Approximate truth and causal strength in science

Robert David Northcott

London School of Economics and Political Science

PhD thesis

Department of Philosophy, Logic and Scientific Method

2004

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

In Chapter One I motivate the search for an account of approximate truth as being the way to make sense of how our best scientific theories are simultaneously false but useful, and of how the same theory (even a true one) varies in its usefulness depending on context. I evaluate existing approaches and find that they fail – among other reasons – because they are unable to accommodate how two errors of similar *logical* seriousness nevertheless may have greatly different implications for approximate truth. The only way round this is some form of weighting scheme across logical statements that is motivated by extra-logical criteria. The little existing work along these lines suffers from insufficient generality, and I suggest instead a weighting scheme based on the notion of causal strength.

In Chapter Two I develop the details of following such a prescription. It turns out to be crucial to highlight a hitherto underappreciated dichotomy between what I label the 'ontological' and 'empirical' senses of approximate truth. After outlining the practical advantages of my approach I discuss a number of technicalities, including several that confound all previous approaches. (I also outline an exact formal definition in an appendix.) Finally, I tackle the vexed issue of comparing two models with incommensurable ontologies. One of the results of the complicated discussion is that no sense can be made of science in general getting nearer the truth, only sense made of particular models getting nearer the truth of particular explananda.

In Chapter Three I flesh out the notion – key for my scheme – of causal strength, giving a formal definition and sorting through the numerous necessary technicalities. I also explain how straightforward śénse can be made of causal strengths even in cases of interactive effects and also even in cases where two causes are apparently incommensurable.

Table of Contents

Dedication

To my parents,
for their exceptional love and encouragement

Chapter One – How best to approach approximate truth?

1-1) Introduction – the view from economics

Initial motivation. Introductory example. A first notion of approximate truth: we want a context-specific account. Different motivations.

1-2) The logical-similarity approach

General agenda. First attempts. Oddie's approach for the propositional case. Niiniluoto's approach for the propositional case. Beyond the propositional case. Recent extensions.

1-3) Weaknesses of the logical-similarity approach

Language-dependence: Miller's objection. Pure logic alone is not enough. Language and reality. Stubbornly false theories. Two measures in one. Summary.

1-4) Miller's objection revisited

Relevance to our own eventual scheme. Mormann and conventionalism. Where does this leave us?

1-5) Some other possible approaches

Possible worlds. Structurelikeness. Partial truth.

1-6) First ontological approaches

Introduction. Giere as a precursor. Type-hierarchies and verisimilitude. Smith's geometric ontology. Barnes's approximate causal explanation.

1-7) Idealisation, laws of nature

Idealisation and approximate truth. The problem of legisimilitude. Our solution.

1-1) Introduction – the view from economics

Initial motivation

I started with a basic intuition brought with me from economics. When tackling a real-world problem in that subject, in practice the most important thing is not to try to devise an *exactly* true model. Rather, the key is to be more modest and to settle merely for a true *enough* one. The success of a piece of economic analysis depends on a careful selection of the right model for the right problem, often borrowing piecemeal from several different models. It is a matter of getting sufficiently good – not necessarily perfect – matches. (The design of the 1994 US spectrum auction is a good example of how successful applied economics works in this way [McMillan 1994] [McAfee and McMillan 1994] [Guala 2001].)

The aim of this thesis is to develop a coherent metaphysics to support this intuition, because at the moment none exists. This will enable us to make philosophical sense of economic practice. Moreover, it will also prove to be a usefully original route into the assessment of approximate truth for science generally. In particular, we shall argue that the conception of approximate truth appropriate to economics also turns out to be much the best one for the rest of science too.

Introductory example

Imagine being Gareth Southgate at the Euro 96 football championships, about to take a penalty for England in the semi-final shoot-out against Germany. The problem facing Southgate is how to shoot his penalty. Let us simplify and give him just two options – either he can shoot at the left side of the goal or at the right. (In reality of course he would also be considering perhaps shooting at the centre of the goal, whether to shoot the ball low or high, whether to shoot it with pace or with accuracy, and so on.) The other person in this confrontation is the German goalkeeper, Koepke. He too has a choice, of diving to the left or to the right. Koepke hopes that Southgate shoots the ball to the same side that he dives, since this increases his chance of making a save. Of course for the

same reason Southgate hopes that Koepke dives the opposite side to the one he shoots. The question is: taking all factors into account, which side should Southgate shoot in order to maximise his chances of scoring the penalty, and which side should Koepke dive in order to maximise the chances of saving it? This is our real-life problem.

Thinking about this problem intuitively, it is hard to reach a definite answer. Suppose that Southgate prefers to shoot to the left rather than to the right. Then he might reason: I can shoot better to the left so therefore I should do that; but then again Koepke, knowing this, may therefore deliberately dive to that side and so I should in fact shoot to the right (the bluff); but then again Koepke might in turn anticipate this way of thinking and so dive to the right, and so I could in fact outwit him by shooting to the left after all (the double-bluff); but then again Koepke might also anticipate that, so... (triple-bluff); and so on and on *ad infinitum*. For his part, Koepke might also follow similar reasoning and so end up without any clear recommendation as to what to do either. Thus at the moment our analysis does not seem to have got us very far.

Suppose now we employ a new tool – game theory. A game theoretical analysis of this situation (or *game*) might proceed in the following way. We define two *players* in the penalty game, Southgate and Koepke. We define the *strategies* available to each of them, in this case respectively either shooting or diving to the left or right. We define the *payoffs* to each of them for each possible combination of strategies, in this case the probability of a goal being scored if Southgate shoots left and Koepke dives left, if Southgate shoots left and Koepke dives right, if... etc. Technically we also need to make other assumptions, such as that each player knows the parameters of the game (e.g. that Koepke knows that Southgate prefers shooting to the left, and knows by how much Southgate prefers it); that each knows that the other knows (and that the other knows that he knows, and that the other knows that he knows that the other knows, and that... etc); and that each player is *rational* in that he does his best to score (Southgate) or save (Koepke) the penalty, and is able to perform any necessary calculations for this.

Once all this 'structuring' of the original penalty problem has been done, and precise

values assigned to the various parameters, then game theory can, as it were, get to work. It is able to calculate precise optimal strategies for both Southgate and Koepke, cutting through the seemingly infinite chain of bluff and double-bluff described above. Of course this is the *raison d'etre* of game theory: to provide a rigorous treatment of just such seemingly intractable situations of strategic interaction. The particular optimal strategies in the penalty game would likely be probabilistic for each player. For example, Southgate's optimal strategy might be to shoot left two-thirds of the time and right one-third of the time, while Koepke's should perhaps be to dive left and right equally often. The exact probabilities would depend on, for example, just how much Southgate prefers shooting to the left, and just how much the goalkeeper can reduce the chance of the penalty being scored by diving to the correct side.

What precisely do we mean by an 'optimal strategy'? Here it means that given what Koepke is doing, Southgate can do no better than his optimal strategy; and the same in reverse, i.e. that given Southgate's strategy, Koepke can do no better than his optimal strategy. In plainer words, each is doing the best given what the other is doing. This is therefore the *only* kind of outcome in which neither player could be doing better by deviating to some alternative. For this reason, that outcome (known as a Nash equilibrium) would usually be reckoned a sensible solution for the penalty game. (There is much technical debate as to what in different contexts should be considered 'sensible'.)

Summing up so far, there are two sides to this whole exercise. On the one side is the real-life problem of the penalty, waiting to be analysed. On the other is our model (derived from game theory) of this penalty situation. The pattern of this example – trying to apply an internally rigorous but rather abstract theory to some real-life problem – is fairly typical of applied economics.

So what is the pertinence of the example to this thesis? Suppose we wanted to criticise this application of game theory to football. Criticism would be unlikely to focus on the rigour of the calculations. Rather, it would likely focus instead on issues *external* to the actual nuts and bolts of the analysis. Perhaps it might be argued that each player's choice

of strategies had been unrealistically simplified to just two options each. Although game theory analysis with, say, nine strategies each is possible in principle, in practice an analytical solution would likely be too complex actually to calculate. Or what about psychological games, such as Koepke putting off Southgate by delaying the taking of the kick? Again, in principle this too might be analysable with game theory. But in practice the problem would be so difficult to parameterise that only many arbitrary assumptions would enable an analysis at all, thereby reducing the relative significance of the analysis itself. Suppose next that neither player knows precisely or for certain what all the parameters are. For example, Koepke might think: 'I know Southgate prefers shooting to his left, but by how much? And how is that affected by the pressure he is under? And did he injure his right leg in that tackle five minutes before the end? And has his confidence been shaken by the mistake he made five minutes before that?' etc. A game theorist might retort that in fact uncertainty can been incorporated into game theory, enabling a formal analysis even of these considerations. But again it might be argued in return that to do this we need to postulate subjective prior belief distributions, requiring arbitrary assumptions. The analysis's conclusions would in turn be critically sensitive to this arbitrariness. Hence the real driver of the conclusions would again not be the game theory so much as the arbitrary judgments that preceded it. Perhaps game theorists might riposte that although all these objections *could* be important, in fact in this example they are not. And so on.

We do not opine here whether game theory actually is useful in this particular case. Rather, the important point is *what determines* whether it is or not. The crucial consideration is not whether this or that internal game theoretical procedure is rigorous or clever. The attributes of game theory in isolation are not the issue. Instead what matters is whether those attributes are *applicable* to this particular problem. Thus game theory is good at studying rational strategic interaction. But our hypothetical critics above were arguing that perhaps players' emotions, rather than their rational strategic interactions, were an important element in the penalty problem, and that the game theory model was poor at incorporating these emotional factors. In other words, the criticisms amounted to pointing out a poor match between the attributes of the model and the attributes important

to the actual real-world problem. In the terms we shall use later, the model was not picking out the most important causes actually at work.

Meanwhile, our hypothetical defenders of game theory were accepting this *same* agenda for dispute. Thus they argued that game theory was in fact able after all to incorporate emotional factors, rather than disputing why it should even be required to do so. No one was arguing about what the model *ought* to be able to do. Rather, the disputes concerned the substantive point of whether or not it actually *could* do those things. There is thus a certain methodological agreement between all sides in this hypothetical case – namely, that when assessing a model's usefulness, we should be concentrating above all on whether or not it is well-suited to capturing the particular attributes actually important in the real-life problem. What matters is not a model or problem in isolation, but rather the degree of match between the two.

A first notion of approximate truth: we want a context-specific account

This example suggests that the key to the success of a piece of (applied) social science is the *appropriateness* of the chosen theory for the problem being studied. What matters is not the theory or problem in itself. Rather what really counts is the connection or otherwise between the two. This implies that we cannot directly judge a theory's usefulness in the abstract, but instead only individual applications of that theory to specific problems. It follows that the usefulness or otherwise of a theory must be *application-specific*, or *problem-specific*. Put another way, usefulness must be assessed at the level of a context-specific *model*, not general theory.

The point becomes obvious if we consider trying to apply game theory to other problems. It is arguable whether it is or is not a useful tool for analysing the penalty case above. But suppose we take instead a different issue, such as identifying the winner of the next general election. Some might argue that even here game theory in practice can make a useful contribution, but it seems clear that this new election problem is considerably more complicated than the penalty one. Accordingly it seems considerably more difficult to

'get a grip on' on it using game theory. Or consider how to mend a faulty disk drive on a computer. It is surely uncontroversial that an appropriate model here would be one of the relevant computer hardware, and that knowledge of game theory would be no direct help. The point of course is that the usefulness of game theory cannot be judged in isolation, but rather is problem-dependent. And the same applies to the usefulness of models of computer hardware, which presumably would not be so helpful when analysing our penalty problem.

Of course, ultimately any model can capture only some aspects of the world, leaving other aspects out. If these left-out aspects are important to a particular problem, then it follows that that particular problem may not be much elucidated by that particular model. So no tool (at least in social science) is applicable universally. But if a theory is not a theory of everything, then it can only be a theory of some things. And then it must follow that its usefulness is not universal, but is in fact problem-dependent.

Trying to apply economic theory often leads to a situation similar to our penalty example. There is a contrast between abstract mathematical theory on one side and a perhaps 'messy' real-world problem on the other. We typically derive, from background theory and other considerations, a formal model of the particular real-world situation. No one believes this model to be a literally or wholly true representation of the reality, yet at the same time it is hoped that it might somehow capture what is important or interesting about it. So there is an apparent tension. On the one hand, an economic model seems to be aiming at something less than the whole truth of a situation. So already an economic modeller seems to be aiming low. Yet on the other hand, there is still the hope that these imperfect models nevertheless might somehow tell us something about the world, despite not being wholly true. It seems that (full) truth has been dropped as a necessary requirement for success in economics. Something weaker is considered sufficient.

Let us illustrate this vital argument schematically:-

Point 1: Economics is sometimes useful.

For example, at the height of the BSE beef scare the price of British beef decreased. The economic explanation is that a drop in demand led to a drop in equilibrium price. It seems intuitively clear that this theoretical economic account captures well a certain part of what was actually going on in this case. Health worries did indeed lead many consumers to avoid buying beef, at least until big price reductions came on stream.

Point 2: Yet all economic models are empirically false.
All (if taken literally) are easily falsified – this point is not controversial. For instance, almost any model in the mainstream modern neoclassical tradition postulates human agents to be rational maximisers of particular utility functions. Yet it is well established that real humans do not always behave in such a way. Hence experiments could easily be set up (if desired) which would contradict this basic assumption.

Conclusion: Full truth is not the key issue.
Points 1 and 2 together imply that while abstract mathematical economic theory empirically is never literally true, it still (sometimes) seems interesting and insightful nevertheless. In other words, it can be simultaneously false but useful. Consequently, a focus just on truth/falsity seems largely to miss the point.

It is evident that economic theory is sometimes useful, sometimes not. If a theory is typically more useful in some contexts than in others, then it seems there is a sense in which it is capturing more of the truth of some contexts than of others. So there must be some entity, connected to truth, which is varying between different applications of a theory. Yet this entity cannot be simple truth/falsity itself since we saw that, strictly speaking, economic theory is always false and that does not vary. It cannot be some kind of 'local truth' either, since even in contexts of successful application a theory is still, strictly speaking, false. So it is natural at this stage to follow intuitive usage and to label our concept **approximate truth**. (No substantive philosophical definition is yet being offered here, we are merely establishing the motivation for this thesis.) Then in applying economic theory we are in fact seeking to maximise its approximate truth with respect to some particular problem. This degree of approximate truth will of course vary from

application to application.

Some notes: first, we are still interested in truth as the goal of science, albeit more as an ideal target to aim for than as something actually achieved. But a model can still be useful even if only approximately true. Second, approximate truth is a goal of models rather than general theories. And third, nowhere in this thesis shall we offer any particular analysis of truth itself. It turns out that none of the main debates turn on our position with regard to that issue – which is probably why, with minor exceptions (sections 1-5 and 1-6), no one else in the literature has addressed it either.

Summing up: the usefulness of an economic theory is *context-specific* since the same theory is often successful in one area but unsuccessful in another, depending on how *approximately true* of any particular context it is. Since (in social science at least) theories are never wholly true, a narrow concentration on truth/falsity fails to capture the factors that actually determine the extent of scientific success. Therefore what we want is a context-specific account of approximate truth.

Different motivations

It is important to realise that there are other possible motivations for studying approximate truth, which indeed have been the actual ones of previous work in the field. In particular, a lot of that work has had its source in the general debate between realism and anti-realism. Many realists would like a way of expressing the intuition that our best theories do actually refer to genuinely existing entities in the world. Since it is also accepted that most if not all of our theories are not literally true, the only way to salvage a realist interpretation of them is then to claim instead that they are approximately true. In other words, a satisfactory account of approximate truth is seen as important, perhaps even essential, to the realist position (or at least to realism about theories). (See for example [Putnam 1975], [Newton-Smith 1981], [Laymon 1982], [Miller 1987], [Boyd 1990] and [Psillos 1999].) However, although noting the current lack of a satisfactory account, many authors have nevertheless maintained their realism even in its absence,

and often even when not really offering any substantial new account themselves. But it is providing just such accounts for just this purpose of buttressing realism that has been the motivation for perhaps the most cited works in the literature [Oddie 1986] [Niiniluoto 1987] [Weston 1992].

Related to the general realism debate, is a notion of scientific progress that sees successive theories as achieving closer and closer approximations to the truth – 'convergent realism'. Thus Newton's laws are taken to be somehow closer to the truth, and hence better, than predecessors such as those of Aristotle. A convergent realist might see the sequence from Aristotelean through Galilean and Newtonian and finally relativistic mechanics, as representing better and better guesses at the actual true mechanics. But of course, in order to put philosophical flesh on the intuitive bones of what it means for these guesses to be 'better and better', we then need some account of approximate truth. Authors sceptical of such convergent realism can point to the lack of any such agreed account as a sign that their scepticism is justified [Kuhn 1962] [Laudan 1984]. And not surprisingly, a desire to underpin the notion of scientific progress (and thence realism) has in turn been a spur behind much work in the field [Aronson, Harre and Way 1994] [Barnes 1995].

A variant on this motivation is the critical rationalist position. This of course disputes that we are actually gaining any positive knowledge with each successive theory, preferring to speak instead merely of theories being tested and then discarded if falsified. Nevertheless, it has remained concerned with trying to explicate a realist sense in which science might be said to progress. This was the motivation of Popper, the founder of the modern literature on approximate truth [Popper 1963] [Popper 1972], and also of Miller, one of its main practitioners since [Miller 1975] [Miller 1994].

So the intellectual provenance of most of the work on approximate truth has lain in concerns drawn from general philosophy of science, which have tended to reflect a certain concentration on natural science, and especially physics, rather than social science. These various motivations matter because they suggest a natural agenda and set

of methods with which to approach the issue. Can these methods deliver a satisfactory result? What can we learn from them? In contrast, we derived our own desiderata from a consideration of the practice of economics – a social science. We shall eventually argue that, ironically, these latter desiderata may in the end prove the most appropriate ones for natural science too. But first let us see how far we can get by following, so to speak, the more traditional agenda.

1-2) The logical-similarity approach

General agenda

The first modern recognition of the issue of approximate truth, and the first serious attempt to address it, came from Popper in the 1960s [Popper 1963], [Popper 1972]. On the one hand his general philosophy of science had a strong realist view that the aim of science is truth, but on the other he recognised that even our best theories are usually false. Therefore to make sense of the notion of progress in science, it becomes necessary to be able to speak meaningfully of science moving 'closer to the truth' or of achieving a 'better approximation to the truth'.

Popper termed the concept he was looking for *verisimilitude*, and thus pictured progress in science as being a tale of our theories' increasing verisimilitude. This original motivation of Popper's – of giving substance to the notion of scientific progress – immediately gave verisimilitude two important characteristics. First, it was perceived as being an objective logical notion, and hence nothing to do with epistemic notions such as uncertainty. Popper wished to find a way of expressing the intuition that a theory actually is nearer the truth, independent of whether or not we happen to know that it is.

Second, verisimilitude was to incorporate both accuracy and comprehensiveness, that is

to say incorporate two different concepts in the one measure. A trivial tautology is perfectly accurate but scientifically useless; by contrast, Newton's laws are often inaccurate but scientifically hugely valuable. So in order to avoid ranking the tautology above Newton, we need to take on board the notion of comprehensiveness as well as mere accuracy – how *much* does our theory seek to explain? It is in this latter consideration of course that the worth of Newton's laws reveal themselves since they help to explain a huge range of phenomena, albeit imperfectly. In Popper's terminology, good science should make 'bold' conjectures. At the same time accuracy must still remain an important criterion too, of course, since no one is interested in conjectures that are bold but fantastical.

This agenda, although arising originally only out of the specific philosophical concerns of the Popperian program, has proved very persistent. In particular, most work has been done in a formal context, providing definitions of distance measures between logical statements and so forth. In addition, most authors have tackled both accuracy and comprehensiveness together, so as to be able to provide a measure by which to adjudge general theories as opposed merely to specific applications of them.

Our own approach will follow Popper in aspiring to analyse the issue as an objective feature of the world. Nevertheless, it will also eventually deviate in important respects from the Popperian exemplar (section 1-1). First, we shall prefer to conceive of approximate truth in *ontological* rather than logical terms. Second, our scheme covers only accuracy not comprehensiveness. This is because, guided by our conclusions from section 1-1, I am interested not in the approximate truth of general theories but only in the approximate truth of specific applications of those theories. I therefore do not attempt to make sense of the notion of comprehensiveness. Instead our approach will lead us more generally to a rather closer analysis of the concept of (application-specific) *relevance* than has been given before. By this route we can reach a definition of context-specific approximate truth. (The details of our scheme are not presented until chapter 2. This is merely to note from the start its deviation from some of the fundamental assumptions of previous work.) But first we must examine more closely just how the

traditional approach comes unstuck.

First attempts

Consider two theories, for instance N ('Newtonian') and E ('Einsteinian'). Suppose we focus on their deductive consequences, that is to say the theories are *defined* logically speaking by a deductively closed class of statements in some language. If we wish to find a way of saying that E is closer to the truth than N, one strategy would be to examine the number of true and false consequences of each theory. Then we could say that E is closer to the truth than N if it has more true, and fewer false, consequences than N (with mere equality of course permitted in one of these cases). Unfortunately this strategy fails immediately since the cardinalities of all consequence classes even in simple languages is countably infinite, and so no theory ever has more or fewer true consequences than any other and no false theory more or fewer false consequences. (Moreover, we would only be counting atomic consequences and this is strongly language-variant – on which issue see section 1-3.) Therefore any definition of verisimilitude that works by comparing numbers of true and false consequences, would seem to be hopeless.

Alternatively, we might restrict our definition to those cases where we can express the same intuition set-theoretically instead. Thus E has greater verisimilitude than N if and only if all N's true consequences are a subset of E's true consequences, and all of E's false consequences are a subset of N's false consequences, at least one of the subset relations being strict. That is, for T (F) the set of true (false) statements in some language, E has greater verisimilitude than N if and only if:

$$N \cap T \subseteq E \cap T \quad \text{and} \quad E \cap F \subseteq N \cap F$$

again at least one of the subset relations being strict [Popper 1972].

Unfortunately, it turns out that on this definition no false theory can ever have more verisimilitude than any other false theory [Miller 1974] [Tichy 1974]. So, for example, verisimilitude would give us above no means to show the (presumed) superiority of relativistic to Newtonian theory. This devastating result of course strikes at the heart of

the very purpose for which the definition was formulated. True theories would still have more verisimilitude than false ones, but this of course is no real help since perfectly true theories seem to be unhappily thin on the ground (especially in social science). (At the end of [Miller 1974], Miller also raises what has become known as the 'canonical objection' [Brink 1989, p186] to the whole verisimilitude enterprise, namely the problem of language invariance. But we save discussion of that until section 1-3.)

However, there remains an apparently rather more promising line of attack here: instead of counting consequences, we can try instead to construct definitions of logical *similarity*. (This is what [Niiniluoto 1998] calls the 'second period' of the verisimilitude research program.) The idea is based on the assumption (which may be questioned – see section 1-3 below) that the truth, or at least the relevant truth, can be expressed in any language as a set of true propositions. A theory can be expressed as another set of propositions. The theory's verisimilitude is then seen as the logical 'distance' between the theory and the truth, that is the distance between these two sets of propositions. The goal is to construct definitions of distance that incorporate desired notions of similarity. This idea is most easily illustrated, and hence later assessed, by outlining the two main developments of it, taken from [Oddie 1986] and the exhaustive [Niiniluoto 1987].

