is: material conditions cannot capture causal direction: it being 5 pm is the cause, while the sounding of the hooters is the effect but both INUS conditions for each other. But since it being 5 pm is also an INUS condition for the Londoners to leave work, the sounding of the hooters is an INUS condition for that effect, too. But obviously the sounding of the hooters in Manchester does not cause Londoners to leave the factories. One can see therefore that the sophisticated regularity account suffers from the same difficulties as Hume's original account.

Mackie thinks that the regularity account fails with respect to both goals that a theory of causation might aim at: it neither provides a satisfactory statement of what we *mean* by causal statements, nor does it show what causation *in the objects* consists in. Let us look at what he says about the problem of meaning first.

If I strike a chestnut with a hammer and thereby flatten it, we say that my striking the chestnut with a hammer was the case of its becoming flatter. What do we mean by that statement? Mackie suggests that a statement of the form

X caused Y

can be analysed as

X was necessary in the circumstances for Y,

and that, in turn, means

If X had not happened, then Y would not have happened.

For Mackie, *counterfactual conditionals* such as this cannot be objectively true. They are only assertable with more or less reason, depending on the amount of evidence we have for the generalisation "All Xs are Ys". Evidence thus plays a double role: first, it confirms inductively general propositions and second, it gives us reason to assert a counterfactual statement.

The problem with this suggestion is that it implies that whether X caused Y is an epistemic matter rather than an objective one. In fact, the account is fairly similar to Hume's. For Hume, the necessity of the causal relation consisted in our subjective feeling of expecting Y upon observing X (or vice versa). That feeling, in turn, is caused by the frequent observation of Ys that follow Xs. In Mackie's account, the necessity of the relation comes along with evidence we have for the counterfactual had X not been, Y would not have followed. But that evidence, too, consists of merely the observation of Xs that have followed Ys. Thus Mackie also regards the causality of the relation as an extrinsic feature.

Another problem, which will discussed in more detail below, is that counterfactual accounts cannot deal with cases of so-called causal overdetermination. Adolf Hitler both poisoned and shot himself. Let us suppose the actual cause of death was the bullet in his brain. Had he not shot himself, he would have died anyway—through the poison. So the shot isn't the cause of his death on the counterfactual analysis.

A potential reply Mackie has is that the death through poisoning isn't the death that actually occurred. Hence, the death, as it actually occurred, would not have occurred had he not shot himself.

Unfortunately, this reply backfires at Mackie. If we define events in this way, causes are not only sufficient but also necessary in the circumstances. This, in turn, implies that causes are counterfactually dependent on their effects. But that means that we cannot distinguish between causes and effects. Mackie notices this, and attempts to solve this problem with a notion of *causal priority*. Mackie defines "x is causally prior to y" if one of the following condition obtains (*cf.* Beauchamp and Rosenberg 1981, pp. 220f.):

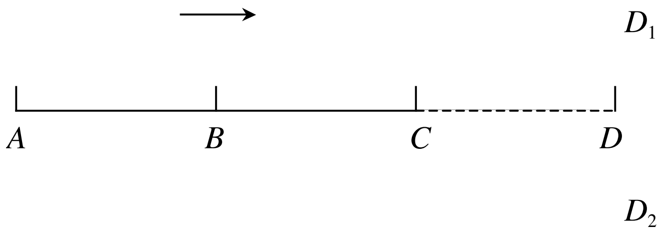
(I) x was fixed at a time when y was unfixed;
(II) x was fixed only at the time of its occurrence, but y was fixed as soon as x occurred; or
(III) there is a causal chain linking x, y and some other event z, such that x is between y and z, and z was not fixed until the occurrence of x.

This account of causal priority leads Mackie into a dilemma depending on the notion of "fixity". Beauchamp and Rosenberg suggest understanding it in terms of Mackie's counterfactual analysis of causation: an event is fixed at time t, if it or its sufficient cause has occurred at or before time t (p. 221). That is, x is fixed at t if and only if either it has occurred at t or if a different event c has occurred such that had c not occurred, x would not occur. Understood in this way, the account is open to the following counterexample, among other things. Suppose a common cause a at t1 causes b to occur at t2 and c to occur at t3. Since causes are necessary and sufficient in the circumstances for their effects, effects are also necessary and sufficient for their causes, and thus b comes out as a cause of c: had b not occurred, a would not have occurred, but had a not occurred, c would not have occurred; therefore, had b not occurred, c would not have occurred, which makes b a cause of c.[7]

The other horn of the dilemma is to take the notion of "fixity" either as primitive or to understand it in causal terms. But the advantage of taking "fixity" rather than, say, "causal priority" as primitive is not at all clear, and taking "causal priority" as primitive renders the account circular. Either way, therefore, there seem to be deep difficulties with our ordinary notion of causation if it is to be understood in terms of counterfactual conditionals.

As an account of causation *as it is in the objects*, Mackie suggests that we amend the regularity theory with an account of a causal mechanism. The crucial thing that is left out the Humean picture is that of the *necessary connection* between cause and effect (as we have noted, the necessity of the causal relation in Hume's account is supplied by the mind, it is not in the objects). Mackie thinks that a causal mechanism can constitute "the long-searched-for link between individual cause and effect which a pure regularity theory fails, or refuses, to find" (1974, pp. 228-9). He then fleshes out the idea of a mechanism in terms of the structural continuity or persistence of certain processes. An example he considers is that of a single particle moving in space free from interference. According to Newton's first law, the particle will move in a straight line.

[7] Despite appearances, we do not need transitivity of the counterfactual conditionals in this argument. Lewis 1973, p. 33, shows that an argument of the following form is valid: (i) Had x not been, y would not have been; (ii) Had y not been x would not have been; (iii) had y not been, z would not have been; Therefore, had x not been z would not have been. This is important because there are good reasons to believe that counterfactual conditionals are not transitive in general.



Mackie goes on to argue that if the particle moves continuously from A to B and from B to C (these being equal distances) in equal times, it is—barring interventions—more expectable that it continues to do so also beyond C. Motion towards, say, D1 would be *prima facie* surprising, since D1 is placed asymmetrically with respect to the line ABC, and the particle might as well go to D2. Since it cannot go to both, and ought not to go to either, it should continue on the straight line. Thus, motion along a straight line is (again, in the absence of interferences) more intelligible than any other motion.

It is unclear to me what the remarks about expectability and intelligibility do here, but what is clear is that Mackie does introduce the notion of "persistence" to cash out Hume's necessity as it is in the objects. One problem that he will encounter is that processes are persistent only as long as no interferences occur, but it will be hard to understand the notion of an interference without causal concepts or at least counterfactuals (and these, in turn, will be hard to flesh out without reference to causality). Though it might be the case that many cases of causality are characterised by persisting structures or processes, it is not at all clear that a reductive account of the kind Mackie seeks will be successful at that. But this idea of Mackie's was well received in later days. In particular David Fair (1979), Wesley Salmon (*e.g.* 1984) and Phil Dowe (*e.g.* 1992) tried to develop accounts along similar lines. They will be discussed in detail in Part II.

Part II: Contemporary Accounts

7 Counterfactuals

We have learnt from the discussion of Mackie's views that the regularity account neither captures the ordinary concept of cause nor causation "as it is in the objects". According to Mackie, the meaning of "A causes B" relates to the counterfactual claim "Had A not been, B would not have been". But his own analysis of the counterfactual claims runs into trouble. On the other hand, causality as it is in the objects seems to involve a continuous process or mechanism that connects cause and effect. Mackie's account of this aspect is also not quite satisfactory. However, his basic insights appear to be promising because a large number of philosophers have tried to cash out the meaning of cause in terms of counterfactuals while others, in objective causation have worked on process or mechanistic theories. In this section, we will examine counterfactual account, reserving process/mechanistic accounts for later.

Princeton philosopher David Lewis has spend much of his life defending a theory of our concept of cause that stresses the counterfactual aspect. An early motivation for developing a counterfactual theory was that Hume himself used the following definitions of cause:

> We may define a cause to be an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second. Or, in other words where, if the first object had not been, the second never had existed.

From the point of view of modern logic, these two definitions are, of course, not equivalent. The first is a version of the standard regularity account and mentions only actual entities, objects in this case. The second, by contrast, is cast in terms of counterfactual states of affairs: an object followed by another such that if the first *had not been*, the second *would not have followed.* This definition thus moves us from actuality to possibility: situations that are not actual but could be. One presupposition for assessing the adequacy of a counterfactual theory of causation, then, is that claims involving counterfactuals can be evaluated. Before going into the details of Lewis's treatment of counterfactuals, however, let me mention a number of limitations of his theory.

Causal relations may obtain at the singular and the generic level. "My girlfriend's falling from her chair during dinner caused her embarrassment" and "Folic acid reduces the risk of neural tube defects" are examples from each category, respectively. Lewis focuses on *singular* causation.

The second limitation concerns the relata of causation. In Hume's definition quoted here, both cause and effect are *objects*. Although in some cases we can think of objects entering into causal relations (singular: "The cue ball caused the eight ball to move", generic: "Seatbelts save lives"), in general this seems unsatisfactory. Consider ordinary causal claims such as smoking causes lung cancer or my striking of the match caused it to light, where the relata are not objects. Lewis takes them to be *events* such as "flashes, battles, conversations, impacts, strolls, deaths, touch-downs, falls, kisses, and the like" (p. 195). For singular causal relations there is, however, a longstanding debate about whether events or facts or tropes are the proper relata. In my view, this debate is not particularly fruitful, and therefore I will not enter it here. Suffice it to say that Lewis regards causes and effects as events of a particular kind but whether that is the best way to think of causal relations is a matter of dispute.

Third, Lewis does not single out a particular salient factor as "the" cause of an effect but treats all contributing factors as causes. In ordinary parlance, people tend to emphasise one factor (*e.g.* "my striking of the match") at the expense of others (the presence of oxygen, the dryness of the match *etc.*) when questions regarding the cause of a given effect. Lewis treats all contributing factors equally; in other words, his theory is a theory of *contributing* causes rather than *total* (or "full" in Mackie's jargon or "real" in Mill's jargon) causes.

With these caveats in place, we can examine Lewis's account in more detail. Lewis defines the concept "causal dependence" in terms of counterfactual dependence. Essentially, event E depends causally on event C if and only if the set of events {E, ¬E} depends counterfactually on the set {C, ¬C}. That is, C depends causally on E if and only if:

C → E and
¬C → ¬E,

where the boxed arrow signifies counterfactual dependence. The first clause is automatically true since both events actually occurred. The issues is whether the second clause, or "Had the cause-event not occurred, the effect-event would not have occurred" is true, too. Lewis explicates counterfactual dependence in terms of "similarity between possible worlds". The concept of a "possible world" has been introduced above in Chapter 3. Now suppose that possible worlds can be ordered according to their degree of similarity with the actual world (or with each other). In the actual world (call it @) I am typing this text. In some possible world (say, w_1) I am sitting in a bar having a beer. In another possible world (w_2) I am a poached egg swimming in sauce hollandaise, just waiting to be eaten for someone's brunch. Denote relation "more similar (to the *actual* world) than" $>_s$. Then we can order the three worlds: @ $>_s$ w_1 $>_s$ w_2: the actual world is most similar to itself, and w_1 is more similar than w_2.

In the original 1973 paper, Lewis takes the concept of similarity as primitive and says that the proposition expressing that event C depends counterfactually on E is true if and only if either both events do not occur (which makes the proposition vacuously true) or some world in which both C and E occur is closer to the actual world than any world in which C occurs but E does not. In other words, the proposition expressing that C depends counterfactually on E is non-vacuously true if and only if it takes less of a departure from actuality to find a world in which both events occur than to find a world where the antecedent event occurs but not the consequent event.

Finally, the causal relation itself is defined in terms of *chains of causal dependence*. Thus, "C causes E" means that there is a chain of causal dependencies of events from C to E (*e.g.* if there is a chain of events C → C1 → C2 → C3 → E where each subsequent event is causally dependent on the precedent event, then C causes E). This last proviso is inserted in order to distinguish the causal relation, which is transitive for Lewis, from the relation of causal or counterfactual dependence, which can be but does not have to be transitive. The reason for this is to avoid certain counterexamples of so-called early pre-emption. As we will see in greater detail below, so-called cases of early pre-emption belong to a class of cases of *redundant causation*. Cases of redundant causation have in common that two or more causes, each of which is capable of bringing about the effect, compete in being the cause which is actually responsible in a given case.

One example concerns a desert traveller. Suppose that two assassins attempt to kill a desert traveller. One poisons his water, the other drills a small hole into his water bottle. As it happens, the water flows out and the traveller dies of dehydration. Here two causes, poisoning and drilling a hole in the flask, compete in bringing about the traveller's death.

The drilling *pre-empts* the poisoning from occurring. Such cases of redundant causation pose a danger to counterfactual account of causation because they falsify the sentence "had the cause-event not occurred, the effect-event would not have occurred" due to a "backup" cause that would cause the effect if the actual cause did not occur.

In the case of early pre-emption, the analysis can be fixed relatively easily though. Despite the fact that there is no direct counterfactual/causal dependence of the death on the drilling—had the second assassin not acted, the traveller would have died anyway, there is a *chain* of counterfactual/causal dependencies from the death to the drilling via dehydration and the outflow of the water (say). There is no such chain between the event of the death and the poisoning of the water. Thus the theory captures adequately our intuitions that the second but not the first assassin's action was responsible for the desert traveller's death.

Lewis's account has had a curious history. Ever since he published it, it has been widely discussed, and counterexamples have been constructed to show that the theory cannot be generally true. Lewis (and his followers) have in response amended the theory in order to avoid a certain kind of counterexample. The amendment, in turn, has then provoked new counterexamples to be found. This has continued till shortly before Lewis death in 2001, where Lewis published a new theory, which differs from the 1973 theory in important respects but still tries to cash out causal dependence in counterfactual terms. Let us go through some of this history.

The first group of counterexamples concerns the apparent vagueness of the concept "similarity between possible worlds". Which possible worlds are most similar to the actual world? In response to an apparent counterexample to his theory, Lewis developed a ranking of weights for his similarity measure. The counterexample concerns the counterfactual "If Nixon had pressed the button, there would have been a nuclear holocaust". In order for this proposition to come out true, Lewis's semantics demands that some possible world in which Nixon presses the button and a nuclear holocaust follows must be closer to the actual world than any world in which the button is pressed but the holocaust somehow fails to obtain. This seems intuitively incorrect: a world in which the holocaust is prevented by some intervening factor looks much more like our world than a world destroyed by a nuclear war where the intervening factor does not occur. But using the following priorities (*cf.* Lewis 1979, p. 472):

1. Avoid big, widespread, diverse violations of law
2. Maximise the spatio-temporal region throughout which prefect match of particular fact prevails
3. Avoid small, localised, simple violations of law
4. Do not worry about approximate similarity of particular fact, even in matters that concern us greatly,

Lewis thinks he is able to solve that problem. The world where the button is pushed but no holocaust ensues requires at least two violations of laws or miracles: the miracle that alters Nixon's course of action and the one that stops the nuclear machinery from running. On this similarity measure, it is therefore *less* similar than the world in ruins.

There are various problems with this measure of similarity. It has been argued that it is insufficient to get the direction of causality right. Like Hans Reichenbach (1956), Lewis wants to keep the possibility open that temporal phenomena are analysed in terms of causality (or counterfactual dependence) and that backwards causation is a conceptual possibility. Hence he cannot built temporal asymmetry into his account. *Prima facie* it seems that the account succeeds. Suppose someone throws a brick against a window and it shatters. Had he not thrown the brick, it would not have shattered. There are various ways in which we can make the antecedent true. We can imagine, for instance, a world w1 which is identical to the actual world @ up to a short period before the throw but then a small, local miracle occurs (*e.g.* some neurons in the thrower's brain are made to fire differently) and the two worlds diverge afterwards. In @, fragments of broken glass are flying through the air, people hear the bang, the thrower gets blamed for his action and so on. In w1, the window is still in place, the counterpart of the thrower has a good conscience, other people have no memories of hearing a bang and so on. In other words, in both worlds the action and the absence of the action leave many traces. Another strategy to implement the antecedent is to make modifications such that the brick is not thrown without the violation of laws. With the assumption of determinism this implies that this world, call it w2, is different from the actual world at all times. Lewis's analysis yields that w1 is closer to @ than w2: though in w1 there is a violation of a law the violation is small and localised, while w2 differs in matters of particular fact from @ throughout its history; a perfect match of particular fact matters more than a small local miracle, and w1 is identical to @ until shortly before the brick is thrown in @. A third way is to let the violation of law obtain *after* the antecedent-event takes place. Here, too, the resulting world (call it w3) is identical to @ over a large spatio-temporal region (throughout its history after the time of the throw). The question is how big the miracle is that is required to make the history converge. Lewis insists that it is big and widespread. Causes, typically, leave many traces in the form of their effects. Effects, by contrast, do not have many causes. This Lewis calls the asymmetry of overdetermination: effects overdetermine their causes; but causes do not overdetermine their effects.