Oddie's approach for the propositional case

The full versions of Oddie's and Niiniluoto's accounts are generalised to arbitrary first-order languages (see below). However, following Brink, the basic ideas can be conveniently explained with reference to what has become known as the canonical example [Brink 1989]. Imagine a very simple weather-language featuring no predicates and only three primitive sentences. These sentences are 'it is hot' (or 'h'), 'it is raining' (r), and 'it is windy' (w). Assume all three sentences are true, so that the whole truth, so far as this language can grasp it, is h&r&w. This is the one true basic conjunction. Of course there are also seven other possible basic conjunctions, corresponding to negation signs in front of one or more of the three sentences. This gives eight possibilities in all, and we shall identify each of them with a possible 'world', or state of the world. Thus we

can draw up the following list for reference, noting respectively the basic conjunction and the name by which we shall label the associated world:

$h\&r\&w = w_1$

$h\&r\&\sim w = w_2$

$h\&\sim r\&w = w_3$

$h\&\sim r\&\sim w = w_4$

$\sim h\&r\&w = w_5$

$\sim h\&r\&\sim w = w_6$

$\sim h\&\sim r\&w = w_7$

$\sim h\&\sim r\&\sim w = w_8$

Begin with Tichy's original proposal (applied to this simple case). Consider the distance between one basic conjunction and another. The basic idea is that we define the distance between them as the sum of the basic states over which they disagree, normalised to be in the interval [0,1]. So for example the distance between w_1 (h&r&w) and w_2 (h&r&~w), which we denote $d(w_1, w_2)$, would be 1/3, since they agree on two out of the three basic states while disagreeing on one. Similarly $d(w_1, w_4) = 2/3$, $d(w_7, w_8) = 1/3$, $d(w_1, w_8) = 1$, and so on.

Note immediately two aspects of such a scheme that we return to later. First, note again the assumption that in any given language the exact truth is specifiable. This implies that any theory in that language can potentially be made exactly true, analogously to how a proposition (or theory – see next paragraph) in the toy example above can be made exactly true simply by rearranging negation signs. It will turn out that this assumption is dubious – see discussion later of stubbornly false theories. Second, we have already implicitly had to introduce a scheme of *weights*. In the above calculations, each basic state was assigned an equal weight – one-third – without any particular motivation. Later, we shall return to weights as a crucial way of incorporating necessary extra-logical information into our definitions.

Returning to our example, note next that a *theory* typically is consistent with many

different basic conjunctions, indeed can be expressed as a disjunction of them. Using an example of Brink's, consider the theory ('t_1' say) expressed by the proposition 'if it is hot then it is rainy and not windy', i.e. $h \Rightarrow (r \& \sim w)$. Now this single proposition is consistent logically not just with w_2 ($h \& r \& \sim w$) but also with all of w_5, w_6, w_7 and w_8 featuring $\sim h$. Thus the theory reduces by disjunctive normal form to the set $\{w_2, w_5, w_6, w_7, w_8\}$. Now each element of this set of course has its own distance from the true world w_1, which can be worked out according to our procedure above. What Oddie, the chief subsequent developer of Tichy's original approach, proposes here is that we take the arithmetical mean of the distance from w_1 of each of these elements. Thus in this case we would calculate the overall distance from the truth of our theory $d(t_1, w_1)$ as follows:

$$\begin{aligned} d(t_1, w_1) &= 1/5\ [d(w_2, w_1) + d(w_5, w_1) + d(w_6, w_1) + d(w_7, w_1) + d(w_8, w_1)] \\ &= 1/5\ [1/3 + 1/3 + 2/3 + 2/3 + 1] \\ &= 3/5 \end{aligned}$$

We can then if we wish define the theory's verisimilitude to be just $(1 - d(t_1, w_1))$, or in this case 2/5. (Note again that by taking the arithmetic mean we are implicitly assigning equal weights to each of the theory's disjunct conjunctions.)

Taking the arithmetical mean only works in the simple case where the truth is a single basic conjunction. Oddie's objective (in this context) is to define a general distance function between any two propositions. He proceeds to outline several more conventions by which he deals with all possible variations of the propositional case. First consider two propositions, 't_2' and 't_3' say, which (as sets) have the same number of elements. Again using Brink's example, suppose in our canonical weather case that t_2 is the proposition $\sim h \Rightarrow (r \vee w)$, and that t_3 is the proposition $h \Rightarrow (r \vee w)$. Then the elements of t_2 are all of w_1 to w_8 with the exception of w_8, and the elements of t_3 all of the basic states except w_4. So we might write:

$$t_2 = \{w_1, w_2, w_3, w_4, w_5, w_6, w_7\}$$

and $$t_3 = \{w_1, w_2, w_3, w_5, w_6, w_7, w_8\}$$

How should we compute a distance here between t_2 and t_3? *A priori*, there would seem to be more than one possible method. Perhaps using our previous procedure we should

compute individually the distance between w_1 and t_3, then the distance between w_2 and t_3, and so on for each element of t_2, before taking some grand average of all these individual distances in order to reach the overall distance. Oddie instead chooses a simpler computation: we should compute the distance between corresponding elements of each theory or proposition, and then take the average of this. Thus we would compare the two first elements, here w_1 and w_1, then the two second elements w_2 and w_2, and so on for all the elements, before dividing by the total size of each set. Thus:

$$d(t_2, t_3) = [d(w_1, w_1) + d(w_2, w_2) + d(w_3, w_3) + d(w_4, w_5) + d(w_5, w_6) + d(w_6, w_7) + d(w_7, w_8)] / 7$$
$$= (0 + 0 + 0 + 2/3 + 1/3 + 2/3 + 1/3) / 7$$
$$= 6/21$$

Of course, the problem with this technique is that we can achieve a different final result simply by changing the (arbitrary) order in which we write the sets' elements. Thus suppose we write t_2 as before, but now write t_3 in the following order:

$$t_2 = \{w_1, w_2, w_3, w_4, w_5, w_6, w_7\}$$
$$t_3 = \{w_1, w_2, w_3, w_8, w_5, w_6, w_7\}$$

Now our calculation would proceed as follows:

$$d(t_2, t_3) = [d(w_1, w_1) + d(w_2, w_2) + d(w_3, w_3) + d(w_4, w_8) + d(w_5, w_5) + d(w_6, w_6) + d(w_7, w_7)] / 7$$
$$= (0 + 0 + 0 + 1/3 + 0 + 0 + 0) / 7$$
$$= 1/21, \text{ and no longer } 6/21$$

So our measure of the distance between t_2 and t_3 has altered simply through an arbitrary reordering of how we write their elements. In order to avoid this, Oddie proposes that we always take the ordering that gives us the *lowest* value (in his terminology the 'narrowest linkage'). In our example this is the second ordering, which gave a distance measure of 1/21. In the simple case here at least, this is an intuitively appealing solution in so far as our intuition perceives that t_2 and t_3 'really' only differ by one element.

Again it is worth pausing, in order to note early instances of what are common criticisms

against all the purely logical approaches to approximate truth. First, consider Oddie's decision always to take the narrowest linkage, that is to say the ordering of set elements that yields the lowest value for the distance measure. The justification Oddie gives for this is an appeal to an *extra*-logical criterion, namely that some orderings (such as the first one considered above for $d(t_2, t_3)$) 'artificially' exaggerate the difference between the two theories and so should be rejected. Yet rather than smuggle in extra-logical factors under cover of a rather vague and perhaps *ad hoc* intuition like this, I think it is far preferable to try to incorporate them systematically. This is what we shall eventually try to do in our own definition (chapter 2).

Second, when computing $d(t_2, t_3)$, why did we compare each element only with the element in the same place in the other set's ordering, rather than compare each with *all* the elements of the other theory's set? This really boils down to the same point as before, since comparing with all the elements would in effect amount to taking as our official measure the average rather than minimum (across orderings) distance between the two theories. Again, the point is that on purely logical grounds the choice of procedure is arbitrary so we are forced to incorporate extra-logical criteria, in which case it is arguably better to do so more systematically.

Oddie's choice of procedure gives rise to further problems if the sets associated with two theories do not have the same number of elements, since then a simple one-to-one comparison of elements is impossible. These more complicated cases necessitate further stipulations. If the number of elements of one theory is an exact multiple of the number of the other, then Oddie proposes essentially that we multiply up the number of elements of the smaller set appropriately. For instance if some t_1 had six elements and some t_2 three elements, then for the purpose of computing the distance between them we could take each of t_2's three elements as occurring exactly twice. There still remains the awkward other case where the numbers of elements do not exactly divide each other. Here Oddie proposes a system of multisets, that is sets in which some elements appear more than once. (Of course these procedures are each themselves open to similar charges of arbitrariness as before.) Overall, by these means he is able to get for the propositional

case a complete definition of logical distance, that is to get $d(t_1, t_2)$ to be some number between 0 and 1 for all pairs of theories (t_1, t_2).

Niiniluoto's approach for the propositional case

Similar remarks could be made regarding any choice of procedure here, not just Oddie's. The point is not to criticise Oddie's particular choices so much as to note the kinds of decisions and assumptions that any such choices entail. But perhaps there are other aspects of this approach that raise further issues? We can get a sense of these by looking at the work of Niiniluoto. In contrast to Oddie, [Niiniluoto 1987] spends much time considering different possible distance measures. Nevertheless he ends up by selecting one very similar to Oddie's. In the context of our canonical weather example, Niiniluoto would define the distance $d(w_i, w_j)$ between two descriptions of the world as being the number of negation signs on which the two states differ, divided by the total number of basic states. So for instance on this measure

$$d(w_1, w_2) = 1/3$$

since w_1 and w_2 disagree on only one sign (whether we should have w or ~w) out of the three basic states they must sign in total.

Where Niiniluoto differs from Oddie is in his treatment of how to extend this basic distance function into a more general definition of verisimilitude. The issue arises when considering propositions, that is sets of possible worlds. Niiniluoto considers various possible distance measures, which can be illustrated as before by positing some theory t_1 and letting w_j be the (finite number of) elements of t_1. He concentrates particularly on the following three possiblemeasures:

1) $d(w_1, t_1)$ = the *arithmetical mean* of the distances $d(w_1, w_j)$ for all the w_j in t_1. (This is of course just Oddie's definition from the previous subsection.)

2) $d(w_1, t_1)$ = the *minimum* of the distances $d(w_1, w_j)$ for all the w_j in t_1

3) $d(w_1, t_1)$ = the *sum* of the distances $d(w_1, w_j)$ for all the w_j in t_1, divided by the sum of all the distances $d(w_1, w_j)$ for all possible w_j.

We can illustrate these with the example of t_1 we used earlier, namely where t_1 expresses the proposition $h \Rightarrow (r \& \sim w)$. This corresponds, recall, to the set $\{w_2, w_5, w_6, w_7, w_8\}$. What is the distance between t_1 and w_1 on each of the above three definitions?

1) Arithmetical mean:

$d(w_1, t_1) = [d(w_1, w_2) + d(w_1, w_5) + d(w_1, w_6) + d(w_1, w_7) + d(w_1, w_8)] / 5$

$= 1/5\ (1/3 + 1/3 + 2/3 + 2/3 + 1)$

$= 3/5$ as before

2) Minimum:

$d(w_1, t_1) = \min [d(w_1, w_j)$ for $j = 2,5,6,7,8]$

$= d(w_1, w_2)$

$= 1/3$

3) Sum:

$d(w_1, t_1) = [d(w_1, w_2) + d(w_1, w_5) + d(w_1, w_6) + d(w_1, w_7) + d(w_1, w_8)] / [d(w_1, w_1) + d(w_1, w_2) + d(w_1, w_3) + d(w_1, w_4) + d(w_1, w_5) + d(w_1, w_6) + d(w_1, w_7) + d(w_1, w_8)]$

$= (1/3 + 1/3 + 2/3 + 2/3 + 1) / (0 + 1/3 + 1/3 + 2/3 + 1/3 + 2/3 + 2/3 + 1)$

$= 3/4$

So here the three different definitions of verisimilitude yield three different answers.

Notice first that definition 2 in fact corresponds to what we might think of as accuracy or degree of truth, in that it captures how near to the true state of affairs a particular theory gets at its best, while ignoring its general scope. And notice second that definition 3 in turn corresponds to what we might think of as a theory's content or informativeness, since it captures to what degree a theory illuminates the issue at hand, as opposed to remaining just a truistic generality. Niiniluoto then picks up on Popper's original two desiderata for theories mentioned earlier, namely accuracy and content. Accordingly, he eventually concludes that the preferred measure should be some weighted average of definitions 2 and 3, since in his view this corresponds to a weighted average of accuracy and content. He does not specify what those weights should be precisely. But recall that in Popper's scheme neither accuracy nor content alone is sufficient, so we do know at least that both weights should be greater than zero.

It is clear in what sense Niiniluoto sees the logical-similarity approach as being a continuation of Popper's project of verisimilitude. It thus appears that this approach can be used to define the verisimilitudes of general theories rather than just those of singular cases. Of course, it could also be adapted to the more context-specific task by selecting just definition 2 and ignoring the considerations of content represented by definition 3.

Note though that some of the criticisms of arbitrariness levelled against Oddie also apply here. And indeed we shall see (next section) that most of the problems with the logical-similarity approach are common to both. Note also that the use of definition 2 to represent context-specific accuracy would require us to view models just as deductions, given certain initial conditions, from general theories. But such a view of scientific modelling now looks rather too simplistic [Morgan and Morrison 1999].

Beyond the propositional case

For any of these measures of verisimilitude ever to be applicable to actual scientific theories, it is necessary (all agree) that we be able to extend them beyond just the propositional case. In particular, we want to be able to extend them to first-order languages so as to be able to include relations and quantifiers. It turns out that Oddie and Niiniluoto each suggest very similar ways of achieving this.

In our canonical example above, a certain number of basic states (three in this case) generated a certain number (eight) of worlds, one of which we took to be the real world. When considering the verisimilitude of some proposition, the strategy was to write it in disjunctive normal form, each disjunct corresponding to a world. Thus a proposition could be expressed as a certain number of guesses as to which world is the real world. We could construct a distance function between a set of worlds and one particular world, and then defined a proposition's distance from the truth according to that function.

The trick now suggested by Oddie and Niiniluoto for the first-order case is to isolate our enquiries to a particular *depth*, 'd' say. Any first-order formula's depth is defined as the

number of bound variables used to express it. The key result is then Hintikka's Theorem, which states that for any particular depth d we can rewrite a proposition as a disjunction of depth-d constituents. Thus we extend from basic states to basic conjunctions as before, and this time also to existential conjunctions. Given Hintikka's Theorem, we are then in a position once again to characterise any proposition as being a set of guesses about which worlds might be the real world. As before, for any depth d we take it that there is one world that actually is the real world. Finally, we then again define a distance function between sets of worlds and a single world, and take this to be our measure of nearness to the truth. Since we can follow this procedure for any depth d, it provides a working definition of verisimilitude for the first-order case. Our definition can also now incorporate variables over infinite domains. Oddie even makes preliminary forays at trying to define verisimilitude for yet higher-order languages.

It is technically rather more complicated to define a distance function now, since the possible 'worlds' are sets of trees rather than the simple constituents of our original canonical case above. Oddie nevertheless sets up a measure essentially exactly analogous to his one for the propositional case. Niiniluoto considers various possible alternatives before again settling for a generalised version of his previous preferred combination of accuracy and informativeness, suitably relativised to depth.

But essentially the same criticisms as made earlier apply equally to these higher-order efforts, notwithstanding their great technical sophistication. In particular, extra-logical criteria are still necessarily appealed to no less than before. Indeed the extra technical difficulties lead if anything to a still greater number of arbitrary definitional decisions. Thus from a philosophical point of view, the higher-order work does nothing to rebut the earlier complaints about the definitions for the propositional case. If we are unhappy with the definitions for the simple propositional case then we shall still be unhappy, and for similar reasons, with the definitions for the higher-order cases too.

Recent extensions

Schurz and Weingartner show that one way to save Popper's original definition of verisimilitude from the standard Miller-Tichy refutation is to limit the content of a theory to its 'relevant' consequences [Schurz and Weingartner 1987]. Intuitively, the motivation is to restrict our attention to what a theory actually says about a problem at hand, and ignore what it happens to imply for other things we are not interested in. In this way one false theory can after all be shown to have more true, or fewer false, (relevant) consequences than another. But the exact definition of 'relevant' here is in effect left a primitive. More fundamentally, I think the way to focus on just the relevant consequences of a theory is to go further and to consider exclusively the accuracy of a context-specific model, foregoing any pretence at also assessing a theory's comprehensiveness. But then we would need to do more than just blithely classify what is relevant and irrelevant exogenously. Rather, we would need instead to be analysing much more carefully the different importance of the theory's different consequences. A large part of our own approach is motivated (chapter 2) by just this task. (Even if, Popper-style, we remain concerned with evaluating general theories rather than particular models, still Schurz and Weingartner's proposal would lead to problems. For if really assessing a theory for comprehensiveness as well as accuracy,, the justification for restricting attention to only a subset of 'relevant' consequences becomes obscure. If one were really concerned with comprehensiveness, then: 'it is not clear why one should be interested in such truncated theories' [Niiniluoto 1998, p8].)

There have been several more technical developments in recent years (for references, see [Niiniluoto 1998, pp8-10]). One idea, for instance, is viewing a statement as a member of the power set of the power set of the set of atomic propositions. Then a statement A is closer to the truth than another statement B if and only if each of A's members has some member of B as its subset, and each member of B is a subset of some member of A. This definition in turn can be extended and reformulated in various ways. Another alternative is to construct a definition in terms of set-theoretic relations between the classes of models of theories rather than the theories themselves. This again can be extended and reformulated in various ways.

The reason none of these developments has caught on as much as the more mainstream similarity and possible worlds (see section 1-5) approaches, is that so far the definitions they have ended up proposing have been plagued by one or more undesirable features. For example, they might imply that a weaker false theory should always be classed as more truthlike than a stronger false one – but as Niiniluoto points out, as a general principle this seems implausible. For instance, if the true number of planets is nine, it seems strange to class the weaker answer '10 or 200' as *more* truthlike than a simple '10'. Or perhaps they might imply that the truthlikeness of a trivial tautology is superior to that of a theory that gets almost but not quite everything right – but it is familiar now that any definition concerned with comprehensiveness should be capable of preferring the latter theory. A slightly different line of attack has been more internal to Niiniluoto's own preferred approach. It concentrates on technical questions such as various ways of incorporating continuous variables.

As with the extensions beyond the propositional case, the key point here is that none of these developments escape the more general critiques of the logical-similarity approach (to come in section 1-3). The judgment criteria driving the literature have, at least ostensibly, been purely internal. Definitions have not been evaluated explicitly by how well they incorporate extra-logical factors, for example. These latter are only ever reflected at best indirectly, perhaps through our intuitions regarding the force of each internal criterion. The agenda of new research has in effect been the satisfaction of those intuitions for more and more complicated logical formulations. But I argue that our main interest should be, recall, not so much the exact definition of similarity we use so much as what actually influences degree of similarity in the real world. In a sense then our real concerns are somewhat tangential to the main thrust of the logical literature here, and for that reason the new developments in the latter do not by our lights really lead to any progress. Rather, progress has been by the literature's own lights only. The real dispute concerns, as it were, whose lights we should prefer. Once again, if we are unhappy with the definitions for the simple propositional case then we shall still be unhappy, and for similar reasons, just as much now as 10 or 20 years ago.

1-3) Weaknesses of the logical-similarity approach

Language-dependence: Miller's objection

At the end of his original 1974 rejection of Popper's definition, David Miller raised what has since become (in Brink's phrase) the 'canonical objection' to the whole verisimilitude enterprise [Miller 1974] [Brink 1989, p186]. This was supplemented a year later by a further paper making an analogous point for the quantitative rather than qualitative case [Miller 1975]. The shadow of Miller's objection has loomed large. Indeed sometimes it has been taken (including on occasion even by Popper himself) on its own to demonstrate the entire literature's lack of success. We illustrate it here with reference to the qualitative case, as outlined in [Miller 1974].

Miller raised his objection in response to an initial suggestion by Tichy for a post-Popperian definition of verisimilitude. Recall the canonical weather example (originally formulated by Tichy), in which we imagined a language containing only three primitive sentences – 'it is hot' (or 'h'), 'it is raining' (r), and 'it is windy' (w). Assume that the true state of the world is that it is all three of hot, raining and windy, i.e. h&r&w. Now we add a further component – Tichy imagined two prisoners, Smith and Jones, each conjecturing what the weather is like outside. Jones thinks that it is cold, dry and still outside, or in other words ~h&~r&~w. Jones therefore gets all three variables wrong. Smith on the other hand thinks that the weather is cold, raining and windy, i.e. ~h&r&w. He therefore gets only one of the three variables wrong (cold instead of hot), and the other two correct. So Jones makes three mistakes but Smith only one. In the words of Tichy, 'it seems hardly deniable that Smith is by far nearer to the truth than Jones' [Tichy 1974, p159]. And accordingly, perhaps we can reach a definition of verisimilitude simply by counting the number of each theory's mistakes in this way. Miller's objection shows why this proposal, although at first sight attractive, in fact suffers from a fatal

weakness. The weakness also applies just as much to the subsequent developments by Oddie, Niiniluoto and others that we have just reviewed.

Consider the following reformulation of the weather example due to Miller. Suppose we invent another simple weather-language with just three primitive sentences. One of these is 'it is hot' as before, but in place of rainy and windy we now introduce two new ones: 'it is Minnesotan' (m) and 'it is Arizonan' (a). Now the key is that we can make this new language fully intertranslatable with our original one. Define the weather to be Minnesotan if it is either hot and rainy, or cold and dry – that is, m = (either h&r or ~h&~r). And define Arizonan to be either hot and windy, or cold and still – that is, a = (either h&w or ~h&~w). The eight possible states of the world described in the old language now correspond one-to-one with the eight possible states described in the new one. Listing respectively the world, the old-language description and the new-language description:

w_1 =	h&r&w	h&m&a
w_2 =	h&r&~w	h&m&~w
w_3 =	h&~r&w	h&~m&a
w_4 =	h&~r&~w	h&~m&~a
w_5 =	~h&r&w	~h&~m&~a
w_6 =	~h&r&~w	~h&~m&a
w_7 =	~h&~r&w	~h&m&~a
w_8 =	~h&~r&~w	~h&m&a

The two languages are thus logically equivalent. That is, logically speaking we could equally have designated the *h-m-a* language the 'original' one, and then defined a 'new' h-r-w equivalent using the transformations r = (either h&m or ~h&~m) and w = (either h&a or ~h&~a). There is complete symmetry. Accordingly, which language we happen to use is completely arbitrary, at least from a logical point of view.

The sting in the tail is the following observation. Recall that Jones's conjecture (~h&~r&~w) we judged to be clearly inferior to Smith's (~h&r&w), since it got three

rather than one variable wrong. But suppose now that Jones and Smith speak the h-m-a language, not h-r-a. If we translate their conjectures, we see that Jones's is ~h&m&a, while Smith's is ~h&~m&~a. But the truth in the new language is just h&m&a. So now it seems that it is Jones who has only got one variable wrong, and Smith who has done worse by getting all three wrong. In other words, simply by translating them into this new language, our verisimilitude ordering of the two theories has *reversed.*

Now if the languages had been logically different, then it is possible that their representation of the whole truth would have been different too, and thus that the translated theories may no longer have been equivalent to the originals. Alternatively put, the 'target' (that is, the expression of the truth in that language) would have changed and hence so might have judgments of truthlikeness. In these circumstances, arguably a change of verisimilitude ranking might still have been disturbing but at least potentially explicable. But since our two languages in this example are completely logically equivalent so the choice between them is arbitrary, and so it is difficult to see how any change of ranking can possibly be explained away. In Miller's words: 'just as truth is language-independent (this is one of the things that Tarski's T-schema insists on), so must judgments of verisimilitude be, if they are to have any objective significance at all' [Miller 1974, p176].