This claim has been challenged. In particular, Adam Elga (2000) has argued that analysing the dynamical properties of thermodynamically irreversible processes (*e.g.* an ice cube melting in a glass of water, smashing a window or frying an egg) shows that Lewis's theory fails to yield the asymmetry. Elga's example involves Gretta who, at 8:00, cracks an egg onto a hot frying pan. Now look at what happens when we run the temporal direction the other way, from future to past: the cooked egg sits in the pan, it uncooks; it coalesces into a raw egg and flows upwards into the shell which closes around it and seals perfectly. This appears to be a very unusual process but the laws of physics entail that processes like this are possible (if rare: time-reversibility implies that they are possible while statistical mechanics shows that they are improbable). The laws of physics also guarantee that a processes like this are very "fragile": a slight change in the initial conditions, that is, in the distribution of the particles and their momenta will, more often than not, result in a very different process. Introduce a (small!) miracle that consists in such a slight change, and the egg will just sit in the pan, slowly cooling and eventually rotting. That is, in the miracle world, the cracking of the egg never occurred. Inside the region affected by the miracle, from the time of the miracle onwards (*i.e.* back in time), things will look thermodynamically normal (*e.g.* the egg becomes more rotten the earlier it gets). Outside that region, things look thermodynamically reversed.

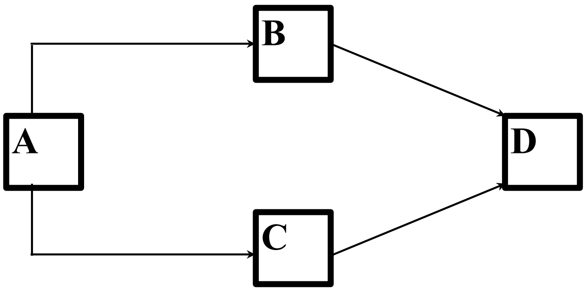
Now, looking at the world forward in time, the situation is reversed. Inside the affected region things look thermodynamically atypical while outside of it they look normal. An interesting question is what happened to the traces that are so important for Lewis. The miracle world matches exactly the actual world shortly after 8:00. Thus, contrary to Lewis's claims, it is full of traces of Gretta cracking the egg. An obvious trace is that there is a cooked egg sitting in the pan in a state that suggests that it had been cracked and thrown into the pan shortly beforehand. But that suggestion is misleading—

in this world, the egg was never raw and was never cracked. Rather, it formed in the pan from rotten slime and reached the cooked stage by a process of reverse rotting (Elga 200, p. S324). Other traces were formed in analogous ways: by thermodynamically reversed processes of particles conspiring in such a way to suggest, after 8:00, that an egg was cracked at 8:00.

There is hence a problem for Lewis's theory if he wants to base causal and temporal direction on the asymmetry of counterfactual dependence. Since past events can depend counterfactually on present events just as present events can depend on past events, causal and temporal asymmetry must have other sources.

Notice the similarity of this difficulty for Lewis with the problem that troubled Mackie's earlier counterfactual account. Mackie, too, could not flesh out causal asymmetry in terms of counterfactuals because his notion of counterfactual dependence was a symmetric notion as well.

There is another source of vagueness which also concerns the evaluation of the counterfactual statement. Suppose we are facing the following causal structure:



A is a mechanism that (in virtue of natural law, say), changes the state of B and C from "on" to "off" such that B is "on" if and only if C is "off" and vice versa. B and C are causes of D. It seems that this structure poses a problem for Lewis's theory. Suppose B fires and D occurs. What if B had not fired? We know that whenever B is set to "off", C is switched on, and thus D would have occurred anyway. Lewis's reply would be that we have wrongly evaluated the counterfactual question "What would have happened to D had B been switched off?" in a "backtracking" manner. That is, to find an answer we took the way by which B happened to be switched off into account—we backtracked to the causes of B in order to evaluate the counterfactual. By contrast, the Lewis-style way to evaluate a counterfactual is by just changing the cause-event, keeping everything else that has happened up to that point constant. Thus, since B in fact fired, C was in state "off" and taking away B's action results in the non-occurrence of D.

Lewis (see his 1979) admits that counterfactual statements are ambiguous in precisely this way but claims that there is a "standard resolution" of the ambiguity, namely to evaluate in a non-backtracking way, *i.e.* without regard to what happened in the causal history of the cause-event. His argument clearly rests on our intuitions as to what rules govern ordinary language usage. I cannot discuss here either his result nor his argument for it; but it is important to point out that there is an inherent ambiguity in the evaluation of counterfactual claims, that Lewis's solution is only one from a number of possible solutions, and that it may not be the most satisfactory solution after all (For further discussion of this matter, see Reiss and Cartwright 2003).

The more serious problems for Lewis's account start when we consider cases of what has come to be called "redundant causation". These cases are characterised by a structure in which two or more causes compete to be responsible for the occurrence of an effect. They subdivide into symmetric and asymmetric cases. Symmetric cases are also called cases of overdetermination. A famous example is about a firing squad that shoots several bullets in the heart of a delinquent. Under Lewis's theory, none of the bullets come out as the cause of the delinquent's death. However, at least between them they do appear to cause the death.

Lewis thinks that cases such as this are not a problem for his theory because his intuition gives out at this point. This is a poor reply, however. Even if we have no intuition as to whether an individual bullet is the cause of death, it seems clearly wrong that the theory yields a definite negative answer.

But asymmetric cases of redundant causation are even more troublesome. Billy and Suzy throw rocks at a bottle. As it happens, Suzy's rock hits first and the bottle shatters. Had it not arrived first, Billy's rock would have broken the bottle a split second later. According to Lewis's theory, Suzy's rock is not a cause of the shattering of the bottle but this seems the wrong conclusion. The literature has labelled cases such as this cases of "late pre-emption". Lewis's theory deals successfully with cases of *early* pre-emption (recall the desert traveller).

The obvious way out for Lewis is to say that the two events (the actual shattering due to Suzy's throw and the counterfactual shattering had Billy's rock got there first) occur at different times and thus are different events. But this reply is not satisfactory. Suppose that the counterfactual cause sends a retarding signal to the actual cause such that the effect-event occurs exactly at the time it would have occurred had it been due to the second, pre-empted cause.

Similar problems beset the account when cases of so-called preventative pre-emption are considered. Billy throws a ball towards a window. Suzy steps into the ball's trajectory, catches it and thus prevents the window from breaking. So far no problem for the counterfactual theory: had Suzy not acted, the window would have been smashed. But now suppose that behind Suzy stood her friend Sally, ready to jump in case Suzy did not catch the ball. We would still argue that Suzy's action prevented the ball from hitting the window but now the counterfactual theory yields a negative answer: had Suzy not jumped, Sally would have and the window was not in danger of breaking.

The example, however, also shows that in some cases it is our intuitions that are wrong or at least dubious, and do not always give us reason to blame the theory. For now suppose that Sally stood in the ball's trajectory already. In this case the ball would have hit her rather than the window anyway. Is Suzy's action still the preventative of the window shattering? If still not convinced imagine that instead of Sally, a solid brick wall stood in between Suzy and the window. Here it seems pretty clear that Suzy's action did not do anything to prevent the window smashing.

Cases of prevention also pose problems for another aspect of the counterfactual theory, *viz.* that they define causal relations are transitive. The simplest of such cases is along the following lines: you walk in the mountains when suddenly a boulder loosens and falls towards you. You see it and duck. The boulder misses and you survive. The boulder causes the ducking, the ducking causes you to survive. But we would be hesitant to say that the boulder is responsible for your survival. Many examples that share the structure of this case have plagued counterfactual theories that regard causality (but not counterfactual dependence) as transitive.

There are indeed conflicting views regarding transitivity. Some, most notably David Lewis himself, regard it as "bedrock datum" which any theory of causality will have to measure up to (see also Hall 2000). Others, including James Woodward (2003) and Judea

Pearl (2000), have argued that it is not an essential property of causal relations that they must be transitive.

A final class of counterexamples concerns cases of so-called "trumping pre-emption". Here is one story that illustrates the counterexample (Schaffer 2000, p. 165):

> Imagine that it is a law of magic that the first spell cast on a given day match the enchantment that midnight. Suppose that at noon Merlin casts a spell (the first that day) to turn the prince into a frog, that at 6:00 pm Morgana casts a spell (the only other that day) to turn the prince into a frog, and that at midnight the prince becomes a frog.

Again, we have apparent causation without counterfactual dependence. Merlin's spell causes the prince to turn into a frog but had he not cast the spell, the prince would have become a frog in the exact same way at the exact same time. In a sense, cases such as this are more worrisome for defenders of counterfactual analyses because (1) there is no chain of intermediate events from pre-empted cause to effect that is somehow broken by the successful cause and (2) there is no difference in the effect-event between actual and counterfactual world (otherwise one could claim that the event as it occurred is a different event than the one that would have occurred had the cause been different).

What are we to make of all these sets of counterexamples? There is a number of possible replies, some of which I want to mention here.

Reply 1: *Deny that the counterexamples endanger the theory*. Though this is not the strategy used by Lewis, I think in the trumping cases there is good reason to deny the plausibility of the counterexample and therefore deny that it puts the theory at risk. In no discussion of trumping cases I have come across there is a mention of a real example where one could see trumping at work. Rather, all cases introduce trumping somehow by fiat: *e.g.* it is a *law* that the spell cast first on any given day is effective; in another example, a major's order trumps a sergeant's *etc*. There is, for example, no detailed causal story about how the successful cause trumps the inferior cause. (Schaffer 2000 does contain a "scientific" example but this is made up just like the others.)

One cannot deny that trumping is a conceptual possibility. But as long as it has not been demonstrated that it is also a possibility *in this world*, we could deny that this constitutes a problem for the theory. This reply is not available for Lewis, however, because he wants his theory to be applicable to *all* conceivable cases of causation, not only actual cases of causation. It is surely conceivable that there are trumping cases, and so Lewis needs a different strategy.

Reply 2: *Fiddle with the details of the analysis*. This strategy has been tried on various occasions by Lewis (see his 1986). For example, cases of late pre-emption he has tried to resolve by means of a new concept of quasi-dependence. Consider the sequence of events linking Suzy's throw with the shattering of the bottle. It is not a chain of counterfactual dependence (the shattering is not counterfactually dependent on any event a split second earlier), but it would be were it not for the second throw. Lewis now thinks that if one considers all regions of the actual and all nomologically identical possible worlds (*i.e.* all possible worlds that in which the same laws are true) which share the intrinsic character of that sequence, one will find that most of them do not have sequences of the kind Billy's throw is made up of; in the majority of regions, the shattering will counterfactually depend on the throw (or an intermediate event). If that is so, Lewis says that the shattering quasi-depends on Suzy's throw in all regions that share the same intrinsic character. In this reading, C is a cause of E iff they are linked by a chain of quasi-dependence.

This strategy, however, is likely to be frustrated by new cases. We just have to take trumping cases to prove the point for this particular amendment of Lewis's theory. There is surely no way to prove it but I would risk the bet that one can find similar counterexamples for any small modification. (Another such amendment is tried by Michael McDermott 1995).

Reply 3: *Draw up a new theory*. Clearly, the distinction between this reply and the previous is not sharp. At what point does an amendment turn a theory into a different theory? However, there is a sense in which Lewis's account of 2000 is a different theory rather than a revised old theory. In the old theory, whether the effect-event occurs depends counterfactually on whether the cause-event occurs. This is an all-or-nothing affair: events are regarded as binary variables with the values "occurs" and "does not occur". In his "Causality as Influence", events are regarded as vectors of continuous variables. Not only whether or not an event occurs is at stake but also when and how it occurs. A central notion is that of an *alteration* of an event, which is an event that occurs at a slightly different time and/or in a slightly different manner from that event. Lewis then defines causation in terms of influence as follows (Lewis 2000, p. 190):

Influence: Where *c* and *e* are distinct events, *c influences e* if and only if there is a substantial range of *c*1, *c*2, ... of different not-too-distant alterations of *c* (including the actual alteration of *c*) and there is a range of *e*1, *e*2, ... of alterations of *e*, at least some of which differ, such that if *c*1 had occurred, *e*1 would have occurred, and if *c*2 had occurred, *e*2 would have occurred, and so on.

Causation: *c causes e* if and only if there is a chain of stepwise influence from *c* to *e*.

Lewis seems to have abandoned his old theory in favour of the influence theory in part in order to be able to accommodate the cases of trumping that have been mentioned above. Recall that Merlin's spell, rather than Morgana's, caused the prince to turn into a frog because his spell was cast first. Lewis claims that he can handle such cases because altering slightly Merlin's spell (*e.g.* to turn the prince into a toad rather than a frog) while holding fixed everything else that happens until shortly before the effect occurs changes the effect but a slight alteration of Morgana's spell does not. If Morgana's spell were to turn the prince into a toad, Merlin's spell would still trump it and the prince would turn into a frog.

Whether the new theory can really handle these cases is a matter of controversy though. First of all, it is clearly the case that some alterations of the pre-empted cause will have an influence on the effect. Since Lewis regards slight changes in timing as admissible alterations, reconsider the case where Merlin's spell is cast at noon and Morgana's at 12:01 PM. Alter Morgana's spell such that it occurs at 11:59 AM and that it is to turn the prince into a toad, and the prince will be a toad by midnight. Lewis argues that certain changes are too distant to count as alterations but it is hard to see how one could work out a metric, which in general will pick out the right event as a cause. A number of further counterexamples have been discussed by Collins 2000 Kvart 2001 and Dowe 2001.

Reply 4: *Split concepts*. This is a more radical approach than any of the previous. It gives up on the idea that a univocal analysis for the concept of causation can be found. Still, the aim of this strategy remains similar to Lewis's: namely, to find conceptual analyses of concepts of causation. Hall 2003 takes the lesson from the counterexamples to be that there is no one concept of causation but two.[8] In a nutshell, his argument begins by claiming that there are four aspects of causal relationships some of which are mutually inconsistent. The aspects are *locality* (causes are connected to their effects by means of spatio-temporally continuous intermediaries), *transitivity* (if c is a cause of d and d is a cause of e, then c is a cause of e), *intrinsicness* (the causal structure of a process is determined by its intrinsic character) and *dependence* (effects counterfactually depend on their causes). Hall then aims to show that the function of some of the counterexamples is to demonstrate that dependence is inconsistent with each of the other aspects of causal relations. As an example, take cases of double prevention: assassin places bomb on your doorstep, friend comes by, sees the bomb and defuses it, you survive.[9] It would follow from dependence and transitivity that the assassin's action caused your survival. But that seems unacceptable. Thus, dependence and transitivity are inconsistent in this case. Hall shows this for all three pairs.

The positive lesson Hall draws is that there is not one concept of causation but two: dependence and what he calls "production". Both require conceptual analyses but different ones. For dependence he suggests the usual counterfactual dependence. He also develops his own analysis of production, linking it to the three aspects of intrinsicness, locality and transitivity.

A similar move is suggested by Chris Hitchcock (2003). He, too, analyses various counterexamples and concludes (p. 21):

> There are a great many cases where we are unclear about what causes what, even though we are clear about all the facts that are supposed to constitute causal relations. The explanation, I contend, is a false presupposition contained in the question: Do events *C* and *E* stand in *the* causal relation? There are many causal relations, and *C* might stand to *E* in some of these relations, but not in others. Here are some candidate causal relations that are brought out by our four central examples: *C* belongs to a causal chain of events leading up to *E*; *C* has a component effect on *E* along some particular causal route; *C* has a net effect on *E* when all causal routes from *C* to *E* are taken into consideration; *C* is a cause of *E* on average in some contingently constituted population; *C* is a cause of *E* as a matter of causal law; *C* is a cause of *E* relative to some particular range of alternatives or domain of variation. The examples show that these relations need to be extensionally equivalent. The time has come to re-direct the resources of theories of causation toward analyzing this collection of causal concepts, and to abandon attempts to characterize *the* causal relation.

Some critics may argue that this strategy goes too far while others may think that it does not go far enough. From the point of view of Lewis-style conceptual analysis, it should be frustrating that there is more than one causal concept. This is because the number of brute facts that one must accept is higher than if there were just one concept. Suppose an analysis such as Lewis's succeeded. We would then have learned a very deep fact about the world or, alternatively, about how our language works. All causal concepts could in principle be eliminated and exchanged for the appropriate counterfactual construction. Thus we would reduce the number of facts we need to know independently. If, on the other hand, no such univocal analysis would be forthcoming but

[8] Though not working within the counterfactual framework, Dupré 1984 and Cartwright 1999 take a similar pluralist stance towards the concepts of causation.

[9] Although almost identical to the abovementioned case where a boulder threatens to hit you, there is a difference relevant for the present context. There is a continuous causal process between the falling of the boulder and your survival as you see the boulder and duck. Here you don't even have to know about the bomb's existence.

rather a number of different analyses, we would have to know, for any given occasion a causal concept is applied (or any given causal relation) under which concept of causation it falls and learn different theories for each different concept.

An alternative to this approach is to restrict attention to a given domain. For example, transference accounts, which will be discussed below, aim to give a reductive theory of physical causation, not causation *simpliciter*.

For realists about causation allowing for different causal concepts, each of which requires its own analysis, would be the wrong move. They believe what the counterexamples show is not that different concepts require different analyses but that the business of trying to analyse causation in terms of other, non-causal concepts is mistaken altogether. By contrast, they take causation to be analytically basic and, if anything, other concepts to stand in need of a causal theory. Here, then, the appropriate strategy would be to give up the business of analysis altogether and seek for alternative ways to characterise causal relations.

Reply 6: *Turn Vice into Virtue—Construct a Realist Theory of Causal Relations.* Proponents of the last strategy I am going to talk about here take the counterexamples to be decisive against any reductive theory of causal relations. There are various ways in which one can be realist about causal relations. A basic distinction is that between monists and pluralists about causal relations. Monists think that all causal relations share an essential element. This essential element may be a theoretical entity, for instance, a universal as in Michael Tooley's and David Armstrong's theories, or a method to test whether the causal relationship obtains, as in Dan Hausman and Jim Woodward's theory. Pluralists, by contrast, think that those relations share at best what one may call "family resemblance": individual members may be more or less similar to others in some respects, but there is no one respect that is common to all members. These realist theories will be discussed after going through another two attempts at reduction: probabilistic and process accounts.

8 Platonism and Thought Experiments

→ yet to be written

9 Probabilistic Causality and Bayes' Nets

There are a number of motivations to develop an account of causality on the notion of *probability*. The first is the apparent failure of (simple) regularity theories that has been pointed out in Part I. Consider the generalisation "Smoking causes lung cancer". It is clear that smoking is neither necessary nor sufficient for the development of lung cancer: there are both cases of lung cancer victims that have never smoked as well as smokers that never developed lung cancer. This much we have known since at least John Mackie. But even the more sophisticated regularity account that he presents does not help much here. Suppose, as a matter of fact, smoking were an INUS condition of lung cancer, *i.e.* in the adequate circumstances A, smoking would be sufficient for lung cancer. Since, however, the circumstances A are not known, it is would be very hard to test whether smoking was indeed a cause of lung cancer or (say) merely a symptom of another cause.[10] A notion of cause that relates smoking to a smoker's *chance* to develop lung cancer appears more promising. In other words, even under determinism it may be useful

[10] But *cf.* Mackie's account of causal inference for "gappy" universal propositions in his 1974, pp. 67ff.

have a probabilistic concept of causality.

The motivation for such a notion under indeterminism is clear. Whatever may be the reason for the belief that indeterminism is an option—be it faith in quantum mechanics or that of other theories in a vast number of sciences including biology, medicine and the social sciences—if it is, then the standard regularity account cannot capture the kinds of causal relations that obtain were it true. This is because under indeterminism there may be no conditions that are sufficient to bring about an effect. It may, for instance, be a law that sixty percent of those who have smoked for at least 20 years will develop lung cancer. There may be nothing more about the circumstances that we could know in order to make that probability one.

One of the earlier attempts to capture the notion of a probabilistic cause was Patrick Suppes's (1970). He, first, defined as *prima facie* cause C a factor which both occurs earlier than another factor E and is probabilistically relevant to it. In symbols,

C is a *prima facie* cause of E iff $P(E \mid C) > P(E)$, where E occurs after C.

A *prima facie* cause can be either genuine or spurious. In order to deal with spurious causes, Suppes uses the concept of "screening off", an idea that was originally introduced by Hans Reichenbach (Reichenbach 1956). Screening off is defined as follows:

S screens off C from E iff $P(E \mid C,S) = P(E \mid S)$.

If there is a factor S, that obtains before C (and E) and screens off C from E, then the correlation between C and E is "spurious". On the other hand, a cause is "genuine" iff it is a *prima facie* cause and there is no factor that screens off its correlation with the effect.

To get a better grip on the idea of "screening off", we need to introduce the concept of probabilistic conditionalisation first. The easiest way to do so is by means of "natural frequencies". Though there are a number of competing theories of probability, here let us think of probabilities in terms of finite numbers in relation to certain populations. Suppose in a population of 1,000 people in total, 400 smoke. We thus say that 400 in 1,000 or 40% smoke, *i.e.* that the probability of smoking (in the total population) is 40%. Denoting the total population by **Ω** and smoking by S, we can write:

$$P(S) = \frac{|S|}{|\Omega|} = \frac{400}{1000} = .4$$

or 40%. Suppose further that in the population in total 50 people contract lung cancer during their lives, hence:

$$P(L) = \frac{|L|}{|\Omega|} = \frac{50}{1000} = 5\%.$$

The quantity of interest now is the probability of lung cancer *given smoking* or the probability of lung cancer conditional on smoking. One can understand conditionalisation as a change in the population of interest. In the unconditional probability $P(S)$, for instance, we relate the number of smokers to the number of people in total $|\Omega|$. Conditionalising on a variable means to shift from the total population to some subpopulation—that part of the population that has the characteristic that is being conditionalised upon. If, for instance, we conditionalise some variable *X* on smoking, we ask what proportion of smokers (rather than people in total) have characteristic *X*. Equivalently, we can relate the *probability* of finding someone who both contracts lung cancer *and* smokes to the probability of finding a smoker. Hence, we can write:

$$P(L \mid S) = \frac{P(L,S)}{P(S)} = \frac{{}^{40}\!/_{1000}}{{}^{400}\!/_{1000}} = \frac{40}{400} = 10\%$$

.

Suppes defines a *prima facie* cause as a factor which both precedes another in time and raises its probability.[11] Let us suppose that smoking precedes lung cancer in time. In this case, *S* is a *prima facie* cause of *L* if and only if:

$P(L \mid S) > P(L)$,

which is true since 10% > 5%. Smoking is therefore a *prima facie* cause of lung cancer. Is it also a cause? According to Suppes, a factor is a cause of another if it is a *prima facie* cause and there is no earlier factor which "screens off" the correlation between the two factors. A factor *C* is said to screen off *A* from *B* if and only if it is true that $P(A \mid B,C) = P(A \mid C)$. Equivalently, we can say that if a factor screens off two factors from each other, it renders them probabilistically *independent*. Two variables *A* and *B* are probabilistically independent if and only if:

$P(A,B) = P(A)P(B)$.

In other words, two variables are independent of they "factor": if their joint probability is equal to the product of their individual probabilities. However, in our example smoking and lung cancer are (unconditionally) probabilistically dependent. That is, their joint probability is unequal to the product of their individual probabilities. They are made independent only if conditioned on a third factor. In general, two variables *A* and *B* are probabilistically independent conditional on *C* if and only if:

$P(A,B \mid C) = P(A \mid C)P(B \mid C)$.

A little bit of algebra shows that *C* screens off *A* from *B* if and only if *A* and *B* are probabilistically independent conditional on *C*. The *prima facie* cause *B* raises the probability of factor *A*. If *C* screens off *A* from *B*, then, conditional on *C*, *B* no longer raises the probability of *A*. If there is such a factor, we also say that *B* is a *spurious* cause of *A*.

Let us say that in the example there is such a factor, a particular genetic condition *G*. In order to find out, then, whether *G* screens off *L* from *S*, we need to relate the probabilities $P(L \mid G)$ and $P(L \mid S,G)$. There are only 20 people which have the factor in

[11] Since a conditional probability is only defined when the variable that is conditioned upon has a non-zero probability, Suppes also demands that a factor has a non-zero probability in order to be a *prima facie* cause.

the population, hence $P(G)$ = 2%. But among those which have the factor, a large proportion contracts lung cancer:

$$P(L\,|\,G) = \frac{P(L,G)}{P(G)} = \frac{{}^{15}\!/_{1000}}{{}^{20}\!/_{1000}} = 75\%.$$

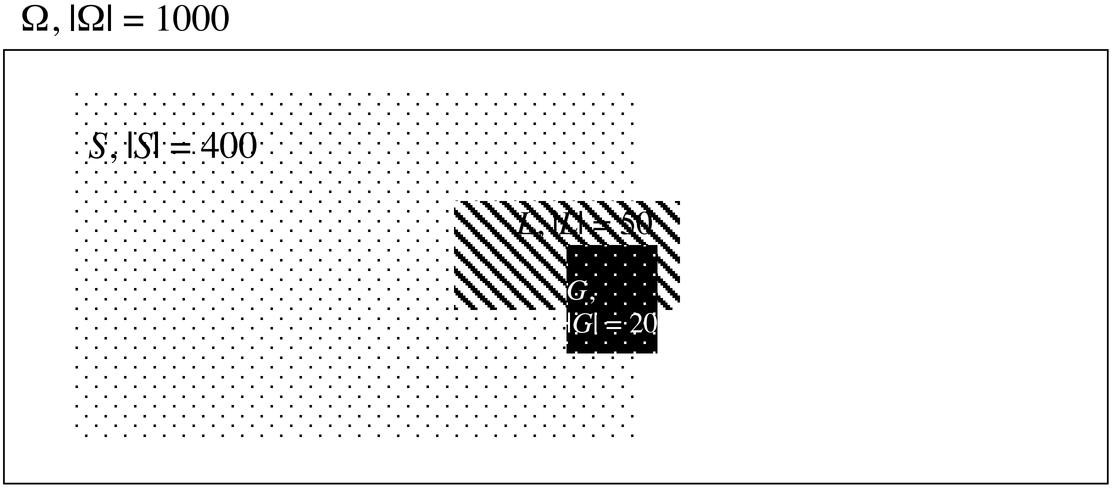
To determine $P(L \mid S,G)$ we also need to know the number of smokers among the bearers of the gene. Let us assume there are 16 gene-bearing smokers:

$$P(S\,|\,G) = \frac{P(S,G)}{P(G)} = \frac{{}^{16}\!/_{1000}}{{}^{20}\!/_{1000}} = 80\%.$$

Now, to find out whether G screens off L from S, we need to check whether S still raises the probability of L once we condition on G. Here we relate the proportion of people that are smokers, have the gene and contract lung cancer (in the example: 12) to the proportion of smokers among the bearers of the gene. Hence we calculate the quantity:

$$P(L\,|\,S,G) = \frac{P(L,S\,|\,G)}{P(S\,|\,G)} = \frac{P(L,S,G)P(G)}{P(S,G)P(G)} = \frac{{}^{12}\!/_{1000}}{{}^{16}\!/_{1000}} = \frac{12}{16} = 75\%.$$

Now, since $P(L \mid S, G) = P(L \mid G)$ = 75%, factor G indeed screens off S from L. Therefore, although smoking is a *prima facie* cause of lung cancer, it is not a genuine or real cause (in the example!). See also Figure XYZ for an illustration of the relations by means of Venn diagrams.



There are a number of counterexamples to this formulation of the probabilistic theory. For one thing, there are causes that appear to *lower* an effect's probability. In a famous example, due to Deborah Rosen, we are asked to consider a beginning golf player who tees off a ball in such a way that it strikes a branch of tree near the green, gets deflected and drops into the hole. If we label the tee stroke A, the collision with the tree D and the hole-in-one result E, we would say that the collision was *negatively* relevant: $P(E \mid A) > \mathrm{P}(E \mid A,D)$—collisions of this kind make a hole-in-one much less likely but on this particular occasion it certainly was the collision that was a cause along the way. Wesley Salmon calls this *the problem of negative relevance* (1984, p. 193).

There is, however, a way out according to Salmon. If one, instead of always conditioning on the initial stroke, conditions on each event in the chain from tee to hole, the problem can be avoided. Let us thus define a number of intermediate events, *viz.* a swing that produces a slice (B), the travelling of the sliced ball towards the tree (C) and the collision with the branch (D). Now, we can see that the following probabilistic relations hold:

$P(B \mid A) > P(B \mid \neg A)$
$P(C \mid B,A) > \mathrm{P}(C \mid A,\neg B)$
$P(D \mid C,B) > \mathrm{P}(D \mid B,\neg C)$
$P(E \mid D,C) > \mathrm{P}(E \mid C,\neg D)$.

The first inequality is trivial as the right hand side is zero: the only way to produce a slice is to swing at all. The second inequality says that the ball is more likely to travel towards the tree if the shot is a slice than if than if the swing is a good shot. The third inequality says that the ball is more likely to collide with the tree if it travels in its direction than if it travels in a different direction. Finally, the fourth inequality says that given the ball travels towards the tree, it is more likely to hit the hole if it collides with the tree than if it does not collide.

Salmon calls this the *method of successive reconditionalisation*. However, there are cases where this method does not succeed. Suppose that there is an atom in an excited state, which Salmon calls the 4th energy level. It may decay to the ground state or 0th level in various different ways, some of which involve intermediate occupation of the 1st level. Define $\mathrm{P}(m \rightarrow n)$ as the probability that an atom in the mth level will make a direct transition to the nth level. Salmon gives us the following values (p. 200):

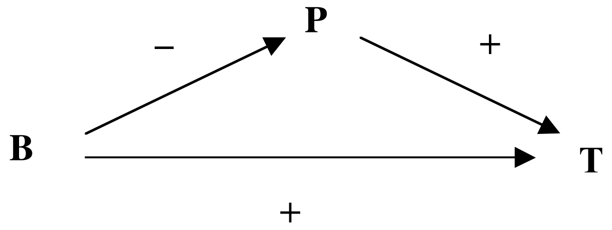
$\mathrm{P}(4 \rightarrow 3) = 3/4$ $\mathrm{P}(3 \rightarrow 1) = 3/4$
$\mathrm{P}(4 \rightarrow 2) = 1/4$ $\mathrm{P}(2 \rightarrow 1) = 1/4$
$\mathrm{P}(3 \rightarrow 2) = 0$.

The probability that the atom occupies the 1st level is 10/16 (= 3/4 * 3/4 + 1/4 * 1/4). If, however, it takes the route via the second level, then the probability that it will occupy the first level is 1/4, *i.e.* occupying the second level is negatively relevant to getting to the first level. Still, it is a member of the causal chain from the initial fourth level to the ground state.

Salmon takes the lesson from counterexamples such as this to be that probability-raising cannot be the essence in a theory of causality (p. 202). An alternative that has been offered is to require probabilistic relevance *simpliciter* rather than positive relevance, *i.e.*, the cause may be either positively or negatively relevant: $\mathrm{P}(A \mid B) > \mathrm{P}(A)$ *or* $\mathrm{P}(A \mid B) < \mathrm{P}(A)$.