We return to Miller in section 1-4. I believe that his objection is less fatal than it at first appears but that in order to deal with it extra-logical factors must be addressed explicitly. Our own account will later do just this, but one of the main weaknesses of the logical-similarity approach is precisely that it does not. Accordingly the Miller objection still remains a powerful one against it, but the underlying cause of that is in turn *other* weaknesses of the logical-similarity approach. So we turn to those now.

Pure logic alone is not enough

Begin by going back to before the verisimilitude literature even really began, to Nelson Goodman. He presents a general analysis of the notion of similarity, the chief conclusion

of which is that 'anything is in some way like anything else' [Goodman 1972, p440]. More precisely, just assuming 'similarity' to be a well-defined relation is unjustified. Rather, to make sense of the notion we have to specify *in what respects* two things might be similar. Many intuitive judgments beg this question and are hard to pin down formally. We need some specific weighting for *importance* over all the properties of a situation in order to give meaning to the similarity relation. Goodman then continues: 'comparative judgments of similarity often require not merely selection of relevant properties but a weighting of their relative importance, and variation in relevance and importance can be rapid and enormous'. But: 'importance is a highly volatile matter, varying with every shift of context and interest', and he supports this thesis with several examples. The conclusion is that 'similarity is relative, variable, culture-dependent' [Goodman 1972, respectively pp445, 444 and 438].

I endorse Goodman's central points that similarity is ill defined without some notion of weighting, and that exactly which weighting we want is highly context-specific and so needs to be adjusted with every application. More interestingly, it seems there is no one in the field who actually disagrees. Niiniluoto himself concedes that some system of weights and a certain (logical) arbitrariness regarding choice of properties is unavoidable, and indeed quotes Goodman approvingly in favour of these arguments. He concludes: 'this means that degrees of similarity and comparisons of similarity are relative to two *pragmatic* boundary conditions: the choice of the relevant characteristics ... and the choice of the weights for the importance of these characteristics' [Niiniluoto 1987, p38, my emphasis]. More explicitly, 'there are not purely logical grounds for these choices' [Niiniluoto 1987, p38]. Oddie also notes that weights can serve to mark the different importance of different aspects of a situation [Oddie 1986, p56]. He too quotes Goodman, largely in the context of an effort to sidestep Miller's objection, and states: '... [it] is surely right that it is a problem for which pure logic cannot legislate the answer' [Oddie 1986, p184]. Miller himself writes: 'what must be important in the assessment of false hypotheses is not simply the amount of error they commit but also the *seriousness* of the errors committed' [Miller 1994, p200, my emphasis]. And with regard to verisimilitude: 'I do not pretend that there is any entirely logical solution to this problem.

In fact, I think that we have gone as far as pure logic will take us' [Miller 1994, p207].

So there in fact exists a measure of consensus on the principle that more than logic alone is needed. Nevertheless there remains a strong difference in practice about how to respond to this state of affairs. In the scheme to be developed in this thesis, the focus is on context-specificity. In the context of logical distance measures such as Niiniluoto's, this can be understood as implying a constant shifting of the weights on the different logical constituents. But as argued at length in this thesis, I think the philosophical interest lies more in what determines these shifting patterns of weights, and less in the surface question of the particular definition of distance. As it were, the choice of measure is only a secondary issue. What is important is not that a model captures a large number of logically correct statements but rather that it captures adequately the handful of important factors in an actual physical situation, even if these equate to only a few logical statements. The weighting, so to speak, is everything; the simple number of logical statements, in itself nothing. Accordingly, we shall focus mainly on issues surrounding what it means to capture the important aspects of a situation, and not on defining a distance measure (although see the appendix at the end of chapter 2 for this latter issue).

By contrast, the main thrust of the similarity literature has been on refining the definition, i.e. on the syntactics. As noted, all concede that extra-logical factors have inevitably to be considered, but the focus nonetheless has remained on logical issues and not on the extra-logical ones. A good example is provided by Niiniluoto himself. In recognising that we need a measure of the seriousness of errors as well as their number, he suggests that we might do this simply by adding a coefficient to each term [Niiniluoto 1987, p314]. (I agree so far, indeed I will use a similar tactic in my own proposed scheme later.) [Niiniluoto 1987, p315] highlights as the kind of thing he has in mind an article of his from ten years before, namely [Niiniluoto 1978]. But in that earlier paper he reveals himself still implicitly to be wedded to a purely logical approach. He specifies two types of error, 'serious' and 'non-serious', a distinction which effectively does introduce a non-uniform weighting system into his basic distance function. Yet his definition of these

two types of error still turns out to be determined by *a priori* criteria [Niiniluoto 1978, p447]. There is no consideration of pragmatic contextual factors. Accordingly, he is in fact still making no attempt to model the influence of extra-logical concerns explicitly. (Of course, in various guises many authors have noted or implied this defect in the similarity approach, for instance [Adams 1990] [Aronson, Harre and Way 1994].)

Oddie meanwhile introduces weights without a passing philosophical glance, almost purely as a technical normalising device [Oddie 1986, p45]. There is no suggestion anywhere of non-uniform weightings, or of imputing to the weights any physical significance. Rather he emphasises instead their technical usefulness, for instance when handling infinite sets [Oddie 1986, p58]. Fundamentally the focus remains purely logical, despite the acknowledged relevance of extra-logical factors. (In fairness to Oddie, he is putting, as it were, all his extra-logical eggs into the one basket of intuition. More precisely, he judges the satisfactoriness of different definitions purely on how well they perform intuitively – according to his intuition – in several stock examples of his. I think we can do better – see chapter 2.)

Summing up, we might say that a logical-similarity approach specifies only a logical formula. This is good in so far as it leaves us free to choose a target and weights as we please, but the drawback is that often it is precisely these choices that are the philosophically interesting bit. It also seems that interest-relativity requires us in addition to take into account pragmatics (an issue discussed fully in section 2-9). So overall, the logical-similarity approach is left at best incomplete.

Finally, a word of defence. The critique in this subsection has been that extra-logical criteria have been left insufficiently examined. A different criticism, and different also from Miller's problem, is that *any* definition of similarity or of logical distance is necessarily arbitrary, on the grounds that the need to invoke extra-logical criteria automatically renders a definition hopelessly shaky or epistemic. But I agree with Niiniluoto when he notes that, once *given* a particular pragmatic weighting, then we can define a verisimilitude ordering wholly logically (see also section 2-10). More to the

point, this latter criticism seems to yearn for an untenable world where pure logic alone can somehow judge approximate truth indisputably.

Language and reality

As [Weston 1992] points out, the notion that we can capture an interesting sense of comprehensiveness by a purely logical definition is also disputable. This is because that definition concerns the logical distance between a theory and the truth, and needs to presume that the whole truth *can* be expressed in the particular language being used. In Weston's words, we need to assume that the vocabulary is known 'to be complete in the sense that no further additions will help give a fuller description of the subject matter' [Weston 1992, p71]. But historically this has rarely been the case, and accordingly it is doubtful whether such an assumption is reasonable. Niiniluoto himself concedes that therefore his truthlikeness 'is not a measure of ... distance from the "whole truth", but from a chosen target' [Niiniluoto 1987, p449]. And again more recently: 'my [measure of truthlikeness] is not intended to be a measure of the descriptive completeness of a conceptual framework, but rather ... is applicable only when the maximum of comprehensiveness is first fixed by the choice of the language' [Niiniluoto 1998, p15]. But then why be interested in this measure at all? This language-relativity calls into question whether a purely logical definition can even assess a theory's comprehensiveness in the first place. [McMullin 1987] too questions whether the truth as conceptualised in a particular language is an acceptable proxy for the 'whole truth'.

These critiques are of the comprehensiveness half of the equation, as it were, rather than of the attempt also to capture degree of accuracy. But an analogous objection can be raised there too. We return to this issue in detail in chapter 2.

McMullin also criticises the very idea of a 'distance' between two propositions, that is the idea that the definitions of distance offered in the verisimilitude literature correspond to anything philosophically interesting. Brink puts the point more gently by noting first that the literature takes verisimilitude to be the similarity between possible worlds and the real

world, and secondly that this similarity is then taken to go hand in hand with a concept of linguistic distance. The second assumption is the one at issue here – why should we take it that the abstract distance relations refer to any interesting aspect of the real world?

Distance between states of the real world is at the heart of the pre-philosophical notion of approximate truth in the first place. The issue here is not how to measure distance. Rather, it is the different question of whether or not we can use distance between logical propositions as a proxy for that distance between real states of the world. Brink calls the assumption that we can, the "linguistic assumption": 'that we can access possible worlds through their descriptions: the syntactic structure of a theory should in some way reflect the common structure of the worlds in which it is true' [Brink 1989, p188]. Despite its importance this assumption has only been weakly, if at all, motivated in the literature.

Our own strategy by contrast will be to define an *ontological* distance measure. This avoids the problem, since by definition it measures distance between states of the world understood *realistically*.

Stubbornly false theories

Often with a false theory, it is apparently promising to examine whether small modifications will make it true. The idea is that a theory is 'approximately true' when it is just a few such refinements away from being strictly true. For example, in a Newtonian universe a simple model of the Earth's orbit of the Sun is approximately true, but in principle could be made strictly true once allowance was made for the small perturbatory (but still Newtonian) effects of the moon, other planets and so on. Unfortunately though, many useful theories in fact appear to be stubbornly false, in that there is no obvious means of correcting them in this way. An example is classical fluid mechanics, which models fluids as perfect continua. Yet this cannot be strictly true of the real world of atomically granular fluids, and since the assumption of continua is a basic axiom of the theory it is hard to see how any refinement could remedy this. Nevertheless it certainly remains a useful theory, for instance governing the design of aircraft wings and

delivering many approximately true predictions. Therefore I think one of our philosophical desiderata here must be the capacity to reflect the worth of such theories.

But as Smith notes, the logical-similarity approach appears clearly to lack such a capacity [Smith 1998]. As we saw, it defines the verisimilitude of a proposition to be some function of the distances between each of that proposition's disjoined basic conjunctions and the one true basic conjunction. It follows that the proposition can be brought nearer to the truth by adjusting the pattern of negation signs (or lack of them) it assigns to each of the postulated basic states. In particular, in this scheme it should always be possible to make any proposition strictly true simply by achieving the correct pattern of negation signs, so that it replicates exactly the one true basic conjunction. But, in Smith's phrase, it is difficult to see how (for instance) classical fluid mechanics could ever be repaired merely by 'the simple expedient of twiddling a few negation signs in the basics in some canonical formulation' [Smith 1998, p265].

A similar situation arises in economics. Typically a modern economic model is built on a foundation of rationally maximising human agents, yet no such humans exist in the real world. It is not easy to see how these ideal agents could ever easily be fully converted into psychologically realistic figures either. Twiddling a few negation signs in the axioms of the theory certainly will not do the trick. Yet notwithstanding this stubborn falsity, economic models often do make accurate predictions nevertheless. Similar remarks apply to the rest of social science, and arguably much of natural science too. Moreover although stubbornly false, there is nevertheless much interest in examining how all these theories vary in their usefulness and applicability. As argued back in section 1-1, one of our most important desiderata for a theory of approximate truth is that it can help us make sense of these patterns of varying usefulness of stubbornly false theories. But if we cannot make sense of their ever being useful in the first place, then presumably we also cannot make any sense of variations in that usefulness.

This whole issue can perhaps be thought of as another example of the language difficulties of the previous subsection. The classical theory of fluid mechanics seems not

so much to be talking the right language and getting some details wrong, so much as to be talking (we presume) a completely wrong language altogether. We cannot be sure that the language of a theory is sufficiently rich to express the truth fully. If it is not, then a logical distance measure alone is insufficient to judge that theory's approximate truth. This would appear to be the case with the classical theory of fluid dynamics, and indeed with all other stubbornly false theories. Neither Oddie nor Niiniluoto provides an example of a real-world application of their schemes that really makes clear how to meet this objection. (See [Aronson, Harre and Way 1994] for similar criticism.)

Finally, two further notes. First, arguably the empirical success of stubbornly false theories could be seen as problem for realism generally, not just for verisimilitude theorists. Second, the way our own scheme handles the issue is outlined in section 2-12.

Two measures in one

Is it possible to capture in a single logical-similarity measure both a theory's accuracy and its range? Oddie's and Niiniluoto's represent the best attempts to do so. As previously mentioned, I prefer instead a context-specific focus just on a particular model's accuracy. Niiniluoto is certainly aware that the truthlikeness of a grand theory needs sometimes to be broken down to become intelligible. For instance, he states: 'If we are asked what degree of truthlikeness a scientific theory like Newton's mechanics has, we should be more specific: relative to which application (solar system, pendulum, free fall, and so on)?' [Niiniluoto 1998, p13]. Moreover, as we saw, one of his sample measures explicitly captures just accuracy, not informativeness. But nevertheless, he still believes that 'for some comparative purposes it may be useful to assess the global truthlikeness of a theory' [Niiniluoto 1998, p13].

Perhaps a root of the logical-similarity approach's difficulties with incorporating extra-logical factors is that it is hard to see how to do this while still talking about general theories. Our own scheme will get round the problem by understanding causal descriptions realistically, and noting that context-specific models can be interpreted as

amounting to such descriptions and hence can be analysed for approximate truth. But abstract general theories *per se* cannot – unless they are concretised in the form of a particular model, in which case of course we are again dealing with a specific context. Thus we cannot use our own eventual method to make any sense of a theory's 'global' truthlikeness. Rather, we need always to specify first exactly what concrete situation in the world a model is trying to capture – so not just 'pendulum', but what aspect of which specific actual pendulum? Not just 'free fall', but what aspect of which specific actual falling object?

It might be wondered too why in any case we should even expect it to be possible to force the two separate pegs, as it were, of accuracy and comprehensiveness into the single hole of just one measure of truthlikeness. As Adams points out: 'intuitively, comprehensiveness seems to be a quality that is desirable in general theories, while accuracy is what is wanted of the particular statements that make them up, and of the data that support them' [Adams 1990, p147]. Moreover, he imagines the possibility of separate theories of accuracy and comprehensiveness in turn illuminating the possible interrelations between the two concepts.

In the context of social science, it is clear that we are not concerned with how close to the truth a general theory is, only with how accurate a specific model is. This thesis originally grew out of trying to make sense of economic theory's variable pattern of usefulness. This different initial motivation is, I think, useful here: nobody thinks economic theories are comprehensively true anyway. Rather, what is of interest is how their degree of accuracy varies across different contexts.

Summary

1) The most commonly cited objection to the logical-similarity approach, and indeed to all other approaches to verisimilitude so far, is Miller's. This complains that all proposed measures generate rankings that are inconstant with respect to arbitrary switches between logically equivalent languages. (We return to Miller in the next section.)

2) I think the most important difficulty is that: a purely logical measure of approximate truth is on its own insufficient since any definition of similarity must take into account the relative seriousness of different errors, and this in turn can only be assessed with the help of extra-logical criteria. This point is not really controversial. Nevertheless, in practice the literature ignores the obvious inference that we should therefore be moving our philosophical attentions away from purely syntactic refinements and over instead onto precisely these extra-logical issues. As it were, when reading from a map anyway distorted by miles it is pointless devoting energy to perfecting measurements to within inches. The first priority is to be able to see the wood for the trees.

3) All the logical distance measures represent truth only relativised to a particular language. Accordingly, it becomes questionable whether distance from such a 'truth' really captures anything philosophically interesting.

4) Stubbornly false theories that are nonetheless useful, are also difficult for the standard approach to handle. Again this difficulty can be seen as stemming from trying to characterise truth in a linguistically relativised way. (Points 3 and 4 can in fact arguably each be seen ultimately as just further examples of Point 2.)

5) Further problems stem from an inherited, but questionable, Popperian attachment to viewing science as a sequence of inductive generalisations. As well as causing difficulties, trying to capture both accuracy and comprehensiveness in a single measure also seems just ill motivated.

(Note also that logical distance measures arguably also struggle with the problem of legisimilitude [Liu 1999], itself perhaps another symptom of Point 2. But we save discussion of that issue until section 1-7.)

1-4) Miller's objection revisited

Relevance to our own eventual scheme

Our own scheme (chapter 2) shall lay heavy emphasis on *ontology*, and in particular on stating models in *causal* terms. But the heart of the Miller problem is really the issue of how to justify privileging some languages or parameters over others. Getting the right ontology is not enough here; what we need in addition is in some sense to get the right vocabulary, if such a thing exists. The two tasks are not the same, since the same ontology can support more than one vocabulary. Therefore even if we believe that certain properties and not others are natural kinds, still that does not uniquely specify the right vocabulary. Natural kinds apply to properties, not to the predicates we use to describe them. For example, predicates such as Miller's 'minnesotan' and 'arizonan' truly apply to some individuals and fail to apply to others. They are not like the predicate 'is a phlogiston', which applies to nothing. (The only way out would be some kind of strong Platonic realism with respect to ontology, such that 'minnesotan' and 'arizonan' do not pick out 'real properties', whereas 'rainy' and 'windy' do.)

In particular, there will in general be many different ways of stating a true *causal* ontology as well. For instance, the conjunction of a cause and an irrelevant factor will itself also typically be a true cause too. So will the conjunction of one cause and another cause. So even, for some instantiations, will be the conjunction of a cause and a counteracting cause. Unfortunately for our own scheme, there seems to be no such thing as a canonical causal vocabulary. In particular, there seems to be no way to rule out vocabulary that leads to Miller-reversals – or at least there is no way to do this purely on metaphysical grounds. This suggests that we must go beyond pure metaphysics for the solution, which is indeed what we shall now do.

Mormann and conventionalism

I think the correct response to [Miller 1975]'s quantitative examples is presented best in [Mormann 1988]. He points out (as have others, for instance Niiniluoto) that Miller's examples are all in fact instances of a single underlying mathematical fact – namely that homeomorphisms or continuous bijective mappings between topological spaces need not preserve metric structure. Definitions of verisimilitude based on distance functions are of course dependent on a particular metric, so an unpreserved metric structure will correspond to varying values for verisimilitude, as per Miller's problem. So it is certainly mathematically possible to generate inconstant verisimilitude orderings. The key question is: are such transformations, just because mathematically possible, thereby also philosophically significant?

Mormann argues that to accept this latter view is in effect to adopt a species of *geometric conventionalism*. In particular, it is to claim that it is entirely a matter of convention which metrical structure we should adopt when measuring two theories' experimental outcomes – just as much as would be the choice of measuring length using metres or yards. More precisely put, physical space is metrically amorphous and may be metricised in many different ways, and Miller's thesis is just the generalisation of this that claims that 'higher dimensional physical magnitude spaces are metrically amorphous as well' [Mormann 1988, p513].

But, following Quine and Putnam, Mormann argues that such a conventionalism is untenable:

> The meaning of a term (in our case a physical metric or distance function) is *not* exhausted by a short list of axioms ... but is rather a function of an extended net of empirical knowledge. That is to say we do *not* fix the reference of the term 'metric of physical space' by convention but by *coherence*. The fixation by coherence involves large parts of scientific background knowledge and proceeds in a series of approximations. A first step for the fixation of a physically meaningful metric of physical space is to impose the condition that a measuring rod is to stay the same length when transported. Reichenbach erroneously thought that this condition would be sufficient to determine the metric of physical space uniquely but at least this condition

> excludes our contrived metric [taken from a Miller example]. Further steps of the approximation process may take into account constraints concerning the form of physical theories, e.g. invariance principles. It seems that in the case of physical space and its metrical structure this narrowing down process is quite successful: for middle-sized objects and distances there is no other candidate left than the traditional Euclidean metric of physical space. The more general case of arbitrary magnitude spaces has not been dealt with in greater depth. But there does not seem to exist a principle obstacle which would prevent us from adopting a similar approximation procedure in this generalised case too.
> [Mormann 1988, p514]

and:

> ... many properties are part of a physical magnitude and ... the meaning of such a term is not exhausted by a simple formal definition. The relevant structure of physical magnitude spaces is much richer and it depends in such complicated ways on other empirical theories and conceptions that it cannot bear just any *prima facie* possible formal manipulation as Miller asserts
> [Mormann 1988, p517]

In other words, not just any metric will do; rather they must also be physically meaningful. When assessing verisimilitude, we must take into account not only our calculation's formal procedures but also the scientific context underlying it. Logic alone is not enough.

Tellingly, Mormann points out that Miller is later forced to reject his own conventionalism himself. Arguing against Good, Miller states: 'reversals of ordering by accuracy can indeed be obtained even in the one-dimensional case if we are prepared to allow discontinuous transformations ... But this cannot be thought to be anything like as interesting, since some topological restraints must be insisted on if our reformulated hypotheses are to be reformulations at all' [Miller 1994, p226]. In other words we must bear in mind at least some extra-formal criteria. But as Mormann asks: 'why is the

metrical structure [of a magnitude space] conventional while its topological structure is not?' [Mormann 1988, p518] And as he argues, there can be no general answer, only consideration in each individual case of what the scientific context tells us should and should not be taken as conventional.

It is worth emphasising the one-dimensional case (regarding accuracy of predictions [Miller 1975]). Can we even say that, for example, 5 (units of some physical quantity) is nearer to 6 than to 7? If he is not anti-conventionalist here, then Miller is hoist on his own petard. His own demonstrations of Miller-reversals themselves *rely* on us being able to make assessments of approximate truth unambiguously for such one-dimensional cases, since the very notion of a reversal implies the existence of some well-established ordering that can *be* reversed. But then this requires that we be able to judge one model's numerical answer more accurate than another's and, as Mormann asks, what justifies being anti-conventionalist only in these cases and not in others?

Of course, anti-conventionalism need not imply that, in Mormann's words, 'a certain set of (traditional) magnitudes is beyond any doubt' [Mormann 1988, p516]. If there exist alternative empirical parameters that are also physically meaningful, then this would present genuine competition as to which parameters to choose. Mormann presents a possible such case for the example of a falling body [Mormann 1988, p516]. One pair of parameters is the distance of the body from the ground, and the square of its momentum. This pair is interdefinable with another pair of parameters that turn out to correspond to the Hamiltonian and Lagrangian of a body in a gravitational field. Therefore both pairs of parameters refer to physically meaningful magnitudes here. Now it turns out that the new pair of parameters give a basis for a metric different from that of the old pair, and accordingly it is entirely possible that the verisimilitude ordering of two false theories may change depending on which metric we choose, just as Miller demonstrates. The difference now would be that this new pair of parameters is not just arbitrary, but rather is physically meaningful and therefore potentially a serious candidate for scientists' attention. As it were, the bottom line is that it is only this latter kind of ambiguity that is philosophically interesting.

Miller himself appears tacitly to accept Mormann's analysis. At least, in his discussion in [Miller 1994] he scarcely rebuts a single one of Mormann's points. Indeed, rather the contrary: 'from a formal point of view false hypotheses that predict values for the same set of quantities are on a par. ... If one such false hypothesis is preferred to another, then presumably one family of quantities is preferred to another. In my 1975, 181-86, I made faces at such preferences, accusing them of essentialism, anyway in the general case. That was no doubt an excessive reaction, as [Mormann 1988] justly observes' [Miller 1994, p231]. In other words we should not accept Miller's own implicit geometric conventionalism.

Where does this leave us?

Mormann provides the only precise diagnosis for [Miller 1975]'s quantitative example. In a similar spirit, regarding [Miller 1974]'s qualitative propositional formalism of the weather example the general response has again been an appeal to pragmatics. Weston lays great stress on the pragmatic use of background knowledge and theoretical criteria for determining an appropriate 'sense of approximation'. For example, in classical mechanics we are justified in disregarding the centrifugal and Coriolis 'forces' which appear in rotating coordinate systems, and preferring instead inertial formulations of the laws, notwithstanding mathematical equivalence [Weston 1992, p68]. In other words, we need to go 'beyond just the mathematics'. In line with this, he argues against Miller essentially by saying that therefore we are justified in privileging some quantities for measuring accuracy, namely the ones that are actually (in a realist sense) causally significant. Niiniluoto concludes similarly: 'real-life applications of the concept of truthlikeness to scientific hypotheses and theories should be made relative to those conceptual frameworks ... that are actually used by scientists' [Niiniluoto 1998, p17]. And Oddie agrees that we 'must grant certain properties, magnitudes or constants a privileged status', and that 'it may be that [extra-logical] considerations will set some properties apart from others' [Oddie 1986, p159].

One way or another, all – Weston, Oddie, Niiniluoto, Mormann – agree that in order to answer Miller's objection, it is necessary to introduce extra-logical considerations. Indeed, from the start so did Miller himself: 'I do not know of any *logical* way of distinguishing the fundamental constants (or parameters) from the remainder' [Miller 1975, p185, my emphasis]. This of course recalls our biggest criticism of the logical-similarity approach (section 1-3), namely precisely its neglect of just such extra-logical factors.

More generally, a concentration on actual scientific practice suggests that we are worrying about the wrong things here. We started with the motivation from economics that a key issue is the applicability of models to particular real-world situations. There is no interest in economics in whether theories as a whole are progressing nearer the truth since it is universally accepted that they deal in fictional simplifications; rather, the methodological interest is in context-specific degree of applicability. We shall see in the next chapter (section 2-6) that similar remarks may apply to the great majority of work in natural science too. Moreover, what is typically *not* in dispute are the choice of ontology and vocabulary. Therefore Miller-reversals are largely irrelevant to the actual practice of science. In a nutshell, the Miller problem is (overwhelmingly) no problem *methodologically*. Therefore any definition of approximate truth seeking to be relevant methodologically should be concentrating not on Miller reversals, but rather on those factors that *are* important methodologically.

It is precisely those factors that will feature in our own definition in the next chapter. We shall emphasise there getting correct causal strengths, where these in turn are defined in terms of certain empirical outputs. Hence there will still be, so to speak, a rigorous connection between our scores for approximate truth and the external world. The only thing missing will be objective metaphysical validation for concentrating on some empirical outputs rather than others. But the point is that from a methodological point of view no such validation is required anyway – we just are interested in whatever parameters we happen to be interested in and this *needs* no justification. Approximate truth relativised to those parameters is what matters; approximate truth in some absolute

metaphysical sense is irrelevant.

Finally, we may note a connection here with more general naturalistic views of science. The claim is that philosophy should focus not on a priori problems, but rather only on those problems that are actually important in the practice of science (see e.g. [Maddy 2000]). On this view Cartesian fundamental scepticism, for instance, sets the epistemological bar too high. In practice, epistemological controversies in science are, so to speak, conducted at a lower level of certainty. In actual science, whether or not some fact is considered sufficiently well-established does not turn on whether it passes Descartes's fundamental sceptical test; rather, the tests that prove useful are rather less strenuous, for instance significant empirical support, coherence with existing background knowledge, plausible causal mechanism, and so on. And this applies even though, logically speaking, Descartes's fundamental sceptical challenge arguably has not been satisfactorily resolved. Analogously, in the actual practice of science some notion of approximate truth is indeed important. But that notion is, I claim, something akin to our own definition of it, that is to a definition that does not assume that choice of ontology and vocabulary is always in dispute. Just as science can progress satisfactorily without solving Descartes's problem, so it can progress satisfactorily without solving Miller's. And so just as useful methodological work does not focus on Descartes's problem, neither – when addressing approximate truth – need it focus on Miller's. The practice of science does not require so perfect a notion as Miller asks for. If we are interested in, as it were, some absolute metaphysical conception of approximate truth then its language-dependence is indeed a significant problem; but methodologically language-dependence is largely irrelevant.

1-5) Some other possible approaches

Possible worlds

The Kripke-Lewis notion of possible worlds, as opposed to the possible worlds of classical semantics, is familiar for instance from its role in the analysis of counterfactuals. It can also be put to use for the study of approximate truth. The starting point is the idea that a sentence is approximately true in our world if and only if there is another world sufficiently close to our world in which the sentence is exactly true. So a central element must clearly be some grasp on what we mean by one world being 'close' to another here. [Hilpinen 1976] develops for this purpose a proposal due originally to [Lewis 1973].

The basic notion is that we can define a set of *nested spheres* of classes of possible worlds, with the actual world being at the centre. A world is less similar to the actual world the further away the sphere on which it lies is from the centre. Any statement implies a set of possible worlds in which it would be true. Then, roughly speaking, the closeness of a statement to the truth is the distance of the spheres containing its possible worlds from the actual world at the centre. And one statement is nearer the truth than another if its associated spheres are nearer the centre. There are various technical issues over the exact definition of this 'distance', for instance whether we should take the distance from the nearest of a theory's spheres, or the furthest one, or some sort of average, and so on. But the philosophical strengths and weaknesses of the approach are more or less independent of these precise concerns, so we disregard them here.

The Lewis-Hilpinen project carries one immediate advantage, noted in [Smith 1998] – it enables us to make sense of the approximate truth of stubbornly false theories. Thus the classical theory of fluid dynamics, for example, cannot it seems be easily adjusted into a true theory. In other words, its linguistic *representation* cannot easily be adjusted into a true one. Nevertheless, it may still be that the theory is true of a (Lewis-type) possible *world* that is very similar to the actual one. But of course we now need to explicate more this new notion of similarity between whole worlds. And this is the possible worlds approach's fatal weakness – it explicitly just takes similarity to be a primitive.

Note first that leaving similarity a primitive leads to a potential equivalence between the

possible worlds and logical-similarity approaches. In particular, this will be so if the former's distance function between nested spheres is equivalent to the latter's distance function between worlds. Indeed [Niiniluoto 1987] remarks on the possible connection, although it remains disputed to what degree a unification of the two approaches could really be feasible. Nevertheless, it is perhaps no surprise that many of the points from section 1-3 against logical-similarity definitions also recur here.

Since all is effectively subsumed by the possible-worlds approach into the black box of the primitive 'similarity', it is clear that a lot of the burden of defining similarity between possible worlds is likely to fall on extra-logical considerations in order to assess the differential seriousness of errors. Thus the real driver of verisimilitude judgments will again be these extra-logical factors, yet leaving similarity a primitive leads instead to a concentration on the definitions' formal structures, recalling our similar objection to the logical-similarity literature. Next comes the problem of relativising the truth to a target in a particular language. This issue is again subsumed by the possible-worlds approach into the black box of 'similarity'. Until forced to judge two worlds' similarity, issues of representation can be kept at bay. What of theories that are approximately true but stubbornly false? As noted previously, the concentration on similarity between worlds rather than between linguistic representations in principle enables the possible-worlds approach to deal with such cases more easily. But although the possible-worlds scheme does provide a potential route out, again leaving similarity a primitive means in effect it tells us nothing about how actually to find that route.

Lewis himself adjudges that any judgment of comparative similarity of worlds must be a 'messy business' [Lewis 1986, p24]. In particular, he agrees with the basic point that it is necessary to weight some aspects more than others when assessing similarity. But he offers no real constructive proposal beyond these observations, except apparently to hint that our choice of weights should be almost intuitive in just the right way as to reflect our original pre-philosophical intuition of closeness to the truth [Lewis 1986, pp24-7]. But I think it is possible (and desirable) to say a little more than that (section 1-6 and chapter 2).

Finally, [Smith 1998] convincingly outlines one further decisive objection – redundancy. His point is that the entire metaphysical detour via possible worlds is in fact unnecessary, since effectively the same definition of approximate truth can equally well be obtained by an ontological approach without any need to invoke all the possible-worlds apparatus in the first place. To illustrate, take the classical theory of fluid dynamics again. This theory is only approximately true when we focus on macro-patterns of fluid flow; in other respects, it is not particularly close to the truth. Now in order to make the possible-worlds approach work here, we need the theory itself to inform our sense of similarity between worlds. So the possible-worlds account of approximate truth would say first, that classical fluid dynamics is approximately true if and only if the world in which it is strictly true is similar to the actual world, and second, that in turn these worlds will be similar only if we are giving weight to the particular phenomenon (i.e. macro fluid flows) that the theory is concerned to explain. But as Smith asks, in which case just when will the actual world be appropriately similar in this way to the model's posited world? The answer – precisely when the model's posited structure of the particular phenomenon of interest is similar to the actual structure of that phenomenon. Therefore once we have weighted our attention exclusively onto as it were just a particular subset of the world (macro fluid flow), we find that the world's background facts and nomological structure are no longer playing any role in our estimation of approximate truth. All that matters is the similarity between the model's posited structure of the phenomenon at hand and the actual structure, regardless of the rest of the world. But now this is precisely the context-specific definition of approximate truth we find in an ontological approach, a version of which we shall develop in chapter 2. In other words, both the possible-worlds and ontological definitions boil down (in this example) to the similarity between actual macro fluid flow and the classical picture of macro fluid flow, ignoring other aspects. So there is no need after all to invoke the grand metaphysical apparatus of possible worlds; a simple ontological definition instead gets us to the same place anyway.

Structurelikeness

Associated particularly with Sneed and Stegmueller, the structuralist program develops Suppes' original conception of scientific theories as abstract structures, which are then taken to correspond to particular phenomena in the real world. The relevance to us is that this offers another way of conceiving approximate truth. Roughly speaking, it conceives it as being the similarity between on the one hand the abstract structure representing the theory, and on the other the second structure that is actually instantiated in the real world. Of course, the task is then to specify more closely what kind of structures these are and what kind of similarity is being invoked. A characteristic of the approach is to reduce all such structures to those of set theory. Depending on the author, different notions of similarity are invoked, sometimes the Bourbaki concept of uniformity, sometimes a specific metric or pseudometric [Niiniluoto 1998]. As Niiniluoto notes, uniformities here 'have arbitrary features and are often insufficient for the comparison of false theories' [Niiniluoto 1998, p12]. In the case of specific metrics on the other hand, often the resultant definition of approximate truth has a formal parallel in the logical-statement approach. Indeed Niiniluoto's work, like that of Kuipers (see shortly), can be read in part as an attempted synthesis of the structuralist and verisimilitude traditions.

[Smith 1998] notes a problem with representing all theories in terms of sets. Suppose that a standard model, for example the ideal pendulum, is held to be strictly true of some system in the real world. This implies that a certain real-world set belongs to the set of ideal pendulums. The difficulty arises if we want to say that the model is only approximately true of some real-world pendulum, because then we would appear forced to say that the real-world set is only 'approximately' a member of the set of ideal pendulums. But this of course in effect passes the definitional buck onto the unexplicated notion of approximate set-membership. And being approximately a member of a set seems no less problematic than was the notion of being approximately true in the first place. Sneed himself frankly acknowledges the difficulty, adding that the solution 'is not clear to me' and 'I have no suggestions how such an account is to be provided' [Sneed 1971, p25].

Problems familiar from previous sections also recur here. Thus it is not always clear how

to understand the posited structures realistically. What realist ontology is thought to correspond to the abstract structures of set theory? An answer to this would help inform us about the extra-formal factors crucial to resolving the conundrum of approximation. As things stand, a concentration just on formal matters can tend instead to direct attention away from the important point. On the other hand, arguably structuralism does tend to concentrate more on context-specific as opposed to general approximate truth, since real-world instantiations of ideal structures are often considered singularly.

The most notable recent development of structuralism in the direction of approximate truth has been the work of Theo Kuipers. In the canonical example, the various combinations of hot, rainy and windy, or their negations, gave rise to eight different possibilities. So far we have taken just one of these possibilities to be the truth, what we can label here the 'descriptive truth'. But [Kuipers 1987] introduces a second concept, what he calls 'theoretical verisimilitude'. Thus suppose that only a subset of those eight possibilities is actually physically possible – for instance, imagine in the canonical example that for some reason it is physically impossible for it to rain when it is also either cold or still. In these circumstances, three of the standard possibilities (cold-rain-wind, cold-rain-still, hot-rain-still) are no longer physically admissible. The remaining five possibilities therefore summarise all the physically permissible states – Kuipers names this set the 'theoretical truth'. His descriptive verisimilitude is then just closeness to the descriptive truth, and theoretical verisimilitude closeness to this theoretical truth. (Kuipers's theoretical verisimilitude is closely akin to what later (section 1-7) we shall discuss as 'legisimilitude'.)

Broadly speaking, for descriptive verisimilitude Kuipers is happy to endorse Niiniluoto's metric. For theoretical verisimilitude, on the other hand, label the set of physically possible worlds W. Then a theory or proposition A can be taken as the assertion that the two sets A and W are equal, in which case any world in A would be in W, and *vice versa*. Kuipers's first suggested definition of distance from the truth here was then essentially the symmetric difference of A and W, i.e. $(A - W) \cup (W - A)$. However, in developing this measure Kuipers ran into a variety of technical problems. As a consequence he went

on instead to *presume* a notion of 'structurelikeness' as a primitive, and to define a qualitative relation of between-ness relative to that structurelikeness [Kuipers 1992]. That is to say, one theory can be thought nearer the truth than another if its structure lies 'between' that of its competitor and the truth. No attempt is made to define this quantitatively. But this approach too turns out to require some strong assumptions, and has also run into various technical problems [Niiniluoto 1998].

This theoretical verisimilitude promises to avoid the problems for the logical-similarity approach pointed out in [Liu 1999] (see section 1-7). And at first sight, it might also seem that one thing its qualitative approach could make sense of is the phenomenon of useful but stubbornly false theories. Perhaps this between-ness relation can place one such theory between another and the truth? But on closer examination, the relation is still defined essentially by counting true and false constituents. Thus it remains difficult to see how it could really capture the usefulness of, for instance, the classical theory of fluid dynamics. More generally, most of the other now-familiar objections to the verisimilitude and related literatures apply equally to Kuipers's work. Thus for example there is no treatment of the differential seriousness of errors, and hence of the role of extra-formal factors.

Partial truth

So far, all the work we have reviewed has accepted the bivalence of truth values into true and false, and attempted some measure essentially of how much truth as opposed to falsity a theory captures. None has proposed that the truth of an individual statement, as opposed to whole theory, might itself be a matter of degree. But one line of research could be to challenge this fundamental assumption of bivalency itself, and to try instead to formulate a notion of 'partial' truth. Many-valued logic is of course not itself a new idea (see section 2-11). But (to my knowledge) none of its research strands has really been used directly to address our issue of approximate truth in philosophy of science – except one. The recent work of Steven French and others attempts to develop rigorously a notion of partial truth specifically applicable to the evaluation of scientific theories and

idealisations. (For details see, for instance, the account in [French and Ladyman 1998].)

The classic Tarskian definition of truth requires a *language* ('L', say), and an *interpretation* P of that language in a *structure*. Then a *sentence* S of L is true or false only with reference to the particular interpretation P. French adapts this scheme in the following way. The heart of his approach is to view the models underlying interpretations not as completely or exactly mapping some given domain D, but rather as only partially mapping it. Following Kripke, to do this he defines a 'partial structure' such that whereas in the classical case all elements of D are either true or false, now they can be true, false or also unspecified. This in turn enables him to define a partial interpretation of the language in the structure, and hence a notion of a sentence S being partially true if it is fully true with respect to the partial interpretation. So the novelty really lies in the definition of the restricted or partial structure rather than in the definition of truth itself.

French's motivation is to make sense of idealisation from the point of view of the semantic view of theories. That means, in his view, explicating a formal definition of how an idealised model might be said to have captured the truth of *some* of a situation. But this formulation of partial truth can be seen as suffering from the same drawbacks as the other purely logical formulations. Imagine a theory deemed to be partially true in French's sense, that is to say strictly true within a partial domain. Then for our purposes the formulation omits a crucial further consideration – just how significant a part of the whole does this partial domain actually cover? Or equivalently, how important are the bits of the total domain in which the theory is *not* true? This query is of course analogous to the seriousness-of-errors point from earlier sections.

French and Ladyman comment: 'Clearly, [the partial structure] is not conceived of as reflecting the (total) structure of D, but as only partially mirroring D. Nevertheless the partial model ... has to capture some fundamental aspects of D, or some "elements of truth", although it does not mirror D perfectly' [French and Ladyman 1998, p57]. Of course, all hinges on what precisely is meant here by 'capture some fundamental aspects'.

Implicitly, the understanding of 'fundamental' must presumably be with reference to extra-logical criteria, so this just amplifies the objection of the previous paragraph. The formulation of partial truth in itself gives no guidance as to just what aspects should be understood to be 'fundamental'. Once again, in effect relevance is left a primitive. Another problem is that if a theory is strictly true within part of a domain, then this implies that it must have successfully captured exactly the true structure of at least some portion of reality. But this seems to leave no scope for the now familiar phenomenon of useful but stubbornly false theories.

1-6) First ontological approaches

Introduction

All approaches we have tried so far have foundered on similar problems, in particular the failure to incorporate extra-logical factors in a systematic way. In order to remedy that, rather than just the definitions of measures themselves I think it is necessary to focus much more on the interpretation of the components of those measures. And in order in turn to perform this semantic task, it is necessary to frame our definition in terms of entities or variables presumed to exist in the real world. Hence we shall term this an 'ontological' approach. By this means, and by also incorporating pragmatics explicitly, we shall also be able to give an exact definition of 'similarity' rather than just take it as a primitive.

The full development of our own scheme follows in the next chapter. But attempts have already been made to tackle the issue in this kind of way. So first we shall look more closely at the three main examples of this, each of which seeks to specify explicitly a way of incorporating extra-logical factors. It is clear that I think this approach is the best way forward, so I take the following papers to be as it were the vanguard of work in the field

and worth examining in more detail. They represent the first attempts at an ontological account of approximate truth – what can we learn from them?

Giere as a precursor

An early precursor of ontological approaches is the work of Ronald Giere [Giere 1988]. He was disturbed essentially by the 'stubbornly false' critique from section 1-3 earlier. That is to say, purely logical metrics have difficulty explaining how a theory like classical fluid dynamics can be approximately true, since it does not consist of mostly true propositions mixed with a few false ones but rather seems to be everywhere false. Giere preferred instead to think of a theory being (exactly) true of some abstract model, and then of that model being *similar* to some specified real-world system. The notion of approximate truth is then contained in this relation of similarity. The key is to move away from general theories and instead to focus on comparing directly a specific real-world structure with a postulated model of it. This way of conceiving of approximate truth is the precursor of all ontological approaches, including our own one. But unfortunately, like the possible-worlds theorists, Giere's own work leaves his 'similarity' a primitive.

So is this approach really so new? Niiniluoto, for instance, allows both for the idea of a theory's best model being close to the real-world system (what he calls approximate truth), and also for the idea of its whole class of models being close to the real-world system (verisimilitude). Hence in Niiniluoto's scheme, if the theory is maximally informative and has only one model then the distinction between approximate truth and verisimilitude collapses [Niiniluoto 1998, pp18-19]. In other words, if we are considering a singular model not a general theory, then the Giere approach of assessing the similarity between that model and the real world is in fact just exactly what Niiniluoto's approach does anyway. Moreover (unlike Giere) Niiniluoto of course also goes on to provide a metric for defining that similarity.

However, notwithstanding the above, I do not believe that Niiniluoto can escape so

quickly here, or indeed ultimately really escape at all. At root, he does not address the stubborn-falsity critique nor the consequential imperative to move away from his purely logical definition of approximate truth, even though it is this critique and imperative that in fact lie at the heart of the motivation for Giere's approach in the first place. Putting the burden for defining approximate truth onto 'similarity' was really intended by Giere as a way precisely of avoiding having to count true and false logical propositions. Instead, perhaps some other way of assessing similarity might be possible and the key is to find this. This of course is precisely what Niiniluoto's scheme does not achieve. The criticism of Giere that really hits home, and which is made by [Aronson, Harre and Way 1994] among several others, is rather that he left his notion of similarity a primitive. In a sense, the challenge since then has been to put extra-logical flesh onto that particular bone. The first really to attempt this in an ontological framework was Aronson.

Type-hierarchies and verisimilitude

This account is only fully developed in [Aronson, Harre and Way 1994], but much of it, especially its application to the issue of approximate truth, is prefigured in [Aronson 1990]. For ease of exposition, we shall therefore hereon refer to the authors just as 'Aronson'. The heart of their scheme is to take scientific theories to be positing type-hierarchies, in turn intended to capture actual structural relationships between natural kinds. These structural relationships, being between natural kinds, are to be understood realistically – Aronson take them to be metaphysical primitives, in other words they are just the way that the world actually is. If two types are close in a type-hierarchy, then it is because they are indeed really close, since the hierarchy represents by definition the true state of affairs. Aronson's main motivation is to formulate an understanding of approximate truth compatible with, and hence supportive of, realism. (An inspiration is that such hierarchy frameworks have turned out to be useful for knowledge representation in artificial intelligence.)

How does this scheme for a realist representation of theories lead to an account of approximate truth? Roughly speaking, a theory is approximately true if its hierarchy

picks out a type that is similar to the actual type. Two types are deemed similar if they occupy similar locations in the hierarchy's tree. An example of one of these hierarchies might be one with higher-level nodes such as 'mammal', intermediate-level nodes such as 'cat', and lower-level nodes for individual tokens such as 'my cat Smokey'. Then 'cat' and 'dog' are similar in that they are both subtypes of the same supertype 'mammal'. To get a full definition of verisimilitude it is necessary to list all the supertypes the two subtypes have in common, plus all those possessed by one but not the other, and then compare the two subtypes for similarity. (Aronson borrow a measure from the psychologist Tversky for this last operation.) Some nodes may be assigned greater weight than others, as determined (potentially non-circularly) by scientific context (p122). Finally, full truth can then be seen as the limiting case of increasing verisimilitude, where the type a theory picks out is identical to the type of the actual object.

An initial query is: where exactly do these type-hierarchies come from? It is disputable whether the world is really objectively carved up in these ways, or whether on the contrary 'natural' kinds are in fact just a matter of convenience or context [Psillos 1995]. For instance, in biology it is now considered doubtful whether the species types used as illustrations above can in fact even be considered natural kinds at all, given their evolutionary mutability. But if the choice of hierarchy might ultimately be just conventional, why then should we assign any objective weight to some particular choice? In fairness, Aronson do explicitly state that the ordering of natural kinds is to be taken as a metaphysical primitive. But Psillos's real claim here is that this still leaves Aronson's definition of approximate truth at risk from a kind of vicious context-dependence.

I think this charge is itself uncompelling, but it is interesting to see why. Note first that Aronson fully agree that similarity is context-specific, and that degree of similarity may well vary depending on which type-hierarchy we employ (p129). Thus suppose that the true type is a dolphin, but that one competing theory thinks it is actually a fish and another that it is a cat. On one type-hierarchy, both the fish and dolphin are subtypes of 'sea animal', whereas the cat is not. But on another, both the cat and dolphin are subtypes of 'mammal', whereas the fish is not. Therefore the ranking of similarity depends on

which type-hierarchy we adopt, which in turn will depend, Aronson say, on the context of the particular problem we are concerned with. But Psillos claims that this leaves Aronson's definition of verisimilitude still open to precisely the charge of arbitrariness that the invocation of objective type-hierarchies was designed to relieve in the first place. So have we actually gained anything?

But Aronson are explicit that they are interested in accuracy rather than comprehensiveness: 'when it comes to truthlikeness, we are simply comparing the thing or system the proposition refers to with the real thing or system; we are not comparing entire possible worlds in which these things or systems exist' [Aronson 1990, p10]. In order to assess accuracy, we must specify the target situation we are tying to model. And any such specification is inevitably 'subjective'. The key thing is: once *given* the specification of a problem, can we still then give an objective treatment of how well a particular model tackles it? And Aronson's scheme does remain objective in that – the important – sense. Once we know what we are interested in, then the posited objective ordering of natural kinds represented by the hierarchy does yield us an objective measure of similarity. *Given* a specification of our problem (in other words, given a particular choice of type-hierarchy), *then* we can objectively define approximate truth (in other words, then we can objectively define relative location within that hierarchy).