This does not work either as the following famous example, due to Germund

Hesslow (1976), shows. Suppose birth control pills (B) cause thrombosis (T). They also prevent pregnancies (P), which are themselves a major cause of thrombosis. The structure of the example is the following:



An arrow means "causes" and the plus and minus signs indicate the direction of the causal contribution. Now, depending on the actual frequencies, it is possible that the causal influence from pills to thrombosis exactly cancels.[12] That is, because pills prevent pregnancies and thereby cases of thrombosis in women that otherwise would have become pregnant, but also cause thrombosis directly, it may be the case that pills are probabilistically irrelevant overall.[13]

This is worrisome on two counts. First, any probabilistic theory of causality seems to require that causes make a difference to their effects. The Hesslow example shows that this is not true in all cases of causation. Second, whether or not pills are probabilistically relevant in this case appears to depend on actual frequencies. That is, it seems that whether or not pills cause thrombosis depends on the number of women taking pills, what other forms of contraception are used, women's actual sexual activity and so on. Inasmuch as the causal relation is thought to be a intrinsic feature, *i.e.* that it depends only on properties of the causes and effects and not on anything else that happens, this is a problem.

For Salmon the way out is to make transmission of energy and transmission of propensities "for various kinds of interactions under various specifiable circumstances" essential to a theory of causality. I will look at the specifics of his theory below.

Bayes' Nets

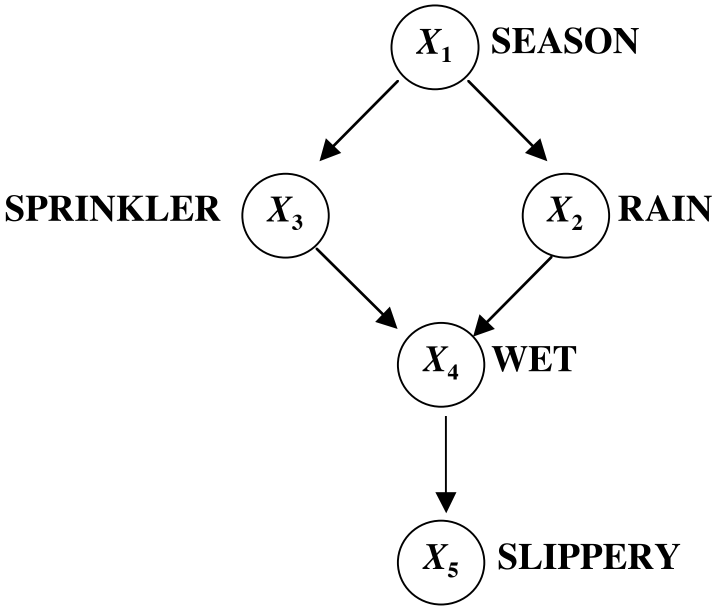
Despite the abovementioned difficulties, ideas similar to those of Reichenbach, Suppes and Salmon have been developed into very influential new tools for causal inference from statistical data, the so-called *Bayes'-Nets methods*. There is a heated and still ongoing discussion about the value of these methods. It is beyond dispute that the methods are very powerful wherever their underlying assumptions are satisfied. The debate concerns, first, the fact that their proponents often appear to advertise them as if the assumptions did not matter and, second, how frequent situations are for which they

[12] Hesslow's use of this example was slightly differently pitched. He meant to show that causes can *lower* the probability of their effects. I use the same example here to show that they can be probabilistically irrelevant.

[13] Salmon solves this part of the problem by an appeal to homogenous reference classes. A cause must be relevant to its effect in a homogenous reference class. In this case, the reference class is not homogenous because pregnancy is relevant to thrombosis. If we partition the total group of women into pregnant (P) and not pregnant (¬P), we can easily see that $P(T|B\&P) > P(T|\neg B\&P)$ as well as $P(T|B\&\neg P) > P(T|\neg B\&\neg P)$. It is however possible to construct similar examples along the lines of the decaying atom where such considerations about the reference class do not matter.

are satisfied. Let us therefore examine the underlying assumptions in detail. Before doing that, however, a number of terms need to be introduced.

It is probably easiest to introduce the important terms by means of an example. Consider one of Judea Pearl's favourite examples (Pearl 2000, p. 15):



A graph such as this one is an entity that consists of a set V of vertices or nodes (in the example X1, …, X5) and a set E of edges that connect pairs of vertices. Vertices correspond to variables, *i.e.* certain measured quantities of interest, while edges correspond to certain relationships between the variables. Edges can be directed, undirected or bidirected. If all edges are directed (as in the example), the graph is called directed graph. It is possible that graphs contain cycles. If for instance, the arrowheads between X1, X3 and X4 pointed in the other direction, there would be a cycle X1 → X2 → X4 → X3 → X1. But if as in the example there are no cycles in the graph, it is called acyclic. A graph that is both acyclic and directed is a DAG or directed acyclic graph.

If an arrow points from, say, X1 to X2, X1 is called the parent or ancestor of X2 and X2 is X1's descendant. We can further define a joint probability distribution P(V) over the variables. The important underlying assumptions concern the relation between the graph on the one hand and the probability distribution on the other. One such assumption is the Markov Condition:

MC: For every X in $\mathbf{V}$, and every set $\mathbf{Y}$ of variables in $\mathbf{V} \setminus \mathbf{DE}(X)$, $P(X \mid \mathbf{PA}(X) \& \mathbf{Y}) = P(X \mid \mathbf{PA}(X))$; where $\mathbf{DE}(X)$ is the set of descendants of X, and $\mathbf{PA}(X)$ is the set of parents of X.

In words, the Markov Condition says that in a graph, a variable is independent of all other variables except its descendants conditional on its parents. Alternatively, we can say that a variable's parents screen it off from all other variables in a graph except its own descendants.

We can easily see that the Markov Condition is a generalisation of Suppes's earlier "screening off" condition. The main differences are that unlike screening off, the Markov Condition does not presuppose a time order of the variables and the definition is relative to larger set of variables of interest rather than just two variables *A* and *B*.

Stated in this way, the Markov Condition is difficult to understand intuitively. That is easier if we give the graph a causal interpretation. We can, for instance, interpret the arrows as causal arrows, showing the direct causal influence of one variable over another. In the example, the season influences the amount of rainfall as well as whether the sprinkler is switched on. The latter two variables determine whether or not the street is wet which, in turn, is responsible for the street's being slippery.

The *Causal* Markov Condition (CMC), then says that a variable is probabilistically independent of all other variables in a graph except its effects conditional on its direct causes. Consider node X4 in the graph. Intuitively, the CMC says that once I know the values of X2 and X3, *i.e.* whether the sprinkler is on or not and whether it is raining, the value of X1, the season, doesn't give me any further information about whether the street is wet or not. There are cases, such as the case at hand, where the CMC appears to make sense. However, one must use it with great care. The CMC is violated in a great number of cases. Those typically cited are in certain kinds of heterogeneous populations; when not all common causes of variables are included in the Bayes' net; when populations are "biased" in certain ways; and in indeterministic systems.

Before discussing some counterexamples to the CMC, I want to draw attention to its intellectual ancestors. It has long been observed that any observed correlation between two variables can be spurious in the sense that they are not directly causally connected and the correlation is due to the existence of a common cause. In the social and biomedical sciences the problem has usually been tackled with the method of "stratification": instead of measuring the correlation between two variables in the total population, the population is divided or *partitioned* into subpopulations according to one way or another.[14]

To get a clear idea of the basic issue, suppose that we live in a very simple world in which only age can matter for recovery from diseases. The "age" variable can take two values: {young, old}. Let there be the rough empirical generalisation that young people tend to recover faster than only people from any given disease. Suppose now that a new drug is tested for its efficacy to relieve constipation. Divide a group of trial participants into "treatment" (the group that receives the new drug) and "control" group (the group that receives a placebo or standard treatment). Unless more is known about the constitution of each group, observing a correlation between treatment and relief is not informative about the efficacy of the drug. This is because it may be the case that more young people are in the treatment group. Therefore, the observed "effect" may be due to the average age of the group (a "confounder") rather than the drug itself. The social science solution now consists in dividing the groups into subgroups or *stratifying*. In this case, the natural strata are the two age groups "young" and "old". If we observe a correlation between treatment and recovery in either stratum or both[15], we have reason to believe that the effect is due to the drug rather than spurious.

As we have seen above, conditioning on a variable is equivalent to moving from population to subpopulation. Therefore, "screening off" is also the same as "stratifying".

[14] "Stratification" is also a term used to denote the social phenomenon of a society being divided into classes, usually across economic characteristics such as income or access to means of production (*economic* stratification) or other social characteristics such as race, religion or caste (*social* stratification). Though obviously linguistically related, stratification as it plays a role in statistics has nothing to do with these social phenomena.

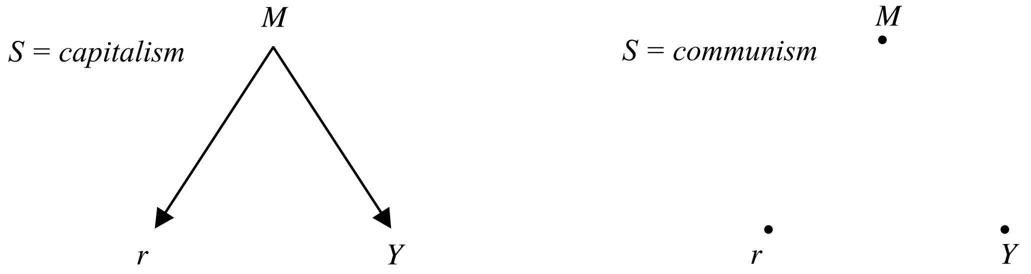
[15] There is a debate about whether the cause-variable should raise the probability of the effect-variable in *all* strata or in at least *one* stratum. For now these difference do not matter.

But the big question is how to partition the population such that causal inference is unbiased. According to Suppes, there should be *no partition* in which there is a variable C that screens off the correlation between A and B (Suppes 1970, p. 28). According to Cartwright, by contrast, the strata should reflect all factors that have a causal influence on the outcome and only these factors (Cartwright 1979). The use of the CMC implies that the strata constitute all causal variables which are modelled in the Bayes' net.

Let us then consider situations where the CMC fails. We can easily see that it fails when common causes of variables in the net are omitted—as should be expected. Suppose there is an omitted common cause of X3 and X5 (we could think of it as a kind of machine that, whenever it turns off the sprinkler it turns on a soap bubble maker (which also makes the street slippery but not wet). Then, knowing that that the value of the sprinkler variable does give me information about the slipperiness of the street even if conditioned on the street's wetness.

"Mixing" poses another problem for the CMC (see Spirtes *et al.* 2000, ch. 2, Cartwright 1999, pp. 130ff.). The problem arises because different causal structures among a set of variables (which, individually, may satisfy the CMC) obtain depending on the value of a "switch" variable where this variable cannot be regarded as a cause of the other variables in the system. Suppose that in capitalist economic systems (S = *capitalism*), the stock of money M is a common cause of interest rates r and aggregate income Y. By contrast, in communist systems, the three variables are probabilistically and causally independent.

In these systems, then, the following structures obtain:



To make things really easy, further suppose that all variables are binary, that is, they can assume only two values (in our example: h for "high" and l for "low"). Our fictitious study concerns 68 countries, of which 40 are capitalist and 28 are communist. The values of the three variables are summarised in the table below.

Capitalist countries:

	$M = h$					$M = l$		
$r = h$	2	0	2		$r = h$	0	21	21
$r = l$	8	0	8		$r = l$	0	9	9
	10	0				0	30	
	$Y = h$	$Y = l$				$Y = h$	$Y = l$	

Communist countries:

	$M = h$					$M = l$		
$r = h$	5	5	10		$r = h$	2	2	4
$r = l$	5	5	10		$r = l$	2	2	4
	10	10				4	4	
	$Y = h$	$Y = l$				$Y = h$	$Y = l$	

Total:

	$M = h$					$M = l$		
$r = h$	7	5	12		$r = h$	2	23	25
$r = l$	13	5	18		$r = l$	2	11	13
	20	10				4	34	
	$Y = h$	$Y = l$				$Y = h$	$Y = l$	

We can now demonstrate that CMC holds in the separate data but not in the combined data (let P_{Cap} stand for the probability in the capitalist countries, P_{Com} for the probability in the communist countries and P for the probability in the total):

$P_{Cap}(r = h \;\&\; Y = h|M = h) = 2/10 = P_{Cap}(r = h|M = h)P_{Cap}(Y = h|M = h) = 2/10*10/10$
$P_{Cap}(r = h \;\&\; Y = h|M = l) = 0/30 = P_{Cap}(r = h|M = l)P_{Cap}(Y = h|M = l) = 21/30*0/30$

$P_{Com}(r = h \;\&\; Y = h|M = h) = 5/20 = 1/4 = P_{Com}(r = h|M = h)P_{Com}(Y = h|M = h) = 10/20*10/20$
$P_{Com}(r = h \;\&\; Y = h|M = l) = 2/8 = 1/4 = P_{Com}(r = h|M = l)P_{Com}(Y = h|M = l) = 4/8*4/8$,

but:

$P(r = h \;\&\; Y = h|M = h) = 7/30 < P(r = h|M = h)P(Y = h|M = h) = 12/30*20/30$
$P(r = h \;\&\; Y = h|M = l) = 2/38 < P(r = h|M = l)P(Y = h|M = l) = 25/38*4/38$.

Hence, using the CMC in the joint data from both capitalist and communist countries, results may be misleading. Spirtes, Glymour and Scheines 2000 (SGS) call this the problem of "mixing": analysing data from mixed populations may lead to wrong causal conclusions. Populations are mixed if the underlying causal structures differ, as in the example. SGS have a solution handy: using mixed data illicitly leaves out a cause, "type of country" in this case. Formally, indeed this does the trick: conditioning also on the type of country restores the validity of the CMC in the example. But there is no reason to suppose that that variable is indeed a "cause" of the remaining variables in the usual sense of the word. The "institutional structure" or "economic constitution" that we use to classify countries into "capitalist" and "communist" may indeed be responsible for the observed causal structure as a whole but not in the same way that money causes interest rates or aggregate income.[16] "Institutional structure" lacks all or most metaphysical characteristics causes are said to have: it does not *precede* the causal structure between money and other variables but rather *constitutes* it; we cannot manipulate it without upsetting the whole system[17]; changes in the institutional structure are not followed by changes in the causal order between the variables of interest in any law-like way.

A less straightforward but possibly more convincing counterexample can be found in Kevin Hoover's econometric work, which will be discussed in more detail in Chapter XXX below. In his methodology for testing causal claims, Hoover explicitly allows for changes in what he calls the "causal field" which can be associated with changes in causal direction. Now, these changes can occur *within regions* over different periods of time. In one application, for example, Hoover concludes (Hoover 2001, pp. 246f.):

> Three principal conclusions emerged from our causal investigation. First, there was a change in the causal field or causal relation between taxes and spending which occurred sometime in the late 1960s and early 1970s. Second, in the period following the change in the causal field, taxes and spending were causally independent. Finally, in the earlier period, taxes and spending were causally linked and there is some mild evidence in favor of taxes causing spending.

It seems very unlikely that we would or could measure the required variable "causal field responsible for the causal relation between taxes and spending" in a way that restores the validity of the CMC in a non-question begging way. Therefore, mixing remains a problem for the CMC.

Mixing is in fact an example of a more general problem for probabilistic accounts of causality: Simpson's paradox.[18] This is a fact about probabilistic dependencies in populations and their subpopulations. Suppose you read in a newspaper the following three statements:

- The likelihood to get accepted in the UC Berkeley Department of Biology is higher for women than for men.
- The likelihood to get accepted in the UC Berkeley Department of Philosophy is higher for women than for men.
- The likelihood to get accepted at the UC Berkeley as a whole is higher for men than for women.

One is inclined to think that at least one of the statements must be false; but in fact they can all be true, as the following numerical example shows:

[16] That this is indeed a severe problem will be appreciated if one considers the widespread use of cross-sectional data analysis in econometrics. Very often, cross-sectional data come from different countries or different socio-economic systems within a country, and thus we can suppose that in many such cases the populations are "mixed".