Psillos complains that such 'contextualism' means that the same theory might score differently, depending just on which type-hierarchy we choose to ground our measure of similarity. In other words, Aronson's method offers no technique for measuring the approximate truth of theories *in toto* (i.e. verisimilitude), only for measuring the accuracy of specific models. This is true, but as already argued I do not think is a weakness. Context-dependence is indeed a problem if we are wishing to rank general theories, but not if we are concerned only with context-specific models. And I think that any context-specific definition of approximate truth must take on board the pragmatic issue of interest-relativity. The important thing is to do it satisfactorily, and I think that Aronson do. (See section 2-9 for a full discussion of our own treatment of this issue.)

In my view, a more potent criticism of Aronson concerns their choice of ontology. In particular, as Psillos points out, the use of type-hierarchies is open to the charge of a certain circularity:

> '[Aronson, Harre and Way] construe similarity in terms of locations in a type-hierarchy. But what determines a type-hierarchy in the first place, if not some similarity relation between the types chosen to stratify the hierarchy? If this is so, then it seems rather trivial – and not explanatory – that type-hierarchies determine similarity relations; they are meant to do so. They insist that they do not take the notion of similarity as primitive but that they try to analyse it in terms of locations in type-hierarchies. I agree that given that type-hierarchies are somehow given, their approach to similarity is cogent and really casts new light on the nature and significance of similarity judgments. But, I think, it is not enough to declare the ordering of natural kinds as a "metaphysical primitive" (p123). For this does not advance our understanding of how hierarchies of natural kinds come to being.'
> [Psillos 1995, p181]

In other words, ultimately their notion of similarity rests on the provenance of the type-hierarchies, yet Aronson take the latter as a primitive. So in fact we hardly seem to have moved the invocation of the primitive much further back than Giere had it.

I endorse this criticism, and think it is better to choose instead an ontology that tracks the causal structure of the world. And what better for doing that than the actual causal structure itself? Just assuming a causal structure clearly does not in itself thereby also smuggle in a similarity ordering, so the charge of circularity can be avoided. Moreover, postulating such a causal structure is surely much less controversial than postulating type-hierarchies of natural kinds, and it is therefore correspondingly less controversial to claim a realist interpretation of it. We can also thereby avoid the need to invoke macro-scale natural kinds in contexts such as biology, and presumably social science, where their existence is doubtful. Furthermore, whereas sometimes it is unclear which type-hierarchy should be invoked, presumably there is unambiguously available (on a realist account) only the one actual causal structure of the world. (The possibility of there being

many different valid descriptions of that structure is not a problem here – see chapter 2.) Formulating weights across these causes is then surely a more flexible method of incorporating interest-relativity than switching between relatively unwieldy type-hierarchies. (And note that even then, Aronson still wished also to be able to weight across the nodes of those hierarchies.) And where really valid, type-hierarchies should presumably anyway be expected to 'fall out' naturally from a correct causal description of the world, so we certainly are not losing anything valuable in the switch. Of course, all this now leaves us the task of defining some notion of similarity in terms of causal structure since, as it were, we are no longer receiving it *gratis*, built-in to the definition of the hierarchies. That is the task for chapter 2 (see especially its appendix).

Smith's geometric ontology

Peter Smith has developed an interesting and original ontological notion of approximate truth [Smith 1998]. It arises from his more general study of dynamical systems, and of how the mathematical models of chaos theory can be thought to explain the actual world. He too is following the lead of Giere in that he proposes first that we understand a model as being exactly true of some abstract description, and then second this description as in turn bearing a degree of similarity to some actual system. Approximate truth is then just this degree of similarity between the real world system and the model's postulated system.

His particular twist is that most dynamical systems theories are what he calls 'geometric modelling' theories. That is, they posit a particular geometrical structure, and when applied to the real world they are in effect attributing a certain geometrical structure to real phenomena. These phenomena in turn presumably do possess some particular, possibly different, actual geometrical structure. The definition of approximate truth is then some measure of *geometrical* similarity ('close-tracking') between these two geometrical structures. Since defining such a measure presents no particular technical difficulty, so neither should defining approximate truth. By thinking of similarity explicitly in these geometrical terms it also becomes easy to define it precisely, since

there are many readily available measures of geometrical similarity. His approach carries other advantages too, to whit the usual welcome benefits of an ontological approach – thus it is inherently context-specific, and it is able easily to accommodate the approximate truth of stubbornly false theories.

Nevertheless Smith's approach has shortcomings, and as with Aronson I think they stem ultimately from the choice of ontology. Geometry seems well-suited to the assessment of models from dynamical systems, but less well-suited to scientific theories in general. Most theories are not specified, or easily re-expressed, in geometrical terms. In fairness, Smith himself concedes that his account applies only to the special case of dynamical systems theories, and indeed explicitly does not present his definition as a candidate for a general measure. But notwithstanding this, he still addresses general issues in the approximate truth literature.

The strategy depends critically on our being able to understand the ontology concerned *realistically*. (Aronson had this point in mind when specifying that their type-hierarchies must represent the actual orderings of natural kinds.) The problem in this case is that Smith is in effect invoking a geometrical ontology, but that such an ontology is difficult to understand in the required realistic way. We already noted above its apparent lack of general applicability. And when Smith speaks of being able 'canonically' [Smith 1998, p264] to express a theory in a particular geometrical form, this seems to me to assume a realism that, as it were, needs to be built into the ontology, not added in from outside theory. Similarly, his abstract geometrical descriptions of an ideal pendulum [Smith 1998, p260] implicitly require additional realist assumptions in order to do their philosophical work – just abstract descriptions on their own are insufficient. And ironically, it is Smith himself who introduces an example that seems to me to illustrate especially well the benefits of a more realistic, and hence more generally applicable, ontology.

Suppose we have two models of planetary motion, one Ptolemaic and one Newtonian. Suppose next the parameters in them are so chosen, as is perfectly possible, that the

Ptolemaic model actually tracks the apparent movement of the planets better than the Newtonian one does. How in this case can we capture the sense that the Newtonian model is still closer to the truth? This example embarrasses a geometric definition of approximate truth, since here it is the Ptolemaic model that would seem to be close-tracking the planets' trajectories better. In response, Smith is forced to appeal to a wider context of Hempelian unification. In particular, we should consider (he argues) the Newtonian model only as part of a package of generally successful Newtonian theory, and the Ptolemaic model by contrast only as part of a relatively more failing package. Viewing matters in this broader light, we become justified in preferring the Newtonian model after all. Although the Ptolemaic one may give a better geometrical track right here, in other contexts a Newtonian approach has shown itself superior. But this kind of holism runs counter to one of the principal advantages of the ontological approach, namely its ability to make sense of singular judgments of approximate truth. It implicitly is saying, for apparently arbitrary reasons of convenience, that the geometric definition should in this case be applied only to general theories, not to particular models.

The point is that the problem may become much more manageable if, rather than geometry, we talk instead in terms of causation. Then we can easily imagine that the Newtonian model more accurately captures the actual causal processes at work in the heavens, and accordingly that on some causal definition it will be assigned greater approximate truth. Admittedly, even when working with causation the analysis of this example turns out to be tricky due to the different ontologies of the competing models, and a holistic element may enter the interpretation of the reference of each model (see sections 2-12 to 2-14 for extended discussion). And admittedly also, we saw that getting round the Miller problem involves holistic considerations too (section 1-4). Nevertheless, it still seems to me desirable to be able to express the sense in which the Newtonian model is preferable without a forced reference *in our very definition* to extraneous contexts and the performances of other models in other problems. At root a causal ontology, on a realistic interpretation, enables us straightaway to target as it were what is actually going on in the world, whereas the geometric ontology is revealed by this example to carry only a more instrumental flavour.

Finally, a quick note on the analysis of interest-relativity. Like almost everyone (including us), Smith agrees that any judgment of approximate truth must unavoidably be relative to the particular aspects of a problem in which we happen to be interested: 'a theory is approximately true if the world exhibits a *relevant* structure sufficiently similar to the abstract structure specified by the theory (though *which* similarities to weight will, no doubt, be interest-relative)' [Smith 1998, p264, first emphasis mine, see also his footnote 9 p259]. But when the issue crops up again later in the paper there seems to be more confusion. In dynamical systems, a very small difference in initial conditions can 'explode' into a huge divergence in final outcomes. Thus the same model may seem to be approximately true of the initial conditions, but not at all approximately true of the final outcome. In order to get round this ambiguity, Smith proposes that: 'we need to stress the fact that chaotic theories prioritize other, more abstract, metric and topological similarity-relations between the chaotic models and worldly behaviour, and the theories can count as getting near the truth in virtue of these similarities' [Smith 1998, p274]. The implication is that there exists a unique answer here and that we can resolve the apparent ambiguity objectively. But my own view is that in such cases there is an unavoidable subjective element – namely, are we interested in the initial conditions, or the final outcome? *Once* we have characterised our subjective interest, *then* we can formulate an objective measure. So here the ambiguity can only be resolved subjectively, and after that – but only after that – we can bring in an objective apparatus. I think this is an important point to be clear about – see section 2-9 for a full analysis of interest-relativity.

Barnes's approximate causal explanation

To sum up so far: both Aronson and Smith suggest interesting, essentially ontological, ways of defining approximate truth. But despite many virtues, each in the end is handicapped by its actual choice of ontology. A causal ontology on the other hand, understood realistically, may be more promising and is the basis of our own scheme in chapter 2. In this subsection we look at the (to my knowledge) one serious attempt in the literature at such a causal explication – Eric Barnes's development of the notion of

approximate causal explanation [Barnes 1995], which he offers as a more reliable instrument for demonstrating scientific progress than bald approximate truth. As he notes, the literature on approximate explanation is remarkably sparse. He gives one reference to [Tuomela 1985], which recalls Giere's work. Otherwise, almost the only precursor is [Pearce and Rantala 1985], but that paper is chiefly concerned with whether an approximately true theory can nevertheless be thought to furnish deductive-nomological explanations in the same way as a strictly true one. No account is given of approximation itself, nor indeed much fresh insight into Barnes's specific concern of scientific progress. (We discuss ourselves the relation between approximate truth and approximate explanation in section 2-8.)

Barnes nowhere mentions Giere, nor indeed any explicit metaphysical program for tackling the issue of approximate explanation. But he takes there to be an underlying causal reality and in effect defines approximation by the extent to which different theories capture it, and this of course is just the ontological approach we have been advocating. Giere takes theories, or models, to posit a structure in a possible world which can then be compared for similarity with one in the real world. If this comparison is of the different worlds' *causal* structures, then Giere and Barnes would seem to be following the same strategy. (This is in fact also the strategy advocated in [Tuomela 1985], but [Barnes 1995] only refers to it briefly.)

Barnes uses for illustration a simple example of his own invention, where we are to imagine a brick being pushed off a table due to the action of three separate forces 'a', 'b' and 'c'. These forces are individually necessary but only jointly sufficient to cause the brick to fall. The complete causal explanation is therefore A&B&C, where the 'A' means the theory states that force 'a' was applied, and so on for all the three forces. Consider now two competing explanations: X_1 states ~A&B&C, while X_2 states ~A&~B&~C. The intuition is then that, although both false, of the two explanations we nonetheless prefer X_1 since it at least gets two of the three contributing causes right, whereas X_2 gets all three wrong. Although, being false, neither theory counts as a full explanation, Barnes's aim is to explicate a sense of approximate explanation which enables us to express the

intuition that X_1 is somehow preferable all the same.

The main burden of Barnes's paper is to demonstrate that we can circumvent Miller's problem by focusing on the underlying causes a, b and c rather than on the linguistic expressions of them A, B and C. To do this, he needs to set up a slightly cumbersome apparatus of 'conditional contributing causes' [p219], which arguably runs into some technical problems (see [Niiniluoto 1998, p17]. But I do not believe that his approach would really circumvent Miller's problem in any case since, as Niiniluoto comments: 'Barnes seems to assume realism about the forces a, b and c, whereas a Millerian translation might be applied to them as well' [Niiniluoto 1998, p18]. I believe that Barnes's concentration on Miller's problem is misguided (section 1-4), and distracts him from addressing other more pressing issues associated with a causal approach. (In fairness, Barnes does explicitly acknowledge that his account is likely to leave much scope for further development.)

Among the basic possibilities that Barnes does not address is what happens when some causes are more important than others. Suppose cause X raises the probability of an effect by 0.7, cause Y by 0.1 and cause Z by 0.01. On Barnes's account, an explanation citing Y and Z should be awarded a higher score than one citing just X, but this seems unsatisfactory. Similarly, what if a model cites a cause but gets its strength wrong, for instance quoting X above as increasing the probability of the effect by only 0.5 instead of 0.7? Our own scheme will be tailored precisely to handle such considerations. [Niiniluoto 1998, pp17-18] also correctly raises questions about Barnes's treatment of the composition of causes, another lacuna in his account. This issue cannot be dealt with satisfactorily here without looking at how to define the strengths of composed causes, a good analysis of which turns out to depend critically on adopting a context-specific approach (chapter 3). Barnes's discussion of his own conditional contributing causes arguably also suffers for want of a greater focus on context-specificity. Therefore, although starting out with an apparently very similar approach to our own, Barnes's paper ends up heading in a rather different direction.

Finally, two other acknowledgements. First, [Weston 1992, p68] also supports the adoption of a realistically understood causal ontology for tackling approximate truth. His actual scheme is rather different to the approach advocated in this chapter though, concentrating as it does on the evaluation of general theories rather than context-specific models. Second, much the closest anticipation comes from Paul Humphreys, who points the way to exactly the kind of scheme we develop ourselves in the next chapter:

> 'I believe that this kind of causal approach also captures rather better than do traditional accounts how we approach closer to the whole truth. Many accounts of ... verisimilitude use a counting measure on the degree of correspondence between correct state descriptions and proferred state descriptions. That can be replaced by a similar counting measure on [the set of causes mentioned in an explanation]. One can make this more precise and include a measure of the relative contributions of causal factors ... but I shall not pursue that here. ... The more factors cited and the more they contribute to the explanandum, the better and more complete the explanation. We may rank explanatory factors according to the degree to which they contribute to the effect.'
> [Humphreys 1990, p115 footnote 28]

However, this passage comes from a footnote, and unfortunately (as far as I know) Humphreys nowhere developed this idea any further than in these words above.

1-7) Idealisation, laws of nature

Idealisation and approximate truth

Idealisation examines the relation between an abstract model and the messier real world. Typically such a model will capture only some aspects of the world, or will only resemble rather than exactly represent it. Clearly therefore idealisation has a connection with the issue of approximate truth. Moreover, many strands of work in this field follows

our own eventual scheme in being framed context-specifically and also causally, and so may appear promising as a source of insights. However, it turns out that its concerns are, as it were, typically somewhat tangential to ours. In particular, much of the focus has been methodological, with a resultant lack of precise proposals – as opposed to general approaches – for defining approximate truth. We shall mention here only that work which most closely anticipates our general concerns.

An early proponent of causal metaphysics was J.S. Mill [Mill 1846]. He saw that economic laws were rarely empirically fully correct, but felt that they did seem nevertheless to capture something of the world – essentially the intuition with which we also started (section 1-1). His solution was the notion of 'tendency laws'. These capture causal factors that do indeed actually operate in the world, but the reason their predictions often fail empirically is the simultaneous operation of *other* causal factors muddying the water. Only when these other factors, or 'disturbing causes' in Mill's phrase, are not present can the operation of tendency laws actually be observed unimpaired.

Mill's account of scientific models describing causal factors that may form only part of the total causal structure actually present in a situation, anticipates in part our own approach. Moreover, his account is also implicitly context-specific, as we desire. However, as it stands we need to add to it in order to reach a useful conception of approximate truth. We want, recall, an understanding of to what degree a particular model does actually capture the truth of a particular problem. Clearly a key factor in determining this must be some way of expressing how serious the disturbing causes are in any specific context. For any particular case, *how much* do the disturbing causes disturb? Mill never really analyses this issue in detail. He therefore correspondingly never really provides us with a detailed way of saying that one model captures the truth of a particular situation more than does another. In the parlance of this thesis, Mill provides neither a metric nor any detailed treatment of the differential seriousness of errors. (Moreover, his scheme also provides no way of incorporating errors in a theory's specification of the tendency law itself, independent of any disturbing causes, making it difficult to accommodate useful but stubbornly false theories.)

Similar remarks apply to other strands in the literature otherwise reminiscent of our own approach. The idealisation-concretisation procedure of Nowak and associates of the Poznan school [Nowak 1992] sees idealisation as a two-stage process: first, an idealised model is abstracted from a real-world situation; then second, concretising factors are added back to the idealised description in order that the idealised model may describe empirical reality accurately. For instance, a model of an ideal pendulum may be abstracted from reality, its properties studied, and then concretising factors such as air resistance and so on re-introduced until the idealised model describes actual behaviour accurately. As with ceteris paribus laws, this requires well-founded criteria to motivate the reintroduction of its concretising factors else the scheme could be applied in support of any theory at all, no matter how apparently crazy. Nowak recognises this, saying that the concretising factors should be those most relevant to the particular real-world situation, but this 'relevance' is in effect left a primitive. For this reason, the scheme gives us no precise measure for how *well* an idealisation matches the real-life situation, and hence no precise measure of approximate truth.

Maki's method of isolation-deisolation for our purposes has similar strengths and weaknesses to Nowak's idealisation-concretisation. And his notion of 'essesimilitude' – closeness to the relevant truth – while encapsulating nicely our idea of approximate truth, must again be taken as in effect a primitive since it is not really analysed in any way [Maki 1991, 1994]. Cartwright's metaphysics and methodology [Cartwright 1989, 1999] also closely resemble those of our own eventual scheme, but again it is not part of her project to furnish a precise definition of approximate truth. Thus: 'Models, I say, *resemble* the situations they represent' [Cartwright 1999, p193, her emphasis]. This shows her support for an ontological conception of resemblance and hence implicitly of approximate truth, but the term 'resemblance' itself is left in effect as a primitive. Where Mill left disturbing causes unanalysed, Nowak the relevance of his concretising factors and Maki his essesimilitude, Cartwright leaves her 'resemblance' unanalysed too. Similar remarks applied also, as we saw, to Lewis's similarity of possible worlds (section 1-5) and to the work of Giere (section 1-6). In all cases, in effect no exact definition of

approximate truth is offered.

The problem of legisimilitude

A perennial concern in the literature is the relation between idealisation and laws of nature. This has led in turn to a concern in the eyes of some with the possible relevance of laws of nature to the problem of approximate truth. Science is sometimes seen as an attempt to 'carve nature at its joints' or, put more mundanely, an attempt to isolate laws of nature. According to this view, what we should be concerned with is therefore not closeness to mere empirical facts but rather closeness to these laws. In a phrase, our goal should be 'legisimilitude' rather than verisimilitude [Liu 1999]. We saw (section 1-2) that the logical-similarity approach expresses propositions in disjunctive normal form, and geometrically we can take this to define a space in which each of the disjuncts forms an orthogonal axis. As Liu shows, it is natural to think of this space as Euclidean, and indeed most suggested measures boil down to some variety of Euclidean distance function within it. But from the point of view of legisimilitude, such a measure of closeness to the truth now gives rise to serious problems.

We adopt an example of Liu's. Suppose for argument's sake that the special theory of relativity is true, and that the speed of light is precisely 300,000 km/s (in vacuum). Suppose next that an object is accelerated, no doubt with great difficulty, to a speed of 299,000 km/s. Now consider the following two statements:

1) the object is accelerated to a speed of 297,000 km/s.

2) the object is accelerated to a speed of 301,000 km/s.

As Liu points out, using a Euclidean metric to represent approximation both these statements are equally close to the truth, each getting the speed wrong by exactly 2000 km/s. But given the truth of special relativity, statement 2 cannot possibly be true since we must presume it to be physically impossible. Liu argues that, when employing an idealisation in science, often the 'truth' we are seeking to model is a particular pattern, or law, of nature rather than a brute one-off empirical accuracy. Therefore when assessing closeness to the truth here we should prefer statement 1 to 2, but on the face of it a

Euclidean metric is unable to capture this. The nature of physical reality seems to imply that our Euclidean space for representing closeness to the truth should, so to speak, be thought of as non-flat.

Now an obvious rejoinder is to claim that the space of basic disjunctions incorporates, so to speak, the laws of nature in its structure in the first place. So in our example the space would somehow incorporate the special theory of relativity so that statement 2 would no longer be measured closer to the truth than statement 1. One issue arising immediately is how much weighting we should assign to law-transgressions – should statement 2 be adjudged ten times further from the truth than statement 1? A hundred times? Infinitely many? Should we expect our logical formulation somehow to give us the answer endogenously? But leaving that issue aside, Liu shows how in any case a little examination suggests we would be hard put to incorporate as desired laws of nature into the space's very structure.

Consider a second example [Liu 1999, p253 footnote 8]. Suppose our universe consists only of balls of plutonium and that none of these balls is above the critical mass required to trigger a catastrophic chain reaction. Hence no such chain-reaction occurs, and all the atoms remain plutonium. Now consider two possible models of this universe:
A) one of the plutonium balls is above the critical mass, and there is a catastrophic chain-reaction.
B) one of the plutonium balls is above the critical mass, but there is no catastrophic chain-reaction.
Just counting the categorical properties of the atoms, statement B is certainly closer to the truth than statement A – according to it, all the atoms remain plutonium, which is correct, and all but one balls are below the critical mass, which is almost perfectly correct again, going wrong only by citing the one overweight ball. In statement A by contrast, because of the catastrophic chain-reaction many of the atoms do not remain plutonium, exactly contrary to the real world. Of course, the difficulty is that if there *were* an overweight ball, then according to the laws of physics there should be a catastrophic chain-reaction. So although on any logical measure closer to the truth, statement B suffers (unlike

statement A) from implicitly violating the laws of physics by postulating a world in which an overweight ball does not lead to a chain-reaction.

The suggested escape route for a logical measure here would normally be to state that the rest of the world under statement B would be extremely different from the actual world, due to the different laws of physics implicitly postulated, whereas this does not apply to statement A. Therefore taking into account the rest of the world, statement A would in fact end up being measured closer to the truth after all. This already begs questions about how we are weighting between different aspects of the world when performing our measure of closeness to the truth. But in any case the manoeuvre does not work here since we took by assumption our universe to consist *only* of these plutonium balls, and hence the 'rest of the world' looks the same for either statement, namely non-existent. Thus it seems that a logical measure must inevitably and unsatisfactorily rank statement B closer to the truth than statement A.

The only remaining remedy would seem to be simply to declare statement B and all statements like it physically impossible, and therefore removed from our consideration. But of course any incorrect statement of laws is 'physically impossible' – yet that does not stop us wanting to be able to judge the closeness to the truth of, for instance, a Newtonian model of the moon's orbit (as opposed to the relativistic model, or whatever the true state of affairs actually is). The whole point of a notion of approximate truth is in part precisely to cover cases of this sort. Moreover, how are we to make the decision as to what models are to be considered inadmissible due to their physical impossibility? Our current ideas as to what is and is not physically possible may be mistaken – they certainly have been many times in the past. Plus when assessing physical impossibility we are clearly again being forced to smuggle in extra-logical considerations.

(Niiniluoto is aware of this line of criticism. He suggests an answer to it in which the initial cognitive problem is taken to include modal statements. The basic framework of his similarity approach can then be employed as before, except this time the distance between a proposition and the truth is in part a distance between modal basic statements.)