[17] That we can manipulate each cause independently is the essential ingredient of causation according to James Woodward's (2003) theory, which will be discussed below.

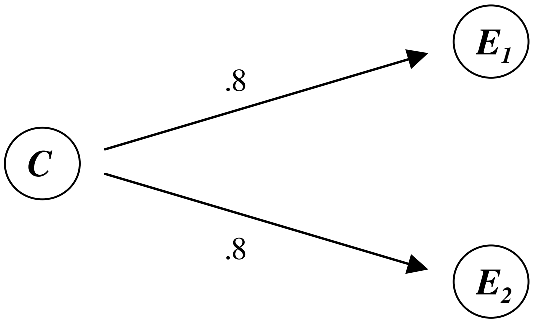
[18] To call Simpson's paradox a *paradox* is misleading. A paradox is a small set of individually plausible but jointly inconsistent statements such as "A single pebble isn't a heap. Adding or subtracting a pebble does not turn a non-heap into a heap and *vice versa*. A thousand pebbles make a heap". Simpson's "paradox" isn't a paradox in this sense but rather an unexpected or counterintuitive fact about probabilistic dependencies in populations and their subpopulations. However, the statistical cases E.H. Simpson drew attention to *look* paradoxical because we can describe them by a small set of individually plausible statements, which *appear* to be mutually inconsistent. The counterintuitive fact about them is that they are actually consistent.

	Men	Women	Departmental acceptance rate
Biology	66/80 (82.5%)	44/50 (88%)	84.6%
Philosophy	12/50 (24%)	24/80 (30%)	27.7%
Total	78/130 (60%)	68/130 (52%)	

Thus, it looks as if UC Berkeley discriminates against women while in fact the chances of being accepted are higher for women than for men—of one analyses the data at the level of departments. The spurious discrimination result obtains because (relatively speaking) more women apply to departments which have a lower acceptance rate (15% of men but nearly 65% of women apply to philosophy).[19]

Important for Simpson's paradox is that gender is associated with a cause, department, but it is inessential whether this is due to a direct causal connection or something else. Paul Holland (1988) gives metaphysical reason why we shouldn't think of gender as a cause in a case such as this. But whether or not Holland is right, more important is the consideration that proponents of Bayes'-nets methods sometimes appear to advertise the methods as if applicable without caveats. This may even be put in cost-benefit-analysis terms: forget about the caveats because it is too costly to learn them all and the risks of misapplying the methods are small. The point of the example is to show that there are these risks. Whether or not they are worth bothering about population homogeneity is something a researcher has to decide for themselves.

The third class of systems where CMC fails comprises indeterministic systems (see *e.g.* Cartwright 1999). Suppose a nucleus C decays with probability .8 into a product E_1 and a by-product E_2. Suppose also that whenever E_1 is produced, E_2 is produced as well (for simplicity; in fact it is enough to assume that the presence of E_1 raises the chance of the presence of E_2 and *vice versa*). The situation is as follows:



If these facts causally exhaust the situation, CMC is violated:

[19] For an early philosophical discussion, see Cartwright 1979. The historical case about UC Berkeley is presented in Bickel *et al.* 1975. For a good introduction and overview, see Malinas and Bigelow 2004.

$P(E_2\&E_1 | C) = .8 > P(E_2 | C)P(E_1 | C) = .8^*.8 = .64.$

Hausman and Woodward 1999 have offered arguments in defence of CMC even in such indeterministic systems and Cartwright 2002 provides an attempt to rebut them. One interesting aspect is how serious we should take a counterexample such as this. We could argue, for instance, that we do not have to take it too seriously because cases such as this obtain only in certain quantum systems. The vast majority of applications of CMC (for instance, in statistical biology, epidemiology, econometrics *etc.*), however, concerns macro systems where determinism is true.

There are at least two replies to this argument. First, it is far from obvious that determinism is true in fields where there is a great use of statistics to help causal inference. Some of these disciplines such as biology, medicine and epidemiology are affected by organisms undergoing mutations, and mutations are often regarded as involving elements of chance. Other disciplines such as economics and the other social sciences involve human action and therefore free will, and it is far from clear that determinism is the best solution to the free will problem. At any rate, it is not the only solution.

Second, Nancy Cartwright argues that "our evidence is not sufficient for universal determinism. To the contrary, for most cases of causality we know about, we do not know how to fit even a probabilistic model, let alone a deterministic one. The assumption of determinism is generally either a piece of metaphysics that should not be allowed to affect our scientific method, or an insufficiently warranted generalisation from certain kinds of physics and engineering models" (1999, p. 115).

The second condition I am going to discuss is called Faithfulness.[20] Again, this is a development of an idea immanent in the earlier probabilistic accounts, namely that it is a necessary condition for a genuine cause to be a *prima facie* cause: if *A* causes *B*, then *A* and *B* are correlated (no matter whether *A* and *B* are causally connected in any other way). Faithfulness is defined as follows (SGS, p. 31):

FC: Let G be a causal graph and P a probability distribution generated by G. <G, P> satisfies the Faithfulness Condition if and only if every conditional independence relation true in P is entailed by the Causal Markov Condition applied to G.

Faithfulness or FC is the converse of CMC. CMC takes us from causes to probabilities: it tells us what conditional probabilistic independence relations should hold in a causal graph. FC takes us from probabilities to causes: it tells us what causal relations should hold given probabilistic independence relations. Like CMC, FC is also violated in Simpson's Paradox cases. This should be intuitively clear by now. If smoking causes heart disease and exercising, which is a preventative of heart disease, is positively correlated with smoking, then it may be the case that in the total population smoking is independent of heart disease. A similar case was mentioned above, where taking birth control was both a (direct) cause as well as a preventive (via pregnancies) of thrombosis. Here too it might be that the causal influence on the two routes just cancels.

SGS argue that such exact cancellations have Lebesgue measure zero (*e.g.* pp. 41f.). That is, though not impossible, they have a zero chance on a Lebesgue measure of occurring. For actual applications this argument is as good as irrelevant however.

[20] There is also a third condition required for some of the proofs, *viz.* "minimality". I will not discuss it here.

Empirical correlations are never exact. It takes a lot of subject-specific background knowledge to determine whether the fact that a measured correlation differs from zero significantly at a given level constitutes evidence for the fact that the actual variable are probabilistically dependent.

Often, such as in the Hesslow example, we can solve the problem by partitioning into homogenous classes. Pregnancy is a cause of thrombosis, so we better partition women into pregnant and not pregnant. We will see that taking birth control will raise the probability of thrombosis in both classes. But this strategy is not always available. Suppose there is a virus that raises the mortality in half of its hosts and decreases it by the same amount in the other half. In each individual case there might be a perfectly well understood physiological process that explains the change in mortality but there is no type-level variable (say, gender or race or age) that one could use to partition the population. Still, in each case we know perfectly well that the change in mortality was due to the virus.

The motivation behind adopting the FC despite its apparent falsifications might be that without it, we would not get theorems and algorithms for causal inference that are as strong as those of the Bayes'-nets methods. It is interesting to note that the probabilistic theories of causality after Suppes (such as Cartwright 1979 and Skyrms 1980) did not rely on the idea that all causes are *prima facie* causes—for exactly the reasons mentioned.

The debate about the significance of such falsifications, I think, carries an important methodological lesson: causal inference requires a substantial amount of background knowledge (*e.g.* are the examined populations homogenous with respect to other causes of the putative effect?; do causes operate deterministically or indeterministically in the system envisaged?; how precisely can variables be measured on the units of the population studied? *etc. etc.*). Importantly, this sort of knowledge is very subject specific. That five percent is an acceptable level of significance in one study does not imply much for others; universal determinism is a bad assumption and so on.

10 Transference Accounts

Bayes' Nets are a development of the probabilistic theory of causality that was introduced by, among others, Patrick Suppes in the 1970s. Since they share their fundamental building blocks, they also suffer from the same problems. Wesley Salmon reacted as early as 1977 to these problems by largely abandoning the attempt to analyse causality in terms of probability-raising. What he thought instead was that the notion of a causal *process* must be central to a theory of causality (Salmon 1977). This reflects one of John Mackie's ideas, namely that what is missing in the regularly theory tradition is an account of how causes are tied to their effects by means of a spatio-temporally continuous process. Mackie, however, has left the idea as a loose suggestion. Salmon instead took up the challenge and attempted to identify the characteristics of that continuous process.

Salmon's theory has three roots: Bertrand Russell's "At-At Theory of Motion", Russell's theory of causal lines and Hans Reichenbach's criterion of mark transmission. To some, the first two ingredients may sound surprising because Russell has been better known for his scepticism about causality than as a positive contributor to a theory of causality. Particularly illustrious is his statement:

> To me it seems that… the reason why physics has ceased to look for causes is that, in fact, there are no such things. The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm. (Russell, 1913, p. 1)

In that paper, Russell argued essentially that the concept of causation is incoherent, and that modern physics has replaced that concept by a concept of a functional law and thus is not in need of the concept of causation.

However, in later work Russell was much more optimistic regarding causation. In 1927, he argued that causality plays a fundamental role even in physics, and in 1948, causal notions are fundamental in four of the five postulates he laid down as basis for all scientific knowledge.

That Russell's ideas contribute to a theory of causal *processes* should be less surprising. He advanced his "At-At Theory of Motion" as a solution to Zeno's arrow paradox. The arrow paradox is one of Zeno's paradoxes of motion with which he aimed to show that Parmenides' theory of oneness is true (this is at least Plato's view; see his *Parmenides*). The paradox can be stated briefly as follows (*cf.* Huggett 1999, pp. 48ff.). Suppose an arrow moves from A to B during a certain interval of time. According to Euclid's theory, time is composed of instants just as a line is composed of points. Now attend to any given instant. Since an instant has no parts (as a point has no parts), the arrow can't move during that instant. But since time is composed of nothing but its instants, the arrow cannot move at all.

Now, according to Russell's "At-At Theory", to move from A to B simply *is* to occupy the intervening points at the intervening instants. Motion thus consists of being *at* a particular point in space *at* a particular point in time. There is nothing else to be said about the situation; there is no question how the arrow gets from A to B.

Salmon combines elements from the At-At Theory with ideas from Reichenbach and merges them into his so-called "At-At Theory of Causal Propagation". A basic element in his theory is the concept of process, which he takes to be similar to Russell's concept of a *causal line* (Russell 1948, p. 459, quoted from Salmon 1984, p. 140):

> A causal line may always be regarded as a persistence of something, a person, a table, a photon, or what not. Throughout a given causal line, there may be constancy of quality, constancy of structure, or gradual changes in either, but not sudden change of any considerable magnitude.

Salmon is critical of the details of Russell's concept of causal line. In a passage immediately preceding the one just quoted, Russell defines (quoted from Salmon 1984, p. 144):

> A "causal line"… is a temporal series of events so related that, given some of them, something can be inferred about the others whatever may be happening elsewhere.

There are three salient elements in Russell's theory of causal lines. First, they exhibit permanence; second, they allow inferences; third, the possibility of inferring is independent of *whatever may be happening elsewhere*. Salmon's main criticism of these elements is that they cannot mark the distinction between genuine causal processes and pseudo processes, which he regards as fundamental. Genuine processes are things such as billiard balls, light beams or radio waves. Pseudo processes are processes such as a shadow moving along a wall or the image of horse moving on a screen. For him, the former but not the latter propagate causal influence. The distinction is important for the purposes of special relativity, among other things. It is a law of special relativity that light is a first signal. That is, no *signal* can travel faster than light in vacuum. Relativity admits, however, that there are processes that travel faster than light. These kinds of processes are incapable of serving as signals—of transmitting information (p. 141). Salmon now criticises that some of the elements that supposedly characterise causal lines (or processes in his terminology) also apply to pseudo processes. Pseudo processes such as shadows or light spots moving along a wall sometimes exhibit great persistence and one can infer properties about later stages from earlier stages of the process. Hence, Russell's theory cannot distinguish causal from pseudo processes.

This is obviously a fallacious argument because only whatever satisfies all three elements counts as a causal process. And Salmon notices explicitly that "the inference from one part of the pseudo-process to another is not reliable regardless of what may be happening elsewhere, for if the spotlight is switched off or covered with an opaque hood, the inference will go wrong" (p. 144). But this means a spotlight moving along a wall is not a causal process according to Russell.

The problem with Russell's theory is rather that inferences even about genuine processes are not reliable regardless of what happens elsewhere. Suppose a light beam travels in a straight line through space. Ignoring everything else in space, we would expect it to keep travelling in a straight line. But suppose that it in fact enters a large gravitational field. Since the light beam would be deflected our expectation would be falsified. And that is true of most causal processes. Think of ordinary objects. Their continued existence is contingent on the absence of a huge nuclear blast in the vicinity. Similarly, radio communication is affected by sunspot activity.

Salmon uses Hans Reichenbach's criterion of mark transmission to distinguish pseudo from genuine process. Intuitively, the mark criterion is very appealing. Consider a genuine process such as a billiard ball. Sometimes, when inexperienced players sink the eight ball at an early stage of the game, they mark a randomly chosen other ball with chalk or pen as substitute eight ball in order not to have to determine the game early. This strategy will usually work because the ball is a genuine process: it transmits the mark from the point of the interaction (in this case, with the piece of chalk or pen). Now consider a pseudo process such as a horse projected against a movie screen. Suppose some evil character intends to shoot the horse. The interaction with the bullet will produce a hole in the screen rather than the horse—once the image moves off, it will not be marked any more.

Intuitively appealing though, the criterion has proved difficult to flesh out in detail. The first difficulty is that the criterion has to be stated in counterfactual terms in order to avoid certain kinds of counterexamples. Consider one of the standard examples, a rotating beacon in the centre of a round building such as the Roman Coliseum. It produces a white spot travelling in a circle on the inner wall of the Coliseum. Mark the spot by putting a blotch of red paint on the wall. The spot will turn red once it arrives at the red blotch, but it will return to white when it leaves the blotch.

So far, so good. But now Nancy Cartwright has pointed out that if a red filter is mounted on the beacon just before the light spot leaves the blotch, it will continue to look red afterwards. If, then, we attend only to the marking of the spot with the paint, it looks as if we had marked that process (*cf.* Salmon 1984, pp. 148ff.).

Therefore, the criterion needs to be formulated in counterfactual terms. It must be made sure that the process would continue to remain unaltered had it not been marked. Salmon's formulation of the criterion is the following (p. 148):

MT: Let P be a process that, in the absence of interactions with other processes would remain uniform with respect to a characteristic Q, which it would manifest consistently over an interval that includes both of the space-time points A and B (A ≠ B). Then, a mark (consisting of a modification of Q into Q'), which has been introduced into process P by means of a single local interaction at a point A, is transmitted to point B if P manifests the modification Q' at B and at all stages of the process between A and B without additional interventions.

A number of objections have been raised against this theory[21], most notably by Philip Kitcher 1989, Phil Dowe 1992 and 1995 and Chris Hitchcock 1995. In response, Salmon abandoned the mark transmission criterion and largely followed Dowe's suggestions. Let us look at the criticism and the subsequent development in detail.

One of the charges Dowe raised in his 1992 paper is that MT makes essential use of the vague term "characteristic". Unless the term is made more precise, the theory is open to counterexamples. His example is the following (p. 201). In the early morning the top edge of the shadow of the Sydney Opera House has the characteristic of being closer to the Harbour Bridge than to the Opera House. But later in the day, this characteristic changes. This characteristic qualifies as a mark by MT, since it is a change in a characteristic introduced by the local intersection of two processes, namely, the movement of the shadow across the ground, and the patch of ground which represents the midpoint between the Opera House and the Harbour Bridge. The example suggests that instead of the vague "characteristic" the account should use something like "non-relational property".

The second problem concerns Salmon's counterfactual formulation of the criterion. Stated without the use of counterfactuals, the theory is open to counterexamples such as Nancy Cartwright's. But with it, it seems to exclude genuine causal processes. Kitcher 1989 pointed out that in reality, processes are subject to interactions continuously. Even a particle travelling in an otherwise empty space is continuously intersecting spatial regions. And even if we required that the intersections be causal, there are still many cases where processes are affected by entirely irrelevant interactions.

Kitcher thinks the villain is not the counterfactual formulation of the theory but the fact that it is a theory of processes and interactions: "What is critical to the causal claims seems to be the truth of the counterfactuals, not the existence of the processes and the interactions… [I]nstead of viewing Salmon's account as based on his explications of process and interaction, it might be more revealing to see him as developing a particular kind of counterfactual theory of causation, one that has some extra machinery for avoiding the usual difficulties that beset such proposals" (Kitcher 1989, p. 472). From the point of view of our preceding discussion (see Section 1), this appears to be a bad move, however. Despite some thirty years of hard work on counterfactual theories of causality, counterexamples multiply.[22] It seems that the (reductive) counterfactual theorist has two equally bad choices. He either uses causal concepts in his analysis of counterfactual statements. In this case the account of causality would be circular and the reductive enterprise would have failed.[23] Or he attempts to avoid causal notions. But then, as we have seen, he will end up with an army of counterexamples to the theory, and the reductive enterprise will have failed too.

Thus instead of getting rid of the notions of causal process and interaction, as

[21] In fact, Salmon provides not only an account of causal propagation—the theory I examine here—but also one of causal *production*. Causal production Salmon seeks to flesh out in terms of causal forks (Reichenbach's conjunctive forks) and causal interactions. The debate that followed focused, however, on the mark transmission criterion, and therefore I omit a discussion of causal production.

[22] I suppose this would constitute a good example of a degenerative research programme in the sense of Lakatos 1970.

[23] This is Pearl's 2000 strategy. He analyses counterfactual statements in terms of possible worlds as does Lewis but then constructs an ordering of the possible worlds in terms of our causal knowledge. For Pearl this is not a problem because he takes causal laws to be analytically basic.

Kitcher recommends, one might try to save processes and interactions by getting rid of the need to formulate the theory in counterfactual terms. This is Dowe's strategy. His theory can be summarised in only two statements:

CQ1. A causal process is a world line of an object which possesses a conserved quantity.

CQ2. A causal interaction is an intersection of world lines which involves exchange of a conserved quantity.

Conserved quantities are physical magnitudes such as mass-energy, linear momentum and charge which our best scientific theories tell us are universally conserved. A world line can be thought of as that which is represented by a line in a Minkowski space-time diagram. Processes (causal and non-causal) are thus "worms" in space-time. An intersection is simply an overlapping of two or more world lines.

Dowe's theory fares very well with respect to the criticisms raised against Salmon's. First, it is very precise about the kind of property a process should possess in order to be regarded causal: a conserved quantity. Science tells us what they are. Second, the theory is formulated exclusively in factual terms. Conserved quantities are possessed and exchanged. There is no need for counterfactual clauses.

In his 1994 article, Salmon accepted Dowe's and Kitcher's criticisms and endorsed Dowe's theory with two important modifications. The first modification concerns the basic physical magnitudes of the theory. Dowe argued that they should be conserved quantities. Conserved quantities are, however, not always *invariant* (under Lorenz transformation) as well. An invariant quantity is one whose value doesn't change when one changes the frame of reference. Suppose you are at rest in a moving train. In the frame of reference of the train, your speed is zero. In the frame of reference of the ground, by contrast, your speed is that of the train. Velocity is thus not an invariant quantity. But now suppose on the train you are playing with an electro butt plug. The charge that occurs on its surface is the same independent of the frame of reference. Charge is thus an *invariant* quantity. It is also conserved. There are also examples for quantities that are invariant but not conserved (Salmon uses *c*, the speed of light, as an example). Salmon argues that because causality is an invariant notion, the theory should require invariant rather than conserved quantities (p. 255):

> We should note, however, that causality is an invariant notion. In special relativity the spacetime interval is invariant; if two events are causally connectable in one frame of reference, they are causally connectable in every frame.

I am not sure whether I fully appreciate Salmon's point here, but *prima facie* it seems to me to be wrong. A number of aspects that we usually ascribe to causal relations—causes precede their effects; causes are contiguous with their effects, to name a few—involve temporal and spatial relations which are *not* invariant under Lorenz transformation. Maybe these aspects are indeed not essential to causation; but it escapes my intuition whether causal relations must be invariant under Lorenz transformation.

Be that as it may, Salmon thinks causal relations are invariant. The second modification concerns the term "possesses" in CQ1. For Dowe, a causal process possesses in the sense of "instantiates" a conserved quantity like an electron has negative charge or a billiard ball has linear momentum. Salmon now argues that the moving spot along the wall of the Coliseum instantiates energy. This world line (the series of spots illuminated by the rotating light), however, does not *transmit* energy.

But since transmission is a causal concept, Salmon must explicate it in turn. Here is his definition (CT for causal transmission) (p. 257):

CT: A process transmits an invariant (or conserved) quantity from A to B (A ≠ B) if it possesses this quantity at A and at B and at every stage of the process between A and B without any interactions in the half-open interval (A, B] that involve an exchange of that particular invariant (or conserved) quantity.

This definition, like MT, embodies the at-at theory of causal transmission; unlike MT, it does not make reference to counterfactuals. The illuminated spot on the wall possesses, but does not transmit, energy because the part of the wall that possesses the energy does so only due to the constant interaction with photons travelling from the beacon. In the absence of further interactions (suppose someone switches off the light) the process ceases to have the quantity.

Chris Hitchcock has shown in his 1995 that it is possible to slightly reformulate this counterexample in order to use it as a counterexample to the modified theory. Replace the wall with a plate that has a uniform non-zero charge density on its surface and the light beam with a shadow that is cast on the plate in such a way that its surface area never changes. This process transmits energy according to definition CT and thus qualifies as a causal process under Salmon's revised theory. But it is a paradigm example of a pseudo process.

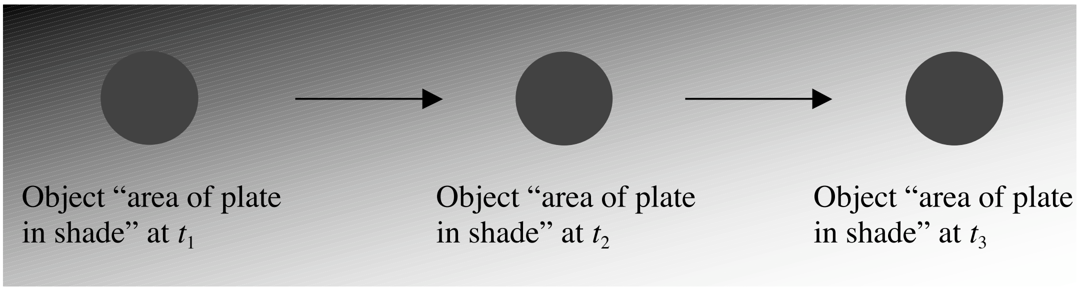
Dowe 1995 defends his theory against Salmon's modifications. He does not mind that causal processes are required to possess invariant conserved quantities but thinks that it is unnecessary to demand invariance. Although the *amount* of linear momentum changes with the frame of reference, the *fact that* an object possesses linear momentum is invariant; and so is the concept of exchange of momentum. On the other hand, invariant quantities that are not conserved will not qualify. Salmon 1994 mentioned the example of "a shadow cast by a moving cat in an otherwise darkened room when a light is turned on for a limited period. This shadow is represented by a world-line with an initial and a final point. The spacetime interval between these two endpoints is an invariant quantity…" (p. 255). He wants to block this kind of counterexample by pointing out that the shadow does not possess the invariant quantity at each space-time point, and therefore cannot be said to transmit it. Dowe argues that this suggestion is unclear; one might as well insist that the shadow possesses its space-time interval at every space-time point.

He also thinks that the requirement that a causal process *transmit* rather than *possess* a conserved/invariant quantity is misguided. The problem is that the direction of influence occurs only on the left hand side of the definition but not on the right. Hence, it cannot distinguish two processes with opposite directionality: one for which the direction of causation is normal from A to B and for the other the direction of causation is backwards in time from B to A. For both, CT requires only that the process has the quantity at every stage between A and B. CT does not solve the problem of the direction of causality. Salmon's transmission amounts to mere possession.

In his 1997 reply, Salmon concedes to Dowe that conserved quantities are the right physical magnitudes as a basis for causation. The residual differences concern the notion of transmission.

There remains a difficulty with the kind of counterexample Chris Hitchcock mentions. The shadow cast on the charged metal plate comes out as a causal process under both Salmon's and Dowe's (1992) theories. Dowe 1995 attempts to rebut this criticism by appealing to a notion of genuine object. The shadow as such does have properties— size, shape and speed, for instance, but no conserved quantities. These are possessed by the plate rather than the shadow. One could (as suggested already by

Salmon 1984 with respect to the rotating light example) instead regard the portions of the surface of the plate in the shadow through time as an object (the Figure).


Object "area of plate in shade" at t_1
Object "area of plate in shade" at t_2
Object "area of plate in shade" at t_3

Dowe now argues that an object like that would not be a genuine object but a "time-wise gerrymander". In order to exist at a time, he says, an object must be wholly present at that time (p. 329). In addition to that he needs a criterion of identity because processes are objects that are connected through time by the identity relation: "the CQ theory identifies genuine causal objects according to the possession of certain properties at a time, and identifies genuine processes over time via the additional presumption of a relation of identity over time" (1995, p. 330). The moving portion of the metal plate thus cannot be an object because the charged molecules in shade at t_1 are not identical to the molecules in shade at t_2.

To Salmon, the concept of genidentity (which is what philosophers have called the concept of identity Dowe needs) carries too much metaphysical baggage and is intuitively unclear. A human body from adolescence to death is arguably the same body. But the molecules it is made up of change constantly, in fact the body undergoes a complete change in about seven years.

(I am not sure whether Salmon's own account of causal transmission rules out Hitchcock's counterexample. There are no apparent interactions when the shade moves on the metal plate.)

Leaving aside these metaphysical qualms, the real difficulty with the CQ approach is that it is severely limited to physics applications, and even within physics it does not answer many of the interesting questions about causality. To demonstrate its limitations, almost any example of a causal relation in social science will suffice. Suppose we are interested in whether a certain kind of event (*e.g.* takeover announcements) causes financial time series (*e.g.* stock prices). Suppose it does: on average, merger announcements raise the stock price of the target and lower the stock price of the bidder. To look for causal processes in the Salmon-Dowe sense would be a completely futile enterprise—as there are myriads of such processes connecting the announcement with the share prices whether or not there is a causal relation between the two (I've written in detail about this point in Reiss forthcoming c).

There are other problems with the process account in the context of causal relations between macro economic entities. Kevin Hoover, for example, argues that the notion of an object at a point in time often does not make sense in macro economics. Consider the GDP, which is an aggregate defined over a period of time, say, a month, a quarter or a year. We might be tempted to shorten the interval and thus arrive at the notion of "instantaneous GDP". But that quantity does not make economic sense. It would behave very weird by, for example, dropping to almost zero over night and during holidays and then increase at fantastic levels the next morning. Still, we do not want to exclude *a priori* the possibility that the GDP stands in causal relations with other aggregates (*cf.* Hoover 2001).

But there is a problem with the account also in a more physical context. Suppose during a game of billiards, the corner pocket, the eight ball and the cue ball lie in a straight line, the player strikes the cue ball and thus pockets the eight ball. What caused the ball to go into the pocket? We would probably say that it was the particular way in which the cue ball was struck by the player. Now consider a less lucky player who misses the pocket. The two processes do not differ in any way with respect to the conserved quantities they possess. Both possess conserved quantities, so both are causal processes. They differ with respect to the amout and/or direction of one of the conserved quantities, *viz.* linear momentum. The CQ theory cannot point to the difference between these two cases (Hitchcock 1995 has argued that the CQ theory does not solve the problem of explanatory relevance).

From a methodological point of view, the CQ theory can thus at best be regarded as providing a sometimes test to distinguish causal from non-causal processes. Hitchcock argues along similar lines:

> I suggest that the conserved quantity theory is best viewed as augmenting rather than replacing the mark-transmission theory. Neither theory provides a reductive analysis of the concepts of causal process and interaction, and neither provides infallible rules for detecting causal processes and interactions. Rather, each provides *guidelines* for recognizing causal processes and interactions, as well as reasons for thinking that these concepts are presupposed by physical science (Hitchcock 1995, p. 316, emphasis original).

11 Agency, Manipulability and Natural Experiments: Back to Bacon?

We have seen a number of properties causal relations (sometimes) have. Effects depend counterfactually on their causes; causes raise the probability of their effects; causal processes but not pseudo processes transmit conserved quantities. In this final section I want to discuss various versions of the idea that we can use causes to manipulate their effects. The intuitive idea is very simple and appealing. If *A* causes *B*, and one can influence *A*, one has the power over *B*. If, for example, aspirins cure headaches, I can take an aspirin to relieve my headache. The equation works the other way around as well: if I can use *A* to manipulate *B*, then *A* causes *B*. If, for example, in a clinical trial I can use a new treatment to increase the chances of headache relief in the treatment group vis-à-vis the control group, then I can judge that the treatment is effective.

Apart from the intuitive appeal, important for the general thesis of this book is the fact that manipulability (and natural experiments) accounts provide a very close tie of metaphysics with methods. According to some of these accounts, experiments and causality are co-extensional: causality is what an ideal experiment tells us. Other accounts do not regard the link to be as strong as this but they too recognise the importance of experiments to elucidate the concept of or concepts of causality.

Within analytic philosophy, this idea has first been introduced by Douglas Gasking (1955). By means of a series of thought experiments, he argued that regularity does not exhaust our concept of causality. Rather, what we mean by relating two quantities as cause and effect is that we can use the cause to change the effect. Knowledge about a cause gives us a *recipe* for changing its effect.

Georg Henrik von Wright's *Explanation and Understanding* (von Wright 1971) contains an elaborate version of this basic idea. He argues that there is a close conceptual relationship between our concepts of causation and action (pp. 65f.):

> I would maintain that we cannot understand causation, nor the distinction between nomic connections and accidental uniformities of nature, without resorting to ideas about doing things and intentionally interfering with the course of nature.

That is, von Wright develops an experimentalist or *agency* account of causation. Action allows us both to *understand* what it means for *A* to cause *B* and to test whether a given regular association is accidental or nomological. First, we understand the concept of causation because we are acquainted with the idea of bringing about a state of affairs by means of doing an action, and a cause, similarly, brings about its effect by "happening" (pp. 73f.):

> Causes do their job [of bringing about an effect] whenever they happen, and whether they "just happen" or we "make them happen" is accidental to their nature as causes. But to think of a relation between events as causal is to think of it under the aspect of (possible) action. It is therefore true, but at the same time a little misleading to say that if *p* is a (sufficient) cause of *q*, then if I could produce *p* I could bring about *q*. For *that p* is the cause of *q*, I have endeavoured to say here, means that I could bring about *q*, if I could do (so that) *p*.

Action, thus, is conceptually prior to causation.[24] It is important to understand von Wright's project in order to appreciate this point of view. Above, we have distinguished a number of questions one may ask regarding causality, *e.g.* epistemological, metaphysical, methodological, semantic questions. This aspect of von Wright's project is clearly semantic or conceptual rather than metaphysical. He does not tell us what causality in the objects *is* but rather what we mean by causal statements.[25] A potential criticism of action theories of causation is that causal relations seem to obtain in situations where we cannot intervene. If we say, for example, that the eruption of Vesuvius caused the destruction of Pompeii, the agency theorist appears to be in trouble as there is no action by which we could make a volcano erupt. But according to von Wright's view, what we are saying when we utter this causal statement is that if we had performed the action of making Vesuvius erupt, Pompeii would have been destructed. The reason that we understand such a claim is that (p. 70):

> The eruption of a volcano and the destruction of a city are two very complex events. Within each of them a number of events or phases and causal connections between them may be distinguished. For example, that when a stone from high above hits a man on his head, it kills him. Or that the roof of a house will collapse under a given load. Or that a man cannot stand heat above a certain temperature. All these are causal connections with which we are familiar from experience and which are such that the cause-factor typically satisfies the requirement of manipulability.