The same issue also arises in connection with other approaches to approximate truth. For instance, if the possible-worlds approach is taken to be equivalent to the logical-similarity one (section 1-5), then the same problems would apply equally as much there. But depending on how exactly 'similarity' is understood, it is possible by contrast that Hilpinen's nested spheres do take nomological considerations into account. In this latter case, the difference between the possible world of an approximately true statement and the actual world, would be factual and within the limits of laws of nature. But as Liu points out, this particular understanding of similarity, while sidestepping one problem, now leaves little scope for capturing the approximate truth of nomologically incorrect theories. Thus we would seem forced to conclude unsatisfactorily that there is no sense in which, for example, Newton's laws are approximately true. Moreover the first problem raised in [Liu 1999] would still remain unaddressed – namely that sometimes a statement may be factually closer to the truth than a rival, but nevertheless seem less preferable if at the same time it is reflecting an underlying law less adequately.

As Liu points out, even [Weston 1992]'s attempt to put flesh on the bones of 'similarity' between worlds still leaves question marks on this point. 'The determination of [Weston's] sense of approximation is left with so much liberty (or so dependent on the contexts of a theory whose statements are under evaluation) that one cannot be sure whether the problems apply' [Liu 1999, p10]. Liu judges that in one concrete example of Weston's [Weston 1987, p60], they still do.

Our solution

I think that the problem of legisimilitude is indeed a serious one for all non-ontological approaches. Perhaps it can be viewed as another example of the seriousness-of-errors problem. The legisimilitude response is to invoke laws of nature, but in this thesis we shall prefer instead to invoke causation. Legisimilitude's focus on general theories is troublesome in the light of our preference for a context-specific approach to approximate truth. Certainly, Liu offers no definition of legisimilitude and the only thing close in the

literature that I know of is Kuipers's 'theoretical verisimilitude', which suffers from its own problems (section 1-5). Accordingly, I think it is best to respond to legisimilitude's criticisms in a different way. The underlying problem they react to – namely the inadequacy of a purely logical definition of approximate truth – can be addressed without the need to invoke laws of nature.

We shall need to anticipate here our own eventual definition of approximate truth developed in chapter 2 – briefly, this consists of comparing for similarity true values for relevant causal strengths with the values postulated by a model. Return now to Liu's special relativity example and its two errant models, only one of which implied superluminal speed. Armed only with the models' numerical predictions, we have insufficient information to determine just what *causal strengths* each is positing. Perhaps the superluminal one is indeed relativistic and has just a minor calibration error, whereas the subluminal one is merely a simple Newtonian model. In that case, legisimilitude might well now prefer the *super*luminal model. A realistically understood accuracy of causal description, not an abstractly understood fidelity to laws, is the more illuminating.

To see this more clearly, turn now to Liu's plutonium example and re-express it in causal terms. Recall, model A posited there was one ball above critical mass and hence a catastrophic chain reaction, model B posited one ball above critical mass but – contrary to the laws of physics – no resultant chain reaction, while the true situation was that there were no balls above critical mass (and hence no chain reaction). It is crucial here to be clear just what our focus of interest, and hence our target causal strength, is. One possibility is that we are interested in the strength of gravitational attraction and hence in the total mass of all the balls, discounting (let us assume) the mass of any energy released as heat or light by a catastrophic chain reaction. Suppose for simplicity that there are 10 plutonium balls of mass 1 each (in some units), that a ball above critical mass would have a mass of 1.1 units, and that the balls are so arranged that the explosion from a chain reaction in one would in trigger chain reactions in the others. Suppose that this catastrophic scenario would leave as a by-product new material totalling a mass of 0.8 for each original plutonium ball. That would mean that the true total mass is 10 units, that

model B posits a total mass of 10.1 units, and model A a total mass of 8 units. Clearly, model B is the more accurate here, and *with respect to this particular causal strength* it therefore should be preferred.

Still, it might seem that we do also want to be able to reward fidelity to the laws of nature. But the point is that we can equally well do this too – if (and only if) it is such fidelity that is our focus of interest. Thus we can imagine a second causal strength, this time one which in effect incorporates a concern with a model's legisimilitude, and then see how close to the truth each model scores this time. So imagine now a causal strength that correlates with the *variance* of the masses of the balls. Then the true situation is that this variance is zero since all the balls have the same mass of 1 unit. Model A also posits zero variance since all the balls now have 0.8 mass, while model B posits a small but positive variance since it posits one ball with a mass of 1.1 and all the rest with masses of 1. Thus with respect to this new causal strength, now it is model A that is correctly judged to be closer to the truth.

In this toy universe the law of nature implies that the variance of the atoms must always be zero, either because there is no ball above the critical mass (as in actuality), or else because one did slip above critical mass and triggered a chain reaction after which all balls are left with the new lower mass. Thus this new causal strength serves as a good proxy for legisimilitude. But the key wider point, and one much emphasised in chapter 2, is that which model is 'closer to the truth' will depend critically on which *aspect* of the truth we are interested in. And, rather than always giving automatic priority to legisimilitude, I think it is desirable that our definition incorporate that flexibility.

Chapter Two – Our own account

2-1) The basic idea

Introductory example

What happens when, standing on the Earth's surface, we drop a ball? Of course, the ball falls to the ground. But if, for instance, we wish to know how fast it falls we need some more detailed picture. In particular, we would need to know the causal strength of the gravity at producing the effect of accelerating the ball downwards. Assume for now that the world is like the simplified one of a Newtonian model. Thus we have a perfect vacuum, the ball is a point-mass, its mass negligible compared to that of the Earth, there are no other forces to take into account, and so on. Assume also that the gravitational attraction of the Earth on the ball can be represented as the sum of the gravitational attractions of the Earth's component parts.

Next suppose that there is a large mountain near where the ball is being dropped. In this case, the true pattern of gravitational pull on the ball is a large amount from the main body of the Earth, plus a little extra from the nearby mountain. Now a standard Newtonian model of the situation assumes among other things the Earth to be a perfect and uniform sphere, and so would neglect the existence of the mountain. Therefore there is a mismatch between the true causal structure, which includes the gravitational pull from the mountain, and the model's posited causal structure, which does not. We might represent the true state of affairs by (1,1), where for (x,y) x is the gravitational pull due to the main body of the Earth and y the extra pull from the mountain. Then the model's posited state of affairs would correspond to (1,0).

So reality is (1,1), the model (1,0). How to assess the closeness-of-fit here? An obvious strategy is to interpret these representations as *vectors*, presumably defined on some *abstract* vector space of causes. Assessing closeness-of-fit then becomes a geometrical matter of comparing the similarity of two vectors. Various measures are possible – which one to choose is discussed in the appendix. The point here is that the assessment of approximate truth would have been reduced to a geometrical issue. And to recap, such an approach to approximate truth promises to fulfil our various desiderata from chapter 1:

it is ontological and context-specific. It is also of general applicability, and moreover easily made quantitative.

Causal strengths: an objective weighting function

Suppose we have a second model which captures the gravitational pull on our ball due to the mountain, but which ignores the pull from the main body of the Earth. Using our previous notation, this model can be represented by a vector (0,1) in our hypothetical cause-space. Now suppose we want to compare the relative performance of the two models, (1,0) and (0,1), in capturing the reality (1,1). By most geometrical measures, (0,1) will presumably be as close to (1,1) as (1,0) is. So as it stands it seems we would have to adjudge the two models to have performed equally well. But this offends intuition badly. The gravitational pull from the mountain is only a tiny fraction of that from the rest of the Earth. Therefore we want to say that a model capturing the Earth's pull but not the mountain's must be much nearer the truth than one the other way round.

The root of this objection lies *in reality itself* – the pull of the Earth really does greatly outweigh that of the mountain. Or, alternatively put, the causal strength here of the Earth is much greater than that of the mountain. (Throughout this thesis, we shall understand 'causal strength' to be a particular, i.e. the strength a cause has in one particular situation rather than generally – for instance the degree of gravitational pull on the ball when it is dropped rather than the general inverse-square law.) Accordingly, we are justified here in putting more weight on the Earth's pull than the mountain's, and this justification stems ultimately from *objective* properties of the actual causes involved. Therefore we can postulate some 'weighting function' that represents these objective properties. In this instance the Earth might have, say, ten million times as much mass as the mountain, in which case we could represent reality not by (1,1) as before, but instead by (10mn, 1). Then the model capturing the Earth but not the mountain would be (10mn, 0), and the model of just the mountain would be (0,1). And so finally the Earth-model would score for approximate truth in proportion to the similarity between (10mn, 1) and (10mn, 0), and the mountain-model in proportion to the similarity between (10mn, 1) and (0,1). We

can now easily imagine measures of geometrical similarity that rank the Earth-model the better fit here, as we would desire.

If we did not use objective causal strengths to license weighting functions like this, we would be left prey to unpleasant paradoxes. Suppose we subdivided the causes in a new way, into the main sphere of the Earth as before, but now splitting the mountain by considering its western half and eastern half separately. Without weighting functions we might represent the three components by (1,1,1) in a new three-dimensional hypothetical cause-space. So now the model capturing all the Earth but none of the mountain would be represented by (1,0,0), and the mountain-only model – capturing both its halves – by (0,1,1). Clearly we would expect (0,1,1) to score higher than (1,0,0) for similarity to the true state of affairs (1,1,1) – i.e. the mountain-model would score more highly than the Earth-model. And yet previously, without weighting functions, the two models scored the same - for *exactly the same physical situation.* In other words simply by changing our arbitrary partition of the different causal elements, the ranking of the two different models changed. This is surely unacceptable – we want our measure of approximate truth to reflect how well a model captures the physical reality 'out there', and this certainly should not be affected by such an arbitrary change in our representation of it. With this new subdivision there would also follow the intuitively unpalatable conclusion that the mountain-model is now ranked *better* than the Earth one.

But by taking into account causal strengths and hence our weights, both these difficulties can be avoided. The objective causal strengths of the Earth and each half of the mountain would now be, respectively, (10mn, 0.5, 0.5). The Earth-model would be represented by (10mn, 0, 0), and the mountain-model by (0, 0.5, 0.5), which clearly suggests that the Earth-model would still score much the better, just as desired. (See the appendix for more on the problem of arbitrary re-description.) Use of a weighting function therefore in effect solves the problem we earlier labelled 'seriousness of errors' (section 1-3). Our approach adjusts to take into account a model's errors and, more importantly, it adjusts *in proportion to* the seriousness of those errors. Thus in the final example above, the mountain-model's single omission was shown to be much more serious than the Earth-

model's two.

Several of the suggested definitions in the literature, for instance Niiniluoto's, do formally allow scope for incorporating a weighting function over their linguistic statements. But none of them relates the role of these weights to ontological considerations, and hence none can motivate any objective interpretation of them. We are therefore already well ahead of the performance of such metrics.

This then is the basic framework of our own approach. In chapter 3 the issue of how to define causal strengths is looked at in detail. In this chapter our examples will, as above, merely quote causal weightings somewhat schematically in order to illustrate the philosophical point at hand. In the next few sections, meanwhile, we shall examine an important conceptual wrinkle in what we even mean by 'approximate truth' and how our definition therefore needs to be refined in the light of it.

2-2) Two different intuitions

I think a central, although rarely emphasised, aspect of approximate truth is that – contrary perhaps to our naïve pre-philosophical expectations – it does not turn out to be a univocal notion, even in principle. Rather, there seem to be two distinct ideas of approximate truth, and it is necessary to disentangle them. Only then can we examine to what extent it may be possible to reconnect them again, as it were.

Imagine two astronomical models. The first is Ptolemaic, comprising a complicated geocentric system of epicycles, and yields a prediction for the movement of a certain planet in the night sky. The second is Newtonian, comprising a heliocentric gravitational system, which also yields a prediction for the movement of the same planet. Assume that

all agree on which patch of light in the sky corresponds to this planet, and that all also agree on how to measure its movement and all other relevant auxiliary assumptions too. Suppose that, as is perfectly conceivable, it turns out that the predictions of the Ptolemaic model are more accurate than those of the Newtonian one. (This example is adapted from one in [Smith 1998].)

Which of the two models should we prefer for approximate truth? On the one hand, it seems that the Newtonian one is clearly preferable. After all, nobody supposes anymore that the other planets orbit around the Earth in accordance with complicated epicycles, whereas we do suppose that they orbit around the sun as the Newtonian model describes. To be sure, the Newtonian theory of gravitational attraction has of course itself been superseded by the relativistic one. Nevertheless, there still seems to be a clear intuition according to which the Newtonian model is closer to the true ontological situation than is the Ptolemaic one.

But on the other hand, if we restrict ourselves purely to empirical predictions about the movement of the planet, the situation is reversed. Now, notwithstanding its weird array of epicycles, it is the Ptolemaic model that scores more highly. That is, the predicted position of the planet is closer to the true position (as viewed from Earth) in the Ptolemaic than in the Newtonian case. Thus, for example, a navigator dependent on knowing the future movement of this planet across the sky would be better advised to consult the Ptolemaic model, notwithstanding its ontological peculiarity.

Perhaps it might be objected that the Ptolemaic model's greater predictive accuracy is a cheap artefact that we should not take seriously, for it is possible to add extra epicycles into the Ptolemaic system in an *ad hoc* way so as to predict any particular planetary movement more accurately. And, the objection runs, it seems strange to reward such *ad hoc* and indeed fictitious additions. Nevertheless, judging purely by predictive accuracy, the Ptolemaic model remains better. The fact that the epicycles are fictitious impinges only on the first, ontological, sense of approximate truth. The epicycles may be fictitious, but the predictive accuracy is not. And while to be *ad hoc* may indeed be undesirable

methodologically, all the same it is irrelevant according to the narrow criterion of predictive accuracy.

No doubt also Newtonian *theory* generally, in addition to its greater ontological appeal, can point to much greater success than Ptolemaic theory generally, even in narrowly predictive terms. (Ultimately, this was presumably the reason it was preferred to it historically, of course.) Thus Newtonian theory may illuminate a huge range of phenomena about which Ptolemaic theory says nothing, including a great range of astronomical phenomena. All this translates as saying that Newtonian theory's general predictive accuracy is much the more impressive. But recall our conclusion (chapter 1) that approximate truth can only sensibly be understood context-specifically. That means here that the greater *general* success of Newtonian theory is irrelevant. All that matters is the approximate truth of this *particular* Newtonian model relative to this particular Ptolemaic one. And according to the narrow predictive criterion, concerning only these models and the position of this particular planet, it is by assumption the Ptolemaic model that fares the better.

There is thus no evading the fact that there exists a distinct sense of approximate truth according to which the Ptolemaic model in this example is preferable. Label this sense *empirical approximate truth*, or EAT. The other sense of approximate truth, according to which the Newtonian model is to be preferred, label *ontological approximate truth*, or OAT. So in this example OAT and EAT give opposite rankings. We have just argued that EAT is indeed distinct from OAT. If desired, this distinctness could also have been seen the other way round, so to speak – it turns out that OAT cannot be reduced to EAT either.

Moreover, and importantly, this is *not* quite the distinction common in the literature between a theory's scope and its accuracy. Newtonian mechanics has much wider scope than a quantum-mechanical model of a laser while being ontologically inferior, so scoring well for OAT does not necessarily correlate with wide scope. And while it is true that scoring well for EAT does seem to correlate with what we mean by accuracy, even

this comes with the caveat that scores for approximate truth must inevitably be interest-relative (section 2-9). Loosely put, perhaps we can think charitably of the OAT/EAT distinction as corresponding to the scope/accuracy one as adjusted for a context-specific scheme such as ours. ([Aronson, Harre and Way 1994], one of the pioneers of an ontological approach to approximate truth, also make an ontological/empirical distinction in exactly this way.)

2-3) More on ontological approximate truth

Suppose we have two models of a situation, M_1 and M_2, say. Then if M1 cites the correct ontology and M_2 an incorrect one, M_1 will always be preferred by OAT, regardless of predictive accuracies. This seems straightforward. But what now if two models cite *different* incorrect ontologies? For example, according to relativity theory Newtonian and Ptolemaic models of celestial mechanics are both ontologically mistaken, but in different ways. (We assume for the sake of this discussion that relativistic and Newtonian mechanics should be thought ontologically incompatible.) In such circumstances, OAT has two possibilities: either both models are just equally wrong, or else we need somehow to find a way of saying that one incorrect ontology is closer to the true one than is the other.

The key question is whether there is any sense, *independent of empirical considerations*, in which the Newtonian ontology is 'closer' to the true relativistic one than is the Ptolemaic ontology? I cannot think of any. To be sure, the accumulated weight of empirical evidence across many different contexts may indeed tend to support a Newtonian over a Ptolemaic approach. But it seems impossible to define any satisfactory purely abstract and non-empirical measure of ontological similarity that gives the same result. Therefore we conclude here that, when two models are both ontologically

mistaken, OAT must adjudge them equally wrong. Hence in the Ptolemy-Newton case, assuming that the real world is non-Newtonian, OAT must adjudge the two models equally far from the truth. The intuition that the Newtonian *ontology* is somehow closer to the true one than the Ptolemaic ontology, we take to be a result of illicitly incorporating from other contexts empirical EAT-type considerations.

What if two models both cite *correct* ontologies? Then again OAT must be neutral, merely awarding full marks to them both. Any finer-grained distinction could again, for similar reasons, ultimately rest only on illicitly imported empirical considerations. For example, suppose that there were three different causes of some event, say that a boulder fell off a ledge as a result of three different pushes on it (all three pushes being actually required to move the boulder). Suppose model M_1 cited the first two of these pushes, and a second model M_2 only the third one. At first sight, perhaps this would suggest that, although M_1 and M_2 both cite the correct ontology, still we would have a reason to prefer M_1 to M_2. But what if the third 'push' was in fact a composite of three individuals pushing simultaneously? Should this not then count as three factors, not just the one? In which case, of course, now it would be the second model that was preferred instead of the first. We conclude that such 'head counting' is too crude a way to compute ontological accuracy – problems of individuation turn out to matter. The only way around these problems is to assign differential weights to each factor, but it is hard to see how this could be motivated except by incorporating some empirical warrant. In our own scheme, we weight different ontological factors by their causal strengths, but it is explicitly acknowledged that this goes beyond purely ontological considerations since our definition of causal strength (roughly speaking) will in turn be defined in terms of the quantity of effect that a cause leads to – which is an empirical, or at any rate extra-ontological, input. The whole issue recalls one of our central criticisms of the similarity approach to approximate truth (section 1-3) – namely the seriousness-of-errors problem. There, the only solution was to incorporate extra-logical factors; analogously, here the only solution would be to incorporate, as it were, extra-ontological factors.

To re-emphasise an important point: it follows that OAT is unable to prefer one model

over another when they each have the correct ontology – *even when one of those models is empirically superior to the other.* For instance, suppose that one ontologically correct model asserts that the Earth rotates about its axis every 24 hours, and so is empirically wrong but only slightly (since the actual period of rotation is fractionally under 24 hours). Suppose a second model is also correct ontologically but this time is grossly incorrect empirically, proposing a period of only 2 hours, say. Although it seems obvious that the first model is preferable, OAT is unable to capture this judgment. Indeed, if the first model were Newtonian and the second Ptolemaic, so that the first now seemed superior to the second both ontologically and empirically, even then OAT must adjudge the two equal – given our earlier remarks about all false ontologies being judged equally false.

So a fine-grained 'pure' OAT definition seems to be impossible. It follows that OAT can therefore only ever offer crude qualitative verdicts. If one model has the true ontology and a second one does not, then (and only then) the first one is preferred. In all other cases, two competing models must be adjudged equal. For this reason, a pure OAT definition alone seems clearly unsatisfactory.

Finally, we should specify just what we mean by a theory or model's 'ontological commitment'. I shall assume a standard Quinean account, so that a model is ontologically committed just to all the things it quantifies over. Note in particular that therefore the entities a model is held to be committed to ontologically will include all the things it cites as causes, since in our scheme the key variables will be the levels of *strength* models assign to each cause. Since we are concentrating on context-specific models, a Quinean approach will tend to mean a model being judged ontologically committed only to variables relevant to the problem at hand. In any case, it turns out I think that the criticisms in this section of a purely OAT approach, and our analyses later of the complicated relation between a model's ontological commitment and what we can say about that model's approximate truth, do not turn out to be unduly sensitive to our precise definition of ontological commitment anyway.

2-4) More on empirical approximate truth

With the Newtonian and Ptolemaic models, from the point of view of OAT we were struggling even to render them commensurable. This of course is the attraction of switching to empirical predictive accuracy – we always have a common currency of accurate predictions (although see section 2-12 later). Thus we could have compared the Newtonian and Ptolemaic models' predictions for the movement of the planet, and thereby adjudged one of them more accurate than the other. Then, assuming some measure of similarity, we would have our score for approximate truth. So is this the panacea? Unfortunately, no it is not. The reason is that the ontological factor is left completely ignored, and this turns out to lead to unsatisfactory consequences. Also, EAT alone is particularly vulnerable to Miller's problem.

Begin with Miller. [Miller 1975], recall, demonstrates that our rankings for approximate truth will vary depending on which empirical parameters we choose to measure it by. More precisely, for any two parameters A and B, we can in general define two new parameters C and D – themselves defined in terms of A and B – for which our ranking of two models will reverse [Miller 1994]. It is therefore not enough just to specify approximate truth in terms of empirical predictive accuracy. Rather, we must also specify exactly which empirical parameters we need to be predictively accurate *about*. Recall that the answer to the Miller problem is to be able to justify privileging one choice of parameters over another. The point now is that just which parameters we so privilege may depend in part on our prior ontological commitments. Perhaps ontological factors may influence one investigator to focus on the parameter pair A and B and another on the pair C and D. In which case, since EAT may depend on this choice, EAT in turn may depend indirectly on our ontological commitments after all. (Miller makes the point that the mere completion of Brahe's data would not on its own render Kepler's laws fully true – or in our terms, that EAT on its own is insufficient for a full account of approximate

truth [Miller 1994, p223].)

More commonly though, I suspect that in practice competing modellers will likely agree on which empirical parameter on which to concentrate, even given ontological disagreements. For instance, our Ptolemaic and Newtonian astronomers may well have agreed on which observations would decide the matter – namely the apparent movement of the relevant planet. Moreover, there is no reason to suppose that they would not also have agreed on what counted as good evidence for this, given the uncontroversial nature of the relevant auxiliary assumptions. Similarly, phlogiston and modern theorists of chemistry might well agree on how to collect a particular reagent in a test tube, measure its weight or volume, and so on. At any rate, it seems plausible that often scientific dispute will not concern the *choice* of the relevant observables, so in those cases the Miller issue will not arise. But even then, still there would remain another important reason why EAT alone is unsatisfactory – namely that often one model may seem much preferable to another even when their relevant empirical predictions agree.

Consider the following example. For the last 17 (or thereabouts) US presidential elections in a row, the following relation has held: if the Washington Redskins American football team has won its last home match before the election, then the incumbent (defined to mean the sitting president, vice-president or candidate from that party) has won the subsequent election; if the Redskins have lost, then so has the incumbent. (The Redskins' loss in November 2000 was thus a reliable indicator that Bush would subsequently be (declared) the winner over Gore in Florida...) Suppose now we cite two models of US presidential election results: M_1 attempts to incorporate all the usual political factors, such as candidates' personalities, state of the economy, and so on. M_2 postulates instead something more direct – that it is the result of the relevant Redskins game that 'magically' determines the subsequent election. Suppose that both models (retrospectively) predict the results of the elections equally successfully. (Assume that M_2 quantifies over its magical relations, so that it is indeed deemed committed to them ontologically. Assume also, plausibly, that each model also agrees on what the relevant empirical parameters are, i.e. on who won each election.)