Second, this conception of causation allows us to distinguish nomic from accidental regularities empirically (p. 71):

> For consider what the assumption of universal concomitance of *p* and *q* amounts to. Either it so happens that *p* is always succeeded by *q* and the causal or nomic character of the situation is never put to the test by doing *p* in a situation in which it would not "of itself" come about. […] Then there is nothing which decides whether the truth of the general proposition is only accidental or whether it reflects a natural necessity.

[24] He also says: "In the 'race' between causation and agency, the latter will always win" (p. 81).

[25] A clear statement of that is the following: "To say that causation presupposes freedom would be misleading. It would suggest that the way in which laws of nature operate were somehow dependent upon men. This is not the case. But to say that the *concept* of causation presupposes the *concept* of freedom seems to me to be right, in the sense that it is only through the idea of doing things that we come to grasp the ideas of cause and effect" (pp. 81f., emphasis added).

Von Wright's account has therefore a semantic and an epistemic aspect. One important limitation of his theory is the deterministic framework he adopts. Peter Menzies and Huw Price have attempted to improve on that deficiency and condequently developed and defended an up-to-date agency theory based on probabilistic ideas (Menzies and Price 1993). They first introduce the notion of agent probabilities, that is, a kind of conditional probabilities which are defined as the "probability that B would hold were one to choose to realize A" (p. 190), or $P_A(B)$. In this framework, then, causation and rational decision making go hand in hand: A is a cause of B if and only if a rational agent chooses A if he has an overriding desire that B should obtain; and the expected utility of A is greater than that of $\neg A$ just in case $P_A(B) > P_{\neg A}(B)$.

The similarities with early probabilistic accounts cannot be overlooked. But Menzies and Price claim that the agency theory does not suffer from the same deficiencies because their concept of agent probabilities takes into account more information than mere observed frequencies. For example, while it is true that $P(A \mid B) > P(A \mid \neg B)$ implies $P(B \mid A) > P(B \mid \neg A)$ (and therefore, under an extremely naïve probabilistic theory of causation, we might be forced to say that A causes B if and only if B causes A), it is not true that $P_A(B) > P_{\neg A}(B)$ implies $P_B(A) > P_{\neg B}(A)$: a rational agent will not choose to realise an effect if he desires a cause. Suppose exercise is a reliable preventative of heart disease. Upon observing a good sportsman we expect also to observe a health heart—and *vice versa*. We also expect to observe a drop in heart disease upon the introduction of government programme to promote exercise (if the programme is successful in making people exercise more, say). But we would not expect an increase in the number of people exercising if the government intervened to reduce the prevalence of heart disease by miracle, say, or through any other channel which isn't itself causally linked with exercise. A similar argument shows that under the agency theory a concomitant effect of a common cause will not be mistaken for a cause (p. 191).

Their strategy in the paper is, then, to rebut a number of standard objections to agency theories of causation by likening causation to secondary qualities such as colour. If the reference to human capacities is uncontroversial in the case of colour, and causation can be understood analogously to (or as they prefer, simply as) a secondary quality, reference to human capacities should be acceptable in this context too. In particular, they consider the claim that agency accounts mistake epistemology for metaphysics, the problem that not all causes appear to be manipulable, the circularity argument and the charge of anthropocentrism. As these are very typical arguments, levelled against all agency or manipulability accounts, and at least some of them have philosophical relevance much beyond thinking about causality, I go through them in some detail.

Circularity. We have seen above that there has been a strong tradition which aimed at *reducing* causal concepts to other, philosophically less problematic notions. A concept's reduction can be achieved by providing a definition that can act as a substitute. I can, for example, get rid of the concept of "bachelor" by substituting "unmarried man" whenever it occurs. A definition is circular when the same concept appears on both sides of it. If reduction is the aim, circularity is a problem because the substitute or *definiens* inherits the defectiveness of the concept-to-be-substituted or *definiendum*. If "red" appears to be philosophically problematic, we do not want to define the concept as "that which looks red to a competent observer in normal conditions". But this is just what the dispositional theory (which Menzies and Price adopt for the purposes of this paper) appears to be doing.

Menzies and Price counter that colours have the advantage of being open to ostensive definition. Someone who does not understand the meaning of "red" can be initiated by pointing to a red object. The dispositional theory and therefore be saved. In exactly the same way, the authors attempt to rebut the circularity argument against their agency theory of causality. The agency theory relies on a notion of "bringing about", as we have seen above, and what is bringing about other than causing? However, in some cases of bringing about we have direct personal experience: when we *act* to bring about a state of affairs. Understanding these cases does not require prior acquisition with a causal notion (pp. 194f.).

Unmanipulable causes. As discussed in relation to von Wright's theory, there appear to be causal relations outside the realm of possible human interventions. And like von Wright, Menzies and Price counter that the counterfactual statement regarding what would happen were we to intervene can be true. What colour does the interior of the sun have? From its physical constitution we can infer that if an observer were to observe the interior, it would look red to him. Likewise, if an agent were to manipulate tectonic plates in the right way, an earthquake would result. And how do we know the counterfactual is true? In the same way that von Wright described: by analogy. Friction between tectonic plates is relevantly similar to patters of events that we can manipulate (in the sense of sharing a number of intrinsic properties), so we infer that the causal relation holds in the unmanipulable case as well.

Anthropocentricity. Agency accounts make causation relative to human capabilities: no (human) intervention in, no causation out. This does not only constitute a problem for apparent causal relation in areas that are beyond reach for us, it also makes the notion of causation subjective in an important sense: agents with different capabilities to intervene would apply the concept to different kinds of cases. Menzies and Price argue here that because we have the ability to extend our notion to unmanipulable cases by analogy, the subjectivity of causation is, though extant, very limited. Only if we had absolutely no ability to intervene—if we were "intelligent trees" so to speak—we would not call the same things causes and effects. And this kind of subjectivity is virtuous rather than vicious since intelligent trees would not have a concept of causation (pp. 199-202).

This last argument contains a leap of faith, however. Why would we think that a species with different abilities would, and would be justified to, extend its concept of cause to exactly the same kinds of cases that we cover? Certain kinds of seabirds pick up crabs and other shellfish from the water or shore, raise it to some height, and then drop it in order to smash it and get to their inside. Suppose, then, that these birds have a concept of cause which is co-extensive (it applies to the same kinds of states of affairs) with (our) "smashing". How would these creatures be able to extend this concept to qualitatively different cases of causing such as "watering the plants cause them to grow"? More importantly, even if they did extend the concept to cover such cases, would they be justified?

There is one way in which they could be justified, namely, if it could be argued that the well-understood or agency cases of causation share the relevant properties with the hitherto not understood or unmanipulable cases of causation. But now the difficulty emerges that it is hard to see how these properties could not be the causal properties of the cases at hand. This is, in fact, the main objection James Woodward raises against the abovementioned agency theories (Woodward 2003). He consequently develops his own manipulability theory in a way as to avoid this difficulty.

In my view there are three major differences and two commonalities between Woodward's theory and its predecessors. Being agency or manipulability theories, they share an understanding of causality in terms of bringing about a result by means of a manipulation. They also share a formulation in counterfactual terms. Causality does not only obtain whenever a factor *is* manipulated and a certain result ensues but also whenever *were* the factor *to be* manipulated, the desired result *would* ensue. The main differences are (1) a complete turning away from the anthropocentric elements of the earlier theories because the main definitions are couched entirely in *causal* terms without reference to a human or other agent; (2) (as a consequence) a farewell to the reductive aim of these theories; (3) a specification of the properties a manipulation must have in order to count as a genuine intervention to test causal claims.

The earlier agency accounts we looked at suffered from the deficiency that they are hard to reconcile with the strong belief we hold that there are causal relations without human agents. In particular Menzies and Price bite the bullet on this issue and understand causation as a secondary quality analogously to colour. What, then, happened when the dinosaurs got extinct?

Once more, we can read this question in at least two ways: what happened in the objects (or organisms), and how do we understand the phrase "a shower of asteroids caused the extinction of the dinosaurs"? If we believe that causation is something in the world, Menzies and Price's answer to the first question seems almost trivially unsatisfactory. But Woodward has also worries about their answer to the second question. Quite obviously we cannot bring about a shower of asteroids. Menzies and Price reply that we understand this case by analogy with cases in which we can bring about the adequate manipulation. How do we know that the analogy is reliable? Menzies and Price argue (p. 197, emphasis original):

> In its weakened form, the agency account states that a pair of events are causally related just in case the situation involving them possesses intrinsic [*i.e.*, non-causal] features that *either* support a means-end relation between the event as is, *or* are identical with (or closely similar to) those of another situation involving an analogous pair of means-end related events.

But this is exactly what Woodward denies can be done. If, say, we try to understand an earthquake using a computer simulation, we better get the *causal aspects* of the earthquake right, otherwise the simulation will be misleading. But if this is true, we have to give up on the goal to reduce the concept of causation to non-causal concepts. Woodward thinks that this is a price we have to pay in return for the adequacy of our theory. Furthermore, it is not too high a price to pay, as a philosophical theory can illuminate interesting interrelations between concepts without being reductive.

The greatest flaw in previous agency theories according to Woodward is, however, that despite all efforts they do not allow us to distinguish accidental from genuine causal regularities. To take a philosophical stock example (and the one preferred by Woodward), take the common cause relationship between the storm and the barometer reading. Agency theorists maintain that they can successfully distinguish between a genuine cause and a concomitant effect by appealing to their notion of an action: manipulate the barometer (*e.g.* nail the pointer to the dial), and the relation between the storm and the reading will be broken. However, in case of a genuine causal relation such as the atmospheric pressure and the storm, the relationship will continue to hold when the cause-variable is manipulated.

But this is not guaranteed, says Woodward. Suppose that our action to manipulate the barometer pointer is correlated with atmospheric pressure (by chance, say, or because our action is itself an effect of changes in pressure[26]). If this so happens, the spurious relationship between a change in the barometer reading and the occurrence of a storm may be mistaken for a genuine causal connection. Thus the fact that a free action that manipulates the putative cause is followed by a change in the putative effect is not a fool-proof sign of genuine causality. Therefore, more stringent conditions must be imposed on what counts as a test intervention.

Woodward's own account is motivated by reflection upon controlled experiments. At least some controlled experiments aim at establishing a causal law. These experiments often function by varying the putative cause-variable, and tracking the response of the putative effect-variable. Controlling here means to make sure that nothing else which can cause the putative effect causes it to change when the experimenter changes the putative cause. This includes both variables that operate independently of the intervention as well as variables which are causally related to it.

Woodward thus suggests the following definition of an intervention variable (p. 98, calling the intervention *I* and putative cause and effect *X* and *Y*, respectively):[27]

> (**IV**)
> I1. *I* causes *X*.
> I2. *I* acts as a switch for all the other variables that cause *X*. That is, certain values of *I* are such that when *I* attains those values, *X* ceases to depend on the values of other variables that cause *X* and instead depends only on the value taken by *I*.
> I3. Any directed path from *I* to *Y* goes through *X*. That is, *I* does not directly cause *Y* and is not a cause of any causes of *Y* that are distinct from *X* except, for course, for those causes of *Y*, if any, that are built into the *I*-*X*-*Y* connection itself; that is, except for (a) any causes of *Y* that are effects of *X* (*i.e.*, variables that are causally between *X* and *Y*) and (b) any causes of *Y* that are between *I* and *X* and have no effect on *Y* independently of *X*.
> I4. *I* is (statistically) independent of any variable *Z* that causes *Y* and that is on a directed path that does not go through *X*.

A "cause" is now defined as follows (p. 51):[28]

> *X* is a … cause of *Y* if and only if there is a possible intervention on *X* that will change *Y* or the probability distribution of *Y*.

The one-million-dollar question now is: What kinds of intervention are possible? An intervention has to be possible in exactly what sense? In order to avoid the charge of anthropocentrism, Woodward clearly rejects the idea that the relevant notion of possibility relates to what humans can do. In this, he follows the earlier agency accounts.

Now, there are different senses of "physical possibility". According to one, possible is what ever is consistent with the actual obtaining laws and initial conditions. But with the additional assumption of determinism (which he sometimes makes), it follows that only what is actual is possible. This would obviously be too strong a notion of possibility. A weaker notion demands only that the intervention be consistent with the laws of

[26] A reader who thinks that free will is incompatible with our actions being caused in this way may consider the reverse case where our action has the side effect of changing the storm variable through either changing atmospheric pressure or some other channel.

[27] A directed path is essentially a sequence of variables (X_1, X_2, …, X_n) where each predecessor causes its successor—a "causal chain" according to Lewis's terminology.

[28] This is in fact Woodward's definition of a "total cause". The difference between a total cause and a contributing cause is relevant in the context of cancellation cases where a variable influences another on two or more different routes in such a way that the different influences exactly cancel. For the present discussion, the distinction is irrelevant however.

nature and *some* set of initial conditions. This notion may still be too strong. In one of Woodward's own examples (pp. 129ff.), the moon is said to cause the tides. Clearly, there is no intervention possible in the sense that humans could do it (at least not at present) to change the moon's position in order to observe the subsequent change of the tides. But there may not even be a change in the initial conditions that results in an intervention that affects only the position of the moon and not the tides on any other route. Suppose there was a big comet knocking the moon out of its orbit into a different orbit. That comet would probably not only change the moon's shape and therefore its mass distribution and therefore have an independent effect on the tides but also have an independent effect on the tides via its own mass and through its effect on other bodies that affect the tides. It is of course thinkable that in this case there is an intervention consistent with the laws and some set of initial conditions that satisfies clauses (I1)-(I4). But, as Woodward himself recognises, there is nothing that guarantees that there always will be such interventions. He then argues that actual physical possibility is not so important but rather (a) whether it is coherent at all to say that there is an intervention that changes putative cause variable and (b) whether we have grounds to say that a resultant change in the putative effect variable was solely due to the intervention on the putative cause variable and nothing else. Thus, his notion of possibility is weaker still (p. 132):

> My conclusion, then, is that at least in circumstances like those in the above examples, we may meaningfully make use of counterfactual claims about what would happen under interventions, even when such interventions are not physically possible in either the strong or weak senses described above, and we can legitimately use such counterfactuals to elucidate causal claims along the lines suggested by **M** and **TC**.[29] In other words, the reference to "possible" interventions in **M** and **TC** does not mean "physically possible"; instead, an intervention on *X* with respect to *Y* will be "possible" as long as it is logically or conceptually possible for a process meeting the conditions for an intervention on *X* with respect to *Y* to occur.

This notion of relevant possibility appears entirely innocuous. But it also renders his theory as good as empty. If Woodward were to say that *A* is a cause if and only if it changes *B* under a possible intervention where "possible intervention" means "actually possible", he would hold a substantial theory (substantial because it could be false). An economics example will illustrate.

In 1958, LSE economists Alban Phillips hoped to find empirical support for the Keynesian idea that wage pressure depends on the tightness of the labour market (Phillips 1958). He investigated "whether statistical evidence supports the hypothesis that the rate of change of money wage rates in the United Kingdom can be explained by the level of unemployment and the rate of change of unemployment" (*ibid.*, p. 284). Phillips found that they were negatively correlated.

Now suppose that the negative relationship is not only statistical but genuinely causal: labour market conditions *cause* inflation (this is at any rate they interpretation Keynesians preferred). Further suppose that knowledge of this relationship gives us the idea that we might exploit it for policy purposes. We might, for instance, be led to attempt to tighten the labour market in order to keep inflation low. Now it is well possible that no such intervention exists. All interventions will have a number of effects on the economy because they influence agents' expectations, and therefore it is more than likely that inflation is affected as well. If in this world, there is no feasible

[29] Woodward's definitions of Manipulability Theory (p. 59) and Total Cause (p. 51). TC has been reproduced above.

intervention that changes unemployment without thereby also changing inflation through a different route, Woodward's theory is falsified.

However, Woodward doesn't require that such an intervention be actually feasible. He wants it only to be logically and conceptually possible. But without restrictions on the range of admissible ideal interventions, the theory is not informative. Suppose we wanted to test whether labour market conditions cause inflation. Our first attempt yields an unsatisfactory answer because we observe that our test intervention shook up the system of causal laws, and now the statistical relation we have been observing hitherto does not obtain any more. Does that falsify our causal hypothesis? Of course not, since our intervention was too "ham-fisted" as Elliott Sober would say (*cf.* Woodward 2003, p. 129). If we are lucky, the economy reverts back to the old situation where the relationship hold, and we try another (series of) test(s). Alternatively, we try other tests on a different economy.[30] Suppose now that these tests yield a *positive* result. What are these tests evidence for?

Woodward would say that they provide evidence for the counterfactual claim: "Had we intervened on unemployment using an ideal intervention (according to IV), the inflation rate would have changed". Hence we are entitled to say that unemployment causes inflation. But this seems a roundabout way of doing things. Why not get rid of the counterfactual detour?

Thus both horns of the dilemma seem unattractive (which makes it a dilemma after all). Under the strong interpretation of "possible" Woodward's theory is false; under the weak interpretation, it is uninformative and baroque.

Luckily, this way of presenting the story immediately suggests a way out of the dilemma. Reductive theories attempt to define the causal relation in terms of non-causal, philosophically less problematic concepts. Among the reductive theories, agency theories in particular suffer from the deficiency that causal relations obtaining in situations where we cannot manipulate putative causes must be said to be either non-existent or incomprehensible. Woodward improves on this situation by withdrawing from the goal of providing a reductive theory. Because he uses the concept of cause on both sides of his definition, he is entitled to use causal knowledge when drawing analogies between situations where we can intervene and those where we can't. However, Woodward still tries to provide a *definition* of cause, that is, a characterisation which aims at being true of all cases that fall under the concept. The deficiency in his account is that the definition is either false or uninformative. The natural thing to do is to go a step further and withdraw also from the aim of giving a characterisation which is meant to be true of all cases.

We could, for instance, stop asking the metaphysical question and focus on methodology instead. This strategy would have the additional advantage of pre-empting any debate about potential counterexamples. Any given test will be valid only under certain conditions. Regarded as a test for causality rather than a definition, Woodward's criteria show for example that the test is only available for situations in which intervention variables of the right kind exist. But in this way he does not have to demand that for any causal relation there must be an intervention of precisely this kind.

This is done, in their own respective ways, by Kevin Hoover and Nancy Cartwright among others. Hoover states explicitly (Hoover 2001, p. 23):

> The central thesis of this book is that what is implicit in the strategy of the probabilistic approach ought to be explicitly embraced: *Causal structures are fundamental.* Probabilistic accounts are misrepresented when they are seen as elucidating the concept of causality. In

[30] How such an alternative test can look like in the case of testing cognitive theories on laboratory rats, see Bogen forthcoming.

> fact, they are useful not for conceptual analysis, but as part of the epistemology of inferring causal structure from observations.

Hoover then goes on to introduce his own theory of causal inference (as opposed to a theory of what the causal relation *is*). But as such, he comes fairly close to Woodward's criteria. Suppose we want to infer the causal direction between taxes and government spending (ch. 9): are the taxes set in order to pay for (a fixed) expenditure, or does spending react to (a fixed) amount of taxes collected? In order to test for either causal claim, Hoover suggests identifying a historical period in which one of the processes was disrupted, but not the other. Suppose, for instance, the Reagan tax cuts affected only the tax receipts but not the expenditure (except, possibly, via taxes). If we can now further suppose that there is no common cause of the tax cut and expenditure (Hoover leaves this implicit but the above discussion has shown that we need this requirement) and we find a disruption of the spending process, too, we can argue that causal direction runs from taxes to spending.

A similar methodology can be found in the so-called natural experiments movement in econometrics. Proponents of this movement attempt to find "natural" situations, that is, situations that have not been manipulated by the researcher, which nonetheless mimic an experimental situation. Card and Krueger 1995, for example, attempt to measure the effect of an increase in the minimum wage level on employment. For this, they exploit a change in the minimum wage legislation that occurred in New Jersey in April, 1992. Card and Krueger note that New Jersey is economically linked very closely to bordering Eastern Pennsylvania. Hence, they can assume that all other relevant causes that may affect employment have the same influence on both areas. In their study, they carefully check whether the introduction of the new legislation is itself correlated with a factor that may change employment (*cf.* Woodward's condition (I4)). The experimental population they are concerned with consists of fast food restaurants. Because fast-food restaurants do often pay minimum wages, and because most restaurants comply with a change in legislation, they can also argue that the legislation (an "intervention") causes (paid) minimum wages (Woodward's condition (I1)). Further, there is no reason to suppose that this change in legislation affects employment via any other route (Woodward's condition (I3)).[31]

Card and Krueger are not the only economists which follow this methodology. There is in fact a whole industry of econometricians which attempt to exploit natural experiments (I have written about the natural experiments movement in Reiss 2003; for a comparison of the assumptions used by proponents of natural experiments and Woodward's theory, see Reiss forthcoming b). Natural experiments, too, come with their limitations. The first and foremost is of course that they require a great deal of background information. In particular in economic matters where it is hard to isolate systems from outside influences, the amount of background information needed can be prohibitive. It can also always give critics cannon fodder. In one instance, a researcher

[31] Fulfilling Woodward's condition I3 would imply that all restaurants perfectly comply. For practical applications, this condition is too strong; it is important only that whether or not restaurants comply isn't correlated with other causes of employment. Woodward indeed regards his set of conditions as idealised and writes: "[M]y intention is *not* to inquire about the most general set of conditions that are necessary for an experiment to provide information about the efficacy of the [putative cause]. Many imperfect or nonideal experiments can provide [this] information if the right conditions are satisfied, as can nonexperimental investigations. Instead, my interest is in formulating a notion of intervention that fits with the project pursued [in his book], that of providing truth conditions for claims about… causal relationships… by appealing to facts about what would happen under interventions", p. 95, emphasis original).

attempted to exploit the Vietnam draft for a natural experiment aimed at measuring the impact of veteran status on civil earnings because draft decisions were made on the basis of a random sequence number. A critic argued that the random sequence number does not fulfil the criteria of a test intervention because employers are more likely to invest in an employee's training if he has got a high number and thus is not very likely to be drafted. Thus the number affects earnings via a channel different from the intervention-putative cause-putative effect link and condition I3 is violated (about this, too, I have written in Reiss 2003).

The next limitation concerns the applicability of results from the case examined to hitherto unobserved cases. Causal claims are usually supposed to underwrite policy decisions. Card and Krueger—eventually—aim to show that raising minimum wages isn't necessarily a bad idea full stop, not merely that it wasn't a bad idea in New Jersey. But are results of that kind exportable to other situations? Are they, in experimental psychologists' lingo, *externally valid*? Nothing in the study referred to above is evidence for this further desideratum (I have written about external validity of natural experiments in econometrics at length in Reiss forthcoming a).

A third limitation is that we learn from these experiments only about the *average causal effect* of one variable on another. Often, however, we will be interested in other kinds of effect. Suppose a drug trial reveals that treatment *T* shortens the length of some disease by 8 days. This means that *on average* patients in the treatment group suffered 8 days less than those in the control group. But there may be large variations within the groups. For example, the drug may be perfectly effective for young otherwise healthy people but ineffective for elderly people with a poor constitution (or vice versa). As a patient I am interested in how the drug will work *for me*. The average effect may be completely uninteresting in my decision whether to take *T* or alternative treatment *T*'.

These ideas bring us back to the beginning of this manuscript. Francis Bacon was very concerned with how we can reliably learn from experience, and he devised a carefully structured schema of experimental research for causal inference. Almost four hundred years later, some philosophers still regard causality as very closely tied to experiments. James Woodward goes as far as understanding (ideal) experimental conditions to be constitutive of causal relations. Others think of experiments more in terms of tests for causality rather than their essence. But this is a good example of the relevance of methodology for metaphysics (and *vice versa*).

12 Conclusion and Outlook

References

Armstrong, David 1978, *Universals and Scientific Realism*, Cambridge: CUP

Beauchamp, Tom and Alex Rosenberg 1981, *Hume and the Problem of Causation*, Oxford: OUP

Bickel, P. J., Hjammel, E. A., and O'Connell, J. W., 1975, "Sex Bias in Graduate Admissions: Data From Berkeley", *Science* 187: 398-404

Card, David and Alan Krueger (1995), *Myth and Measurement: The New Economics of the Minimum Wage*, Princeton: Princeton University Press

Cartwright, Nancy 1979, "Causal Laws and Effective Strategies", *Noûs* **13**, reprinted in Cartwright 1983, 21-43

Cartwright, Nancy 1983, *How the Laws of Physics Lie*, Oxford: Clarendon

Cartwright, Nancy 1999, *The Dappled World: A Study of the Boundary of Science*, Cambridge: CUP

Cartwright, Nancy 2001, "What's wrong with Bayes' Nets", *Monist* **82**, 242-64

Cartwright, Nancy 2002, "Against Modularity, the Causal Markov Condition, and Any Link between the Two: Comments on Hausman and Woodward", *British Journal for the Philosophy of Science* **53**(3), 411-53

Cohen, Elliot David 1977, "Hume's Fork", *The Southern Journal of Philosophy*, Vol. XV., No. 4

Collins, John 2000, "Preemptive Preemption", *Journal of Philosophy* **97**, 223-34

Dowe, Phil 1992, "Wesley Salmon's Process Theory of Causality and the Conserved Quantity Theory", *Philosophy of Science* **59,** 195-216

Dowe, Phil 1995, "Causality and Conserved Quantities: A Reply to Salmon", *Philosophy of Science* **62**, 321-333

Dowe, Phil 2001, "Is Causation Influence?", draft available at http://www.utas.edu.au/docs/humsoc/philosophy/Phil/causation.html

Dupré, John 1984, "Probabilistic Causality Emancipated", in Peter French, Theodore Uehling, Jr., and Howard Wettstein (eds), *Midwest Studies in Philosophy IX*, Minneapolis: University of Minnesota Press, 169-175

Elga, Adam 2000, "Statistical Mechanics and the Asymmetry of Counterfactual Dependence", *Philosophy of Science* **68** (Proceedings), S313-S324

Fair, David 1979, "Causation and the Flow of Energy", *Erkenntnis* 14, 219-250

Gasking, Douglas 1955, "Causation and Recipes", *Mind* **64**, 479-87

Goodman, Nelson 1954, *Fact, Fiction and Forecast*, Harvard: Harvard University Press

Glymour, Clark 1992, *Thinking Things Through*, Cambridge (MA): MIT Press

Hall, Ned 2000, "Causation and the Price of Transitivity", *Journal of Philosophy* **97**, 198-222

Hall, Ned 2003, "Two Concepts of Causation", in Collins, John, Ned Hall and Laurie Paul (eds), *Causation and Counterfactuals*, Cambridge (MA): CUP

Hausman, Daniel and Woodward, James 1999, "Independence, Invariance and the Causal Markov Condition", *British Journal for the Philosophy of Science* **50**:4, 521-584.

Hesslow, Germund 1976, "Two Notes on the Probabilistic Approach to Causality", *Philosophy of Science* **34**, 290-2

Hitchcock, Chris 1995, "Discussion: Salmon on Explanatory Relevance", *Philosophy of Science* **62**, 304-20

Hitchcock, Chris 2003, "Of Humean Bondage", *British Journal for the Philosophy of Science* **54**, 1-25

Hoover, Kevin 2001, *Causality in Macroeconomics*, Cambridge: CUP

Howson, Colin 2000, *Hume's Problem: Induction and the Justification of Belief*, Oxford: OUP.

Hume, David 1928/1739, *A Treatise of Human Nature*, ed by L.A. Selby-Bigge, Oxford: Clarendon

Hume, David 1936/1748, *Enquiries Concerning the Human Understanding and Concerning the Principles of Morals*, ed by L.A. Selby-Bigge, Oxford: Clarendon

Humphreys, Paul forthcoming, "Theories of Causation and Explanation: Contingently or Necessarily True?", *Causalidad y explicación. En homenage a Wesley Salmon*, Barcelona: The Autonomous University Press

Kitcher, Philip 1989, "Explanatory Unification and the Causal Structure of the World", in P. Kitcher and W. Salmon (eds), *Minnesota Studies in the Philosophy of Science Volume XIII*, Minneapolis: University of Minnesota Press, 410-505

Kvart, Igal 2001, "Counterexamples to Lewis' 'Causation as Influence'", *Australasian Journal of Philosophy* **79**(3), 409-421

Lakatos, Imré 1970, "Falsification and the Methodology of Scientific Research Programmes", in Imré Lakatos and Alan Musgrave (eds), *Criticism and the Growth of Knowledge*, 91-196

Lewis, David 1993/1973, "Causation", in Ernest Sosa and Michael Tooley (eds), *Causation*, Oxford: OUP, 193-204

Lewis, David 1979, "Counterfactual Dependence and Time's Arrow", *Nous* **13,** 455-476

Lewis, David 1986, "Postscripts to 'Causation'", in *Philosophical Papers* vol. II, Oxford: OUP, 172-213

Mackie, John 1974, *The Cement of the Universe: A Study of Causation*, Oxford: Clarendon

Malinas, Gary and John Bigelow, "Simpson's Paradox", *The Stanford Encyclopedia of Philosophy* (Spring 2004 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2004/entries/paradox-simpson/>

McDermott, Michael 1995, "Redundant Causation", *British Journal for the Philosophy of Science* **46**, 523-44

Mill, John Stuart 1874, *A System of Logic*, New York: Harper

Pearl, Judea 2000, *Causality: Models, Reasoning, and Inference*, Cambridge: CUP

Noonan, Harold 1999, *Hume On Knowledge*, London: Routledge

Reichenbach, Hans 1956, *The Direction of Time*, Berkeley and Los Angeles: University of California Press

Reiss, Julian 2003, "Instrumental Variables, Natural Experiments and Inductivism", *Causality: Metaphysics and Methods Technical Report* 11/03, CPNSS, LSE

Reiss, Julian forthcoming a, "Social Capacities", in Luc Bovens and Stephan Hartmann (eds), *Nancy Cartwright's Philosophy of Science*

Reiss, Julian forthcoming b, "Causal Instrumental Variables", *Philosophy of Science*, PSA 2004 proceedings

Reiss, Julian forthcoming c, "The Contingency of Theories of Causality" (in Spanish), *Causalidad y explicación. En homenaje a Wesley Salmon*, Barcelona: The Autonomous University Press

Reiss, Julian forthcoming d, *Beyond Spiders and Ants: The Empiricist Stance in Economic Methodology*, book manuscript, LSE

Reiss, Julian and Nancy Cartwright 2003, "Uncertainty in Econometric: Evaluating Policy Counterfactuals", *Causality: Metaphysics and Methods Technical Report CTR 11/03*. Available at http://www.lse.ac.uk/Depts/cpnss/proj_causality.htm

Russell, Bertrand 1913, "On the Notion of Cause", *Proceedings of the Aristotelian Society* **13**, 1-26

Russell, Bertrand 1948, *Human Knowledge*, New York: Simon and Schuster

Ryan, Alan 1987, *The Philosophy of John Stuart Mill*, London: Macmillan

Salmon, Wesley 1977, "An 'At-At' theory of Causal Influence", *Philosophy of Science* **44**(2), 215-24

Salmon, Wesley 1984, *Scientific Explanation and the Causal Structure of the World*, Princeton: Princeton University Press

Salmon, Wesley 1994, "Causality Without Counterfactuals", *Philosophy of Science* **61**, 297-312

Schaffer, Jonathan 2000, "Trumping Preemption", *Journal of Philosophy* **97**, 165-81

Spirtes, Peter, Clark Glymour, and Richard Scheines. 2000,. *Causation, Prediction, and Search*, 2nd ed., New York: Springer-Verlag

Suppes, Patrick 1970, *A Probabilistic Theory of Causality*, Amsterdam: North-Holland

Urbach, Peter 1987, *Francis Bacon's Philosophy of Science*, LaSalle: Open Court

Urbach, Peter and John Gibson 1994: *Francis Bacon: Novum Organum*, Chicago and La Salle (IL): Open Court

Woodward, James 1997, "Explanation, Invariance and Intervention", *Philosophy of Science*, Supplement to **64**:4, pp. S26-41

Woodward, James 2003, *Making Things Happen*, Oxford: OUP

Worrall, John 2002, "*What* Evidence in Evidence-Based Medicine?", *Philosophy of Science* **69** (Supplement), S316-S330