How then should we rank them for approximate truth? EAT would be forced to say that both score equally. Yet intuitively, of course, we want to say that the political model M_1 is much superior to the magic model M_2. It seems clear that M_1 has captured more of what is 'really going on'. Similarly, *any* 'model' that gave the list of election winners correctly would have to be ranked by EAT equally with M_1, no matter how ad hoc or arbitrary the proferred explanation. For instance, we might simply copy the list of winners from a book, and model this by saying that 'fate willed it'. Essentially, the problem is that EAT must rank all competing explanations the same, given the successful 'prediction' of the final explanandum. This leaves it no resource for distinguishing between true and false explanations. A corollary is that it is unable to recognise the concept of 'fluke' correlations (i.e. correlations between A and B even though there is no causal relation between A and B, nor any common cause of them). Yet distinguishing between true and false explanations, or between flukily and 'genuinely' correct models (i.e. where there *is* some causal connection between A and B), is surely at the core of science, else we must take seriously any wild ramblings so long as they happen to predict a relevant variable correctly. In the case of historical explanations the correct value of variables is already known, so such wild ramblings would become still less impressive and the need to deny them high scientific status a still more acute desideratum. Any definition of approximate truth unable to do this must, I argue, be seriously deficient.

The intuition here is of course the same as the one supporting the Newtonian model in our Newton-Ptolemy example, which was our original motivation for formulating OAT. In cases like the Redskins-election correlation though, the motivation is even stronger. Whereas we assumed that the Newtonian model was empirically inferior to the Ptolemaic one, which did provide at least some argument in favour of the latter, here the magic model of elections does not even have the crutch of superior empirical support. Thus there is no positive reason for preferring it, and still the same ontological reason for disliking it; as it were, the balance of arguments is even more in favour of the political model than it was for the Newtonian model.

To sum up where we have got to so far: first, there exist two distinct senses of approximate truth, labelled by us OAT and EAT. Second, OAT on its own is unsatisfactory since it can offer only crude qualitative verdicts and cannot incorporate considerations of empirical accuracy. Third, EAT on its own is also unsatisfactory since it cannot discount false explanations or fluke empirical successes, which stems from its inability to incorporate ontological considerations. So the question now is: is there any satisfactory way we can combine the two? Is it possible to get, so to speak, the best of both worlds?

2-5) Our own scheme

Initially, this is perhaps most easily illustrated via an example. Suppose we are interested in the Earth's atmosphere, and in particular in how it may have been changing in recent decades, perhaps in part due to global warming. One actual study recently looked closely at the altitudes of different layers in the atmosphere, and in particular at the altitude of the tropopause, which is the boundary between the troposphere and stratosphere [Santer et al 2003]. This has risen through the second half of the twentieth century, indeed several hundred metres since 1979 alone. The researchers examined five possible causes of this change: the level of greenhouse gases, sunlight reflected from airborne solid particles, atmospheric ozone concentration, the sun's output of heat and light, and dust injected into the atmosphere by volcanoes. The first three of these five were considered to be man-made, the latter two natural. It is known from previous work that all five of the causes do affect the air temperature at different altitudes, and hence will impact on the heights of different atmospheric layers, and hence in turn on the height of the tropopause. The exact causal mechanisms by which the altitude of the tropopause is influenced are complicated. The researchers built a model of this overall process, drawing of course from previous knowledge of atmospheric dynamics, and on the basis of this model sought to explain the

observed change in the empirical variable of interest, namely the height of the tropopause.

So far, so uncontroversial – just another case of scientists seeking to explain observations via a model. The interesting thing is how we should interpret what their goal is, and how we should conceptualise how successful they may or may not have been in achieving that goal. What the researchers found was that the five causes could fully explain the observations (in their model), but that no combination of just four causes could. That is, all of the five causes had non-zero importance or strength (with respect to this effect, namely the recent change in the altitude of the tropopause). However, the researchers also found that *most* of the effect seen in the latter half of the twentieth century was due to just two of the factors, namely the level of greenhouse gases and the ozone concentration. The headline conclusion was therefore that important changes in the atmosphere were due mainly to man-made rather than natural causes. But the main point for us is that these results essentially boiled down to *claims about causal strengths*. In particular, the overall finding was that, whereas none of the five causes had zero strength, still two of them did have greater causal strengths than the other three.

This, I shall claim, illustrates a general template for understanding the degree of success of scientific work. No one supposes that, for instance, the modellers of the atmosphere have the exact truth, given the huge multi-causal complexity of this real-world system and the relative simplicity even of sophisticated models of it. Nevertheless, we do suppose that some attempts to model the atmosphere are better than others. For example, a model that claimed that the rise in the tropopause was because of unusually windy conditions or because of an excess of rain dancing, would seem to be clearly inferior to the advanced multi-causal model we have just been discussing. Moreover, we often further suppose that any individual model can be assigned a particular degree of *partial* success, or what since chapter 1 we have been terming approximate truth. So there are two desiderata here: that we be able to evaluate models in absolute terms and also in relative terms, that is both quantitatively and qualitatively. (Of course, the former would presumably imply the latter.) Our claim is that a focus on causal strengths can deliver us

both of these (but see section 2-7 below). In particular, once given a causal ontology and a target effect, we assume first that there is a unique fact of the matter about what the true strengths of these causes are, and second that any given model in turn proposes particular values for these strengths. The approximate truth of that model is then the degree of similarity between its assigned causal strengths and the true causal strengths. For example in the atmospheric case, if we take it that there is some true answer as to the relative importance of each of the five causes, we can then compare that with the relative importance the modellers assign to them.

This then is our basic definition: given a set of relevant causes, approximate truth is the degree of similarity between a model's postulated weights across them and the true weights across them. This proposed definition of course immediately raises many issues, and we shall devote the remaining sections of this chapter to discussing them. Note now that our scheme provides a way to define the approximate truth only of particular models, not of general theories. But our original desideratum from chapter 1 was for just such a context-specific scheme, in which case this is not a weakness. And it turns out that causal strengths are (on our definition – chapter 3) anyway also definable only context-specifically. But it does mean that under our scheme the nearness to the truth of a whole theory or set of theories, such as 'Newtonian mechanics' or 'special relativity', cannot be defined. My own view is that such general approximate truths, like also general causal strengths, cannot be made coherent sense of. So if, unlike us in this thesis, one's goal is to make sense of the notion that science as a whole is progressing nearer the truth, then our definition can provide no help. The progress it can elucidate is progress *within* a particular problem, so to speak – for instance, one model of the rising tropopause may well be adjudged better than a preceding one, so progress is possible with respect to that particular issue. But my own view is that we simply cannot make any rigorous sense of progress in the broader sense, so claims that science *as a whole* is somehow 'progressing nearer the truth' should be abandoned as impossible to substantiate. (Perhaps this supports Kuhn's famous contention that we can speak of scientific progress with respect to individual puzzle-solving within periods of normal science, but not in the global sense of one theory or paradigm being closer to the truth than its predecessor.)

The important merit of this approach to approximate truth is that it captures the virtues of OAT and EAT while simultaneously avoiding their weaknesses. Recall, OAT on its own was unsatisfactory, since it could offer only crude qualitative verdicts and made no allowance for empirical accuracy. Our definition now can offer detailed quantitative verdicts and, as we shall see shortly, in effect does incorporate considerations of empirical accuracy, thus capturing the merits of EAT. EAT itself, on the other hand, was also on its own unsatisfactory because its disregard of ontological considerations meant that it could not discount either bad explanations or fluke empirical successes. It is precisely these latter weaknesses that our own definition is, as it were, designed to remedy. This it manages by in effect taking into account just those ontological factors that EAT ignores, thereby also capturing the merits of OAT. A re-worked example will help illustrate these claims.

Return to the correlation between the results of American presidential elections and certain Washington Redskins American football games. To explain the election results, we imagined a model M_1 based on the usual political factors like the candidates' personalities and the state of the economy, and a second model M_2 citing instead some 'magical' influence of the football results. Let us adjust this example, so that now M_2 postulates something more direct and less ontologically unconventional – say, that the result of the relevant Redskins game influences voters' respect for Washington and hence for the governing party, and thereby determines the subsequent election. We can set up the problem so that the two models are now in agreement as to the relevant causal structure – that is, they both postulate the same electors with preferences and various forces that may influence those preferences. They disagree only over how much weighting to attach to these causes. In particular, M_1 assigns high weight to economic factors and so forth and only a low weight to the results of football games, whereas M_2 assigns its weights just the other way round. Assume as before that both models are empirically successful, i.e. that they pick the right election winner each time.

In these circumstances, OAT would not be able to pick a winner between the models – by

assumption now, both have the same ontological commitments and so must be adjudged equal for OAT. And since the two are equally empirically successful, each picking the election winners successfully, EAT would be unable to prefer one over the other either. Yet still there remains the clear sense that M_1 is somehow much closer to the truth than the apparently ridiculous M_2. And our own definition of approximate truth can capture this – even though M_1 and M_2 are, as it were, equivalent both ontologically and empirically, still it may be that they have different weightings across causes from each other, and hence one of their set of weightings may be closer to the true one than is the other's. For example, suppose the true causal strengths were 9 for economic factors and 1 for the football results. Then if M_1 cited 10 and 0 respectively for these causes, and M_2 cited 0 and 10, it would follow that M_1 was adjudged nearer the truth (at least for all likely candidates for a measure of similarity).

In general, bad explanations or fluke empirical successes will likely be picked up by our definition. For example, suppose that two people push a ball, person A with a force of 5 units and person B with a force of 5, so that the total force (and hence acceleration of the ball) is 10 units. Suppose that two models agree that the causes of interest are these two pushes, and that the observable effect of interest is the acceleration of the ball. Suppose that M_1 postulates pushes of force 5.5 units from person A and 4.5 units from person B, while M_2 postulates pushes of 9 and 1 unit respectively. Thus, in obvious vector notation, the true weightings should be (5, 5), M_1 postulates (5.5, 4.5), and M_2 postulates (9, 1). Intuitively, it seems clear that M_1 is closer to the truth of the situation than M_2. But in terms of the simple final result M_2 does equally well, since although it greatly overestimates person A's push and greatly underestimates person B's, these two errors as it were fortuitously 'cancel out' so that the *total* push comes out at the true value of 10. Consequently, EAT is unable to prefer M_1 in this situation. Our definition, by contrast, *is* so able.

We can adapt the same thought-example to illustrate how our definition also avoids a weakness of OAT's. Suppose that M_1 still posits (5.5, 4.5) as before, but that now M_2 posits (2, 1). Clearly M_2's prediction for the total effect of only 3 units is now also

seriously awry, so that EAT will immediately prefer M_1 to it. But since M_1 and M_2 still posit exactly the same universe ontologically, disagreeing only over some particular causal weightings, OAT must still adjudge them equal – *despite* the empirical superiority of one over the other. Our own definition, although not being based simply on the final empirical prediction in the manner of EAT, will still also prefer M_1 to M_2 here (again on any plausible measure of similarity). That is, although incorporating ontological factors sufficiently to avoid the pitfalls of EAT, our definition is in effect still able to incorporate considerations of empirical accuracy too, thereby avoiding an egregious fault of OAT. It does this ultimately because our definition of causal strength in effect incorporates empirical factors (chapter 3), even while our overall focus on component causes serves also to incorporate ontological factors. That is, we combine consideration of surface empirical measurements with consideration too of the underlying causes. It is in this way that we are able, as it were, to get the best of both worlds.

2-6) Methodological utility

Simplicity

Our definition of approximate truth is conveniently *simple*. (Or at least it is given a reasonably simple definition of similarity – see appendix.) It just compares postulated causal strengths with the actual ones. All the causal terms are expressed in the language the scientist is naturally working in, and the causal strengths are also defined in a natural way (chapter 3). Thus, there is no need to mention epistemic probabilities or what the research community regards as reasonable values for them, no need to formulate sets of all conjunctions of empirical regularities perhaps subdivided into those independent or not of the current theory, no need to think about margins of imprecision or maximising some abstract mathematical function, and no need to have to consider what are the minimal points into which the relevant logical space can be divided in order to sum over

them to calculate an expected verisimilitude – merely to quote all those quantities occurring in [Bonilla 2002]'s structuralist definition. [Niiniluoto 1987] and [Oddie 1986]'s definitions – perhaps the standard ones in the field if any are – are no less complex. As [Kieseppa 1996, p424] states, for these reasons 'it would be absurd to suppose that the theory of verisimilitude were valuable as a *sociological theory* concerned with the methods that are actually used in the special sciences.' Nor are these complaints equivalent to claiming that a child catching a ball must know complicated trigonometry, for there are no short-cut pragmatic rules offered by these definitions for their calculation.

Can we do better? I think it is possible to imagine, by contrast, two real-life scientists performing something precisely like our calculation, at least implicitly, when arguing over which of two models is best. Of course, there will be disagreement over what the true causal weightings should be – else there would be no dispute! The point is that the criterion by which disputes are to be judged, namely degree of accuracy about causal strengths, is *not* itself in dispute but rather is the implicit common currency. Unless real-life scientists can be shown to be maximising these other complicated definitions of approximate truth without realising it – which seems to me unlikely – those definitions fail descriptively.

Normative warrant

A similar point applies to the normative issue. It is clear that a high score on our own definition is a desirable thing – it means that a model replicates closely the true causal strengths. And on our definition of causal strength (chapter 3), this in turn licenses more successful causal interventions (section 2-7). Hence, ultimately, our scores for approximate truth are sanctioned *empirically* in the sense that those models scoring highly can as a direct result expect to predict the results of interventions more accurately.

It may be that similar normative arguments can be mounted in favour of the alternative, more complicated definitions, but such a task would seem to have to be based on *a priori*

philosophising and at the least seems likely to prove rather more challenging. Even Popper was sceptical about the prospects: 'I do not suggest that the explicit introduction of the idea of verisimilitude will lead to any changes in the theory of method' [Popper 1963, p235]. And as Kieseppa argues, on the definitions of such as the logical-similarity or structuralist approaches, it is in fact not clear why we should even *want* to choose a theory judged by them to have higher verisimilitude. [Laudan 1981, p30] points out that there is no obvious connection between being successful and being verisimilar in the sense of Popper's early definitions, and [Kieseppa 1996, p432] demonstrates how Oddie's and Niiniluoto's favoured rules can yield contradictory rankings in actual cases.

Applicability

Perhaps our own definition's biggest vulnerability is to the cases when it is not just the strengths across causes that are in dispute but rather also the very causal ontology itself (sections 2-12 to 2-14 – note though that no competing definition performs any better in such cases). If it were common for competing models to disagree ontologically, then there would be a correspondingly powerful case against our definition. In the rest of this section we shall argue that in practice it is, on the contrary, much more common for scientific disputes *not* to concern fundamental ontological commitments in this way. That is, controversy concerning what ontology to pick is rarely (although of course not never) at issue in typical scientific debates. Similar remarks apply with respect to the separate issue of choice of vocabulary – namely, that in practice scientific disputes rarely concern choice of vocabulary either. (This extra remark is relevant because the same ontology can support many different vocabularies and because in turn, as the Miller problem famously shows, rankings for approximate truth may be sensitive to the choice between these vocabularies.) Rather, typically scientific controversies concern what weighting to put on various existing causes, or else concern whether some new cause might be relevant (i.e. should now have a non-zero weighting). In other words, as we have been claiming, usually new work concerns the sizes of relative strengths across an undisputed list of causes, or else simply establishes a new causal connection for future use. This view boils down to seeing science typically to be in the business of furnishing

causal *explanations* (on which see section 2-8 below).

Similar points apply perhaps still more strongly when thinking of *historical* explanations, namely that scholarly controversies again typically concern which causes were or were not at play and important, rather than more fundamental matters of ontology or vocabulary concerning the specification of those causes in the first place. By 'historical explanations' we have in mind not just conventional human history, but also all the other historical sciences, for instance much in medical diagnosis, palaeobiology, evolutionary biology, astronomy, geology and archaeology. A good piece of historical work can be thought of as one that has a relatively accurate weighting of the causes behind the effect of interest, and a bad work as one that has not. (I do not of course deny that there is also a role for pragmatic factors like originality when assessing whether a piece of work is 'good'; still, there remains a sense in which a model is or is not accurate independent of whether it is also original, and it is this sense which our definition is seeking to capture.)

To reiterate an earlier point: often scientific research will be aimed at establishing a previously unsuspected causal relationship, rather than arguing over weightings across already known causes. But this can still be thought of as attempting an accurate causal description. Indeed, one might think of such work as assigning a positive weight to a cause previously thought to have zero weight and thus indeed to be about weightings over causes after all. Of course, nothing in thesis should be taken to be denigrating the methodological benefits of such work in helping supply the raw ingredients for subsequent causal explanations of other phenomena.

Of course, not every piece of science fits the particular template of proposing new weightings over ontologically uncontroversial causes. In particular, cases of science proposing radically new ontologies, like the Newtonian or Einsteinian revolutions, do not fit this pattern. Our claim here though is that most actual science, and hence most scientific dispute, concerns cases where the choice of ontology is common ground and it is the causal weightings *within* that agreed ontology that are at issue. In Kuhnian terms, most science is normal science within a paradigm, not revolutionary science shifting

paradigms. In more concrete terms, my claim is that our example of an investigation into atmospheric chemistry is typical of scientific work in this respect, and thus that our definition of approximate truth will typically – albeit not always – be easily applicable.

As mentioned in chapter 1, this view of science stemmed initially from my background in economics. There, in practice arguments rarely revolve around fundamental issues of ontology or vocabulary. The concern is, say, with whether and how a particular monetary policy will reduce unemployment or increase growth, not with whether different researchers mean the same things by 'monetary policy' or 'unemployment' or 'growth'. And even when those latter issues are debated, still the questions at issue are not really ontological. For example, the various definitions of unemployment in the literature are in no sense *ontologically* incompatible. Similar remarks apply to other areas of economics, for instance single currency areas, industry structures, designing an optimal tax regime or auction format, or identifying a profit-maximising price. Arguably, similar remarks apply to social sciences and history as a whole, or at least apply to them as they are overwhelmingly practised.

Examples from natural science

What of the natural sciences? We do not offer any systematic survey or definition of what a 'typical' piece of natural science consists in, only a rather more anecdotal account. Nevertheless, this may be of some value to illustrate that there is much science that does fit our description, or at least to make that claim seem plausible. So take some stories from a single week as representatives of natural science as it is practised.

Sunlight can be used in medical operations

A prototype solar concentrator was shown to be able to kill a section of liver tissue in anaesthetised rats [Gordon et al 2003]. This represents a report of a causal event in the laboratory. An extrapolated claim is that a similar causal relationship will hold strongly enough, or may do soon after further technological development, for the same technique to be used on humans. A further implicit claim is that this may prove cheaper or more

convenient than existing alternative medical techniques.

Why you yawn when other people do

People tend to yawn when they see others yawning. Psychology experiments now show that those who do not follow this pattern also tend to be bad in other tests at putting themselves in other people's shoes [Platek et al 2003]. This is evidence for one explanation of why we yawn when others do – namely that identifying with another's state of mind when they yawn may cause an unconscious impersonation. It would also explain why schizophrenics, who have particular difficulty doing this, rarely catch yawns.

Threat of overpopulation

New census data of people and wildlife suggest that the most important cause of species extinction will be the sheer number of humans [McKee et al 2004]. Or more precisely put, in order to reduce species extinctions it may be more profitable to concentrate on limiting overpopulation rather than changing the habits of those already alive. 'Even if we live as vegetarian saints we'll still be having the same impact on biodiversity', in the words of project leader Jeffrey McKee. Other researchers dispute the conclusions, opining that the causal mechanisms are more complicated than assumed and that as a result behavioural factors are after all as important as population size. Thus the original research is in effect a claim about relative causal strengths, and the subsequent debate concerns just this issue too. (For discussion of the issue of different levels of causal structure, see section 2-9 below.)

Nickel in foliage

The foliage of a particular plant was found to contain unusually high concentrations of nickel. Of course, this is perhaps more a discovery of a new effect itself requiring causal explanation than of a causal relation itself. But understanding better the causal relation underlying it may help in the development of plants specially designed to 'mop up' polluted areas by absorbing excess nickel and other metals – a potentially very useful new causal capacity. (From *New Phytologist.*)

Catalyst in liquids

A new catalyst has been discovered which dissolves in liquid reactants, converts them to liquid product, then precipitates out as a solid ready for re-use. This is a new sequence of causes, useful of course environmentally, especially as the process requires no solvents so waste production is minimal. (From *Nature*.)

A historical discovery

The diet of ancient dolphin-like ichthyosaurs included birds and baby turtles, contrary to previous wisdom. This was inferred from the discovery of the preserved contents of a fossilised ichthyosaur stomach. (From *Proceedings of the Royal Society*.)

The typical pattern here is that new particular effects or causal relations are established, under certain circumstances. Or, alternatively put, a context-specific causal strength is established. Often the question is whether the same causal strength can be manufactured so as to obtain under different, more practical circumstances ('technological development'). Implicit is often the claim that the causal strength thus obtained will be greater than that of rival causes, or else that this cause will have a lower strength than the others with respect to some different, undesirable effect (e.g. cost or pollution) while being equally strong with respect to the desired effect. In one case (the effect of overpopulation), a comparison of causal strengths was explicit. There was also one clearly historical case, following a similar pattern to the others. All of these examples can thus be seen as establishing particular causal strengths of one sort or another. The key point is that in no case is there any dispute about the appropriate causal ontology; rather, the research concerns the value of causal strengths with respect to certain effects from within an ontology already agreed. Note also that all the causal strengths discovered are context-specific in the sense that their extrapolation to different contexts cannot just be blithely assumed. This chimes well both with our earlier insistence that approximate truth can only be defined context-specifically in this way, and in addition with our insistence later that causal strengths can also only be defined context-specifically.

The examples considered so far are, so to speak, pieces of relatively pure research. When we move on to more applied work, we find that the above patterns are confirmed there too. Indeed, it is even more common to see comparisons of causal strengths explicitly becoming the key issue. It can seem that, as it were, the more pure research furnishes the portfolio of causes to consider, and the applied research then does that 'considering' – i.e. the applied research then checks which of the portfolio of causes has the greatest importance to the issue at hand, much as earlier the rise of the tropopause was ascribed primarily to greenhouse gases and ozone concentration rather than to the other possible causes. (Of course, the divisions are nothing like as neatly-drawn as this in reality; the point is just to note a certain general trend.) We shall note just two cases, for illustration.

First applied example: the Hiroshima bomb

A large team of researchers have tried to reconstruct accurately the precise effects of the tragic nuclear blast [Straume et al 2003]. Apart from pure historical interest this also helps current research into the health effects of radiation, since the exact nature of the explosion has hitherto been uncertain. New evidence has included recently discovered detailed contemporary town maps, enabling for the first time researchers to piece together which of the survivors were partially shielded from the blast by buildings for instance, as well as more accurate estimation of previously suspected effects such as the shelter provided by a small hill. Another novelty was the availability of sophisticated computer models of the blast, with better modelling of, for instance, the way radiation travels through air. And another big improvement came from new radiation measurements taken from old lightning rods and guttering. New techniques of chemical analysis also enabled researchers to refine previous estimates of radiation levels, for instance by taking into account how one particular kind of radiation (so-called fast neutrons) can transmute copper into a particular isotope of nickel – the levels of such nickel had never been measured before. Such analyses showed that previous estimates of radiation levels near the blast were much too high, as had been suspected by many, but on the other hand that they were still fairly accurate for the zones in which most survivors were found. Much detailed data already exists on the bomb survivors and their subsequent degrees of

suffering from radiation-related diseases; the uncertainty lay in knowing exactly what dosages of radiation the bomb had originally delivered them. Finally, in developing the new reconstruction researchers have been able to trace the angle and direction of radiation striking buildings and the ground back to their point of origin: the bomb. This has enabled them to conclude both that the exact point of the (mid-air) explosion was slightly different to what was previously thought, and also that the overall power of the bomb was slightly greater.

The new research reduced many uncertainties from previous data, as well as ironing out several apparent contradictions in it. To perform this reconstruction the scientists had to integrate very carefully many different models, for instance those of the blast itself, of the interaction of radiation with buildings, its interaction with people, its interaction with landscapes, of where people actually were at the moment of detonation and how reliably they remembered that, and so forth. All of these involved estimates of causal strengths. These causal strengths, while of course informed by background theory, had to be calibrated very carefully to this specific instance. That is, what mattered were the net causal strengths in this particular circumstance, not their values in laboratory conditions. The improvement of the new over previous research is captured very naturally by our definition of approximate truth. There are many different causal strengths involved in the estimation, for instance the carcinogenic dose each survivor received, and the detailed reconstruction work enables us now to have much more accurate estimates of these strengths – that is, the new work scores a much higher degree of approximate truth. And once again, there is no disagreement between the different pieces of research as to which causal ontology we should be focusing on, rather only disagreement as to what the relative causal strengths actually were.

Second applied example: Sundarbans

Sundarbans world heritage park in southern Bangladesh is the world's largest mangrove forest and indeed largest coastal forest of any sort, and a major habitat for tigers. Rivers and creeks running through it provide a breeding ground for fish and prawns and a

livelihood for up to 300,000 small-scale fishermen. The forests are also important for honey, wood and leaf collectors. Moreover, more than 3 million people depend on the forests to take the brunt of the intense annual cyclones and tidal surges from the Bay of Bengal. However, the forests are now only half the size they once were and in danger of shrinking further. What is the explanation of this? The forests are silting up, trees are dying because not enough fresh water is reaching them, and a raft of economic problems is driving locals to destroy the forest for wood. This much is not disputed; what is more controversial is the underlying causes behind these proximate factors, or rather which of those underlying causes is of greatest strength. We consider here two possible such causes, one the fault of a past Indian government, the other that of a past Bangladeshi one.

First, in the mid-1970s India built the Farraka barrage across the river Ganges 18km from the Bangladeshi border. As a result the flow of the Ganges and its tributaries has been halved for much of the year, and from January to March you can now walk across it. The effect, claims the Bangladeshi minister of water, has been that: 'there has been salinity intrusion, rivers have lost their navigability, the north of Bangladesh is turning into a desert, the Sundarbans are being seriously affected, we have bad water-logging and when the Indians release the water we have bigger floods. We have too much water or too little' (quoted in *The Guardian* 31st July 2003).

Second, in the 1960s thousands of miles of high coastal embankments were built by the Bangladeshis with international money. The idea was to protect the population from cyclones and allow farmers to grow high-yielding crops such as rice. While for a time this appeared to work well, longer-term negative ecological effects have now become apparent. The tides were denied entry to the protected areas, which meant that silt deposited by the rivers was not washed away and so accumulated. And whereas monsoons used to drain away surface salinity and be released through sluice gates, now the rising levels of silt are blocking those sluice gates so that the rainwater cannot be drained away and thus the monsoons lead to flooding. This combination of rising silt levels and flooding has led many to abandon agriculture for prawn farming, especially as

the increased salinity has killed off much of the vegetation. The switch away from labour-intensive agriculture has created great unemployment, in turn leading many to the forest to log for wood, thereby damaging it yet further.

We do not address here the issue of *why* these particular government decisions were made. Although no doubt interesting in themselves, and likely to touch on wider aspects of the world economy, they are different explananda. (See section 2-9 for analysis of interest-relativity.) Rather, we shall concentrate here just on which of these two decisions had the greater impact on the forest's area. The example is interesting for a number of reasons. First, it is clearly politically controversial – the question of which of the two causes has the greater strength is extremely charged, yet it seems clear that there should be an empirical fact of the matter as to the answer. This fact of the matter may not in itself necessarily immediately mandate any moral or political consequences, of course – no doubt other things will also come into play there. But nevertheless, notwithstanding the political controversy, it seems clear that there is no great *ontological* aspect to the disagreements here. All accept and use the same physical concepts of salination, flooding, siltation, deforestation and so on. In philosophy of social science it is sometimes casually asserted that ideological clashes will render empirical arbitration impossible because of just such ontological incommensurabilities, but in this example that does not seem to be the case. (Perhaps, outside academia and in the real world of actual political disputes, such a conclusion is typical.)

A second interesting aspect is the highlighted need to be very clear on our focus of attention. As mentioned, we are not interested here in the background politics behind the original government decisions. Moreover, the causal processes by which the decisions led to deforestation are themselves clearly complex and interesting, and involve, as it were, many sub-explananda. But again, here we are interested not directly in *how* the government decisions led to deforestation, only in *how much* they did so (although see the next paragraph too).

A third interesting aspect is that we must take into account plentiful *causal interaction.*

For example, the impact on the Sundarbans of the Indians building the Farraka barrage was partly through increased flooding. But the degree of increased flooding it led to was likely also influenced by the Bangladeshi coastal embankments. In other words, the causal strength of the Indian action was greater than it would have been otherwise because of the Bangladeshi action. Similar remarks would apply when trying to isolate the causal strength of the Bangladeshi embankments. At first sight, it might seem that such interactive effects render establishing these causal strengths difficult or even impossible. But in fact it turns out that our definition of causal strength can incorporate interactive effects without trouble (section 3-3). (Note that our original example concerning atmospheric chemistry and the height of the tropopause likely also would have involved extensive interactions between the different causal factors.)

So which is more to blame, the Indian barrage or the Bangladeshi embankments? Which would be the truer explanation? We can now frame this very naturally in terms of our definition of approximate truth – there is a true weighting for each of these two causes, and the two putative explanations will differ on how accurately their accounts reproduce it. As the above paragraphs argued, our definition can capture this happily even in a case of political controversy, many explananda, and complicated causal interactions.

To sum up: of course, this section is in no way a systematic survey of all of science. Nor is it a proper detailed case study of any single instance of science. Rather, it is intended to illustrate that the following claims are plausible: that our definition of approximate truth is sufficiently simple that it can be easily applied to real-life scientific work and disputes. That typically such work and disputes do not revolve around questions of ontology, nor do they revolve around questions of vocabulary; rather, they concern attributions of causal strength *within* a generally agreed ontology and vocabulary. That they often concern very context-specific models, or else combinations of models very specific to one particular circumstance, and also that they often involve the need to allow for many interactive effects between the causes in question. All these latter features are well catered for by our definition. Accordingly the suggestion is that, for the great majority of cases, our definition of approximate truth is adequate for the assessment of

science as it is actually practised. And although the few cases that are exceptions to this have been much discussed in the literature and indeed in philosophy of science generally, still it is well to remember that proportionally they are in fact very few in number.

2-7) Relative versus absolute approximate truth

Relativisations

Our definition of causal strength yields results that are fully quantitative (chapter 3). It is not difficult also to design a measure of similarity that is similarly quantitative (see appendix). Can we therefore say that our measure of approximate truth is likewise quantitative? The short answer is 'yes' – but only relativised to its particular causal scheme. There are several problems with interpreting its results to be in any sense indicators of 'absolute' approximate truth.

To begin with, recall the demand from much of philosophy of science that a good notion of approximate truth be able to make sense of general scientific progress. As already noted, our own scheme cannot do this. Or rather, it can only make sense of progress in a way narrowed to a particular context. It can make no sense of the notion of one general theory being more approximately true than another.

Now if this were the only argument against interpreting our quantitative judgments of approximate truth absolutely, we would still be licensed to claim some authority for its context-specific scores. Thus we might claim that a score of 0.9 (say) indicated that a particular model was '90% approximately true' of one particular bit of reality. But unfortunately there are reasons to doubt even then whether any such confident pronouncement could really carry much authority. First, as repeatedly noted, our definition requires prior agreement on what the correct causal ontology is. Therefore its

scores can only ever be relative to a given ontology. Second, the scores are also relative to choice of vocabulary, even once given a particular ontology – the classic Miller point. Now it may be that in practice, as argued at length in section 2-6, disputes in science hardly ever turn on these particular issues. That is, within any particular context we usually *do* have agreement on both ontology and vocabulary. But even if this were true, it would give us confidence regarding the preference for one model over another only in that particular context. That is, it would only give us confidence regarding *relative* judgments of approximate truth, i.e. regarding *qualitative* judgments between two models already agreeing on vocabulary. But to take the *absolute* numbers seriously, we would implicitly be comparing the approximate truth of a given model with the approximate truths of all other models everywhere, and this is a much larger step. In particular, although it is true that competing models indeed typically do agree on vocabulary, it probably is not true that the same applies to all different models in different contexts. Thus we conclude: methodological practicalities indeed license us to give weight to relative, qualitative judgments of approximate truth, but not to ascribing any deeper authority to the absolute scores.

Note again also that all results for approximate truth are relative to focus of interest (section 2-9), as indeed are the very values of causal strengths to choice of context (section 3-2). There is also a final factor against which any results must be relativised – namely, our exact choice of measure of similarity. In particular, if there exists more than one possible such measure, as of course there does, then there exists a danger that judgments of approximate truth will be unsatisfactorily measure-dependent. It may well be that, for any reasonable choice, in practice almost all our rankings between models will not be choice-of-measure-sensitive in this way. Thus, rather as for choice of ontology and vocabulary, so far as most actual scientific disputes are concerned, the choice-of-measure problem turns out to be, as it were, only potential rather than actual. Nevertheless, again similarly to the previous two cases, the issue still casts doubt on the authority of any absolute scores. For while the qualitative rankings between two models may be unlikely to vary with choice of measure, clearly the absolute scores will. On the other hand, in this thesis we shall end up recommending only one measure as our

preferred definition (see appendix).

Note that even our qualitative verdicts will nonetheless still be considerably more detailed than those available from a pure OAT measure. Recall that *any* two models with the correct ontology must be adjudged equal by OAT, which makes it somewhat useless for practical purposes. But our scheme can and does routinely discriminate between correct-ontology models on the basis of the accuracy of the causal strengths they postulate.

Causal interventions

There remains though an important sense in which our definition does still yield useful quantitative information after all, notwithstanding the earlier caveats. That sense is with respect to causal *interventions*. Our definition of causal strength is (roughly speaking – see chapter 3) such that if a cause has a strength of 2 units, say, that means that, given the background conditions prevailing, implementing the cause will lead to 2 units of the specified effect (in the natural units of that effect). Thus our absolute scores for approximate truth will reflect the absolute accuracy of relevant causal strengths, and hence in turn the absolute impact of particular causal interventions. In this sense therefore the absolute scores are indeed meaningful.

This argument assumes that the cause in question is well specified, which means in turn that the reliability of our interventions will be dependent on having a (sufficiently) correct ontology. To see why, imagine that phlogiston was awarded a causal strength of 2 with respect to creating some gas. This would not enable us then to apply phlogiston and create 2 units of that gas, since phlogiston does not exist and so cannot be a cause of anything. When adding 'phlogiston' we would presumably actually be adding something else which we were mistakenly thinking to be phlogiston. This 'something else' would have its own causal strength with respect to the gas, which presumably might differ from the value we had assigned to phlogiston. Thus having an incorrect ontology could leave us vulnerable to miscalibrated interventions, and it is in this way that our causal strengths

and hence ability to intervene accurately are relativised to ontology.

Note though that it would not *necessarily* follow that our interventions went awry even if our ontology were incorrect. The reason is that in a given context our ontological error might have only a small effect – arguably the case when Newtonian interventions are still effective even though we take them to be ontologically mistaken. Or perhaps the error would not be costly at all if the causal strength were calculated for something that was accurately specified but merely mislabelled with respect to our wider ontology. For example, if we incorrectly labelled some actual chemical as 'phlogiston' but calculated an accurate causal strength for it, then an intervention would still be accurate even though we would have misunderstood what was happening ontologically. Hence accurate causal intervention is still possible even with imperfect science. It would of course have been worrying had we *not* found this to be the case, given the long history of effective human interventions before modern science, not to mention similarly effective interventions by animals and babies, and not to mention the presumed ontological fallibility even of our best current models. (See sections 2-12 to 2-14 for more discussion of how to handle models with incorrect ontologies.)

Nevertheless, the accuracy of our interventions is still in general dependent on choice of ontology. However (once given a correct ontology), it is *not* dependent in the same way on choice of vocabulary, and so is not, as it were, subject to the Miller problem. For example, suppose the effect we were concerned with was 'making uncovered grass wet'. Then rain would likely be assigned a high causal strength. But Miller's predicate 'Minnesotan' in the canonical weather example, defined remember as (either hot&rainy or not-hot&not-rainy), would now be assigned a *different* causal strength (its exact value depending on the context-specific relative frequencies of its two disjuncts). That is, each choice of predicate has its own causal strength. Accordingly, supposing fancifully we had the power to control the weather, we would know the effect of creating rainy weather and the (expected) effect of creating Minnesotan weather, and in *either* case would thereby have an accurate idea of how to calibrate our intervention. So a Miller switch from one vocabulary to the other, which so confounds measures of approximate truth,

would not confound the accuracy of our interventions, since the causal strengths on which we would base those interventions are already defined predicate-specifically, so to speak (assuming all along of course that our scores for the relevant causal strengths were themselves accurate). Therefore although our ability to intervene accurately is ontology-dependent, it is not – once we have a correct ontology – vocabulary-dependent. This of course is fortunate, else the effectiveness of our interventions in the world would be language-dependent; but of course they are not, so it is well that our definitions should reflect this.

When would we even want absolute approximate truth?

So how does all this relate to the issue of relative versus absolute approximate truth? The point of it is to demonstrate that the quantitative elements in our approximate truth calculation – namely the assignments of strengths for each individual cause at hand – do have some objective significance. Given the lack of ontological dispute typical of most of actual science they are in fact *typically* objectively significant. Or at least, they are with respect to intervention. My real claim now is this: that it is *only with respect to interventions* that we actually really care about quantitative or absolute results in the first place. By contrast, we are not really greatly concerned about quantitative judgments *of approximate truth*; in this latter case, qualitative is all that matters. It is clear that we are concerned about exact quantitative results when planning causal interventions and, we have just argued, the results that our definition of causal strength provides us are indeed objective and authoritative enough for that purpose. But our concern with approximate truth in science is typically only *relative*. That is, we are typically concerned to know whether one model is nearer the truth *than another*. (Even the classic concentration on general progress in science is concerned with relative judgments, in particular with the notion that successive theories or paradigms are nearer the truth *than their predecessors*. It is rare for philosophers to ask 'how approximately true is Newtonian mechanics in absolute terms?' Rather, the more typical question is: 'is Newtonian mechanics more approximately true than Aristotelean mechanics, but less approximately true than relativistic mechanics?' Or rather, usually it is assumed that the answer to this is 'yes', of

course, and the real demand is for some satisfactory way of representing the pre-existing intuition.)

Therefore the fact that our definition's scores for approximate truth are rather more authoritative when interpreted relatively than absolutely, is in fact no particular weakness. It is important to distinguish between interventions on the one hand and comparison of models on the other. In practice, our approach in this thesis can only really provide (usually) authoritative quantitative advice in the former case, but the point now is that that is the only time we even *need* such quantitative succour.

To sum up: our scheme primarily delivers only qualitative approximate truth verdicts. Even these are relativised to choice of ontology, language and measure of similarity, as well as to focus of interest, but in practice this qualification is rarely onerous. Our scheme does also deliver quantitative approximate truth verdicts too, but the various relativisations make it hard to attach objective weight to the exact scores. Nonetheless, once given a (sufficiently) correct ontology our scheme can also license accurate quantitative interventions, thanks to our definition of causal strength. Happily, this latter context also seems to be the only one in which we really even *need* such quantitative verdicts.

2-8) Approximate truth or approximate explanation?

Our definition deals exclusively with causal descriptions. This makes it well suited to analysing causal *explanations*, and also as we saw to licensing causal *interventions*. But what of two different non-causal descriptions of entities – perhaps our definition is not so well suited to comparing two of those? For example, suppose that one model says the solar system has 10 planets and another that it has 15. It seems natural that the first

model should be deemed nearer the (let us still assume) true figure of 9 than the second, but it is not obvious how to express this in terms of causal strengths. Similar remarks would apply to estimates of physical distance: suppose the true distance between two objects is 5cm, one model claims it is 6cm, and another model that it is actually 8cm. A slightly different example might involve biology: suppose the true animal that passed during the night was a dolphin, but one model claims it was a shark, while another claims it was an elephant. Or suppose the true colour of some plant is red, one model claims it is orange, and another that it is blue. How should approximate truth be thought of in the context of these examples, and is our definition able to meet the challenge? Should our definition really be seen not so much as an account of approximate truth, but more as one merely of approximate (causal) explanation?

To an extent, I think this point is well taken. However, I do not see it as a serious problem, for three reasons. First, because our causal emphasis carries compensating advantages, in particular it enables us to solve the seriousness-of-errors problem. Second, because if necessary the supposed non-causal counterexamples above can be re-expressed in causal terms after all. And third, because in practice most scientific disputes do concern explanations in any case, and so it is a merit of our definition to be much better suited to them than are rival ones. The first point was discussed in chapter 1, and the third one in section 2-6. Here, let us flesh out the second one.

Our own definition can in fact handle the apparently acausal examples given above; indeed, we shall argue that it handles them *better* than other approaches, due to just the seriousness-of-errors considerations already mentioned. Take the biological examples first. It might well be wondered whether claiming that a red plant is orange is a greater or worse error than claiming it is blue. Most other approaches would probably mark the two models as equal for approximate truth since each has made one mistake – unless perhaps it was argued that since orange is nearer red in the electromagnetic spectrum than is blue, somehow therefore the orange model is preferable. The very fact that it is unclear whether or not to give this latter argument weight is itself in fact a reflection of the undesirable vulnerability to such questions of any approach not giving explicit attention

to extra-logical considerations (section 1-3).

What would our own definition say here? The point is that which model is closer to the truth depends crucially on *context* – and it is a major advantage or our definition that it automatically incorporates this through our context-specific definition of causal strength. Thus suppose that our context was the visibility of these plants to particular insects whose vision works only in the low-frequency end of the visible (to us) spectrum, so that they can see red and orange but not blue. With respect to this effect of visibility-to-insects, the error of the orange model in terms of the causal strength will clearly likely be much less than that of the blue model, and hence the orange model will score better for approximate truth. Now consider a different context, say some war game where one side's positions are marked by yellow plants and another by orange plants, all other colours being deemed neutral. In this new case, with respect to the effect (say) of drawing enemy fire, it is now a much more serious error to label the red plant orange than it is to label it blue, and so now – with respect to this new causal strength – it is the blue model that will be more approximately true. Thus our ranking of the two models *should* be varying by context, as with our definition. In still other contexts, of course, it may be that the error of the two models is equal and so they would score the same.

A similar analysis applies to the other examples – in each case, the problem can readily be re-expressed in causal terms and hence our definition applied, thereby also incorporating desirable context-specificity. Thus, in one context it may be that a dolphin is deemed nearer an elephant, because both are mammals and so, for instance, pregnant females will display much reduced speed prior to giving birth, which would mean a greater causal strength with respect to an effect of increased-easiness-to-hunt at such periods. But in a different context, say if the relevant effect were the bumping of a boat, the dolphin might now be deemed much nearer a shark because both are sea creatures. Thus the elephant or shark models could after all each be given a score by our definition, and moreover these scores vary desirably with context. Likewise, there are many causal strengths that vary with distance, and so we would be able to rank the two models of different physical distances. And we could similarly imagine causal strengths that varied

with the number of planets.

This ease of applicability of a causal approach was precisely one of the advantages of our own definition over previous ontological ones (section 1-6). It would be difficult, for instance, to see how to apply the geometrical ontology of chaos theory, or type-hierarchies of natural kinds, to our examples above. On a similar note, Newtonian equations of force are often described as acausal in the sense that there is nothing in the equations themselves that implies (for instance) the time asymmetry characteristic of causation. But despite this, we have seen how straightforward it is to analyse Newtonian examples in terms of causal strengths, perhaps because causality enters Newtonian models naturally once they are applied to specific contexts of intervention or explanation. Note also that almost all *policy* issues involve interventions of one sort or another, and so our definition of approximate truth will automatically be applicable to all of those too.

Finally, the relativity of our definition to choice of ontology and vocabulary (previous section) sits well with describing it as an account of approximate explanation rather than approximate truth. Recall also our earlier discussion (section 1-4) about adopting a methodological rather than metaphysical conception of the issue, which also sits well with a focus on explanation.

In summary then, in a manner of speaking it is indeed true that our definition reduces approximate truth to approximate causal explanation. However, this does not in practice limit its applicability, and on the positive side confers important advantages.

2-9) Interest-relativity

Does our scheme just reduce to EAT?

One objection might run as follows: our scheme focuses on the strengths assigned to the different causes of an effect, but ultimately those strengths are in turn defined just by the individual quantities of effect these causes each lead to, in other words by a set of empirical variables. That is, although our definition admittedly does not boil down just to the level of headline composite effect as EAT does, still it is only going one level down, as it were – its results just boil down to the levels of effect due to the underlying causes instead. The focus is purely on the value of particular empirical variables rather than anything ontological. Therefore is our definition really distinct from EAT?

The answer is 'yes' because even shifting the empirical focus just one level down is sufficient to circumvent EAT's main difficulties. This is most obvious when comparing two competing explanations. If, as is frequently the case, both explanations yield the same (actual) final effect, then EAT has no way of separating them. But what matters is whether the explanations have identified, and given the right weightings to, correct causes of the effect. To be sure, the weightings on these causes are indeed defined by us empirically, but typically this is no longer a problem since the competing explanations will be offering different weightings – this is why they are *competing*. EAT is helpless given empirical equivalence at the level of the final composite effect, which unfortunately is a common occurrence. Our definition would only be helpless if there was similar equivalence for every postulated underlying cause as well, which is extremely uncommon – or rather, if there was such equivalence all the way down, then it is not clear in what sense the two competing explanations would even be distinct. Similar arguments apply to the case of fluke correlations. Once again, going one level down is in practice enough to distinguish – unlike EAT – between genuine explanations and ones that are correct merely flukily. In this way, our approach is indeed a significant advance on EAT.

Note here also two other points from elsewhere in this chapter. First, when – as is usually the case in science – there is no ontological controversy, the particular weaknesses of EAT remedied above by our scheme are the only major ones *needing* to be remedied (section 2-4). And second, when there *is* ontological controversy, EAT is

arguably equally as vulnerable as our own definition to the difficulties that follow (section 2-12). Thus our scheme is not adding to EAT any weaknesses not already present. In combination, these points imply that we cannot do worse under our approach than under EAT, and in many cases will do strictly better.

Which level down?

On a couple of occasions in the previous section, we mentioned going 'one level down'. Yet this seems an unsatisfactory phrase – surely there is nothing in principle to stop us going *any* number of levels down, and there is left unresolved the issue of just which 'level' of underlying causes we should focus on exactly. To see this, take now a very different and more tragic example – the decision to drop the atomic bomb on Hiroshima in 1945. Suppose we have three competing causal explanations. The first says that since it was fine weather on that day, unlike on the preceding days, so the bombing mission was enabled to proceed. The second states that the USA dropped the bomb because it wanted to induce a quick Japanese surrender. And the third says that the development of human scientific knowledge had reached a level that made use of such a bomb inevitable sooner or later. Now these three models are not mutually contradictory, so therefore it must be possible that *all three* of them are simultaneously correct. Yet if there is apparently only one explanandum – 'the decision to drop the bomb' – it would seem to follow that there is only one target set of true causal strengths, so how could it possibly be that three very different models all simultaneously score well?

Consider the causal sequence that led up to the bomb. It ends with the physical detonation and then, working backwards, to its dropping from the plane, to the decision to undertake the bombing mission, then in turn to the course of the war before that, and so on back into history. And of course there exists not just one simple linear chain, so much as a whole web. Thus the decision to undertake the bombing mission required the joint occurrence of several factors, all of which were independently necessary – the state of the war, the fine weather that day, the technical development of the bomb and so forth. Many different chains lead back, and overall the antecedent causal web is complicated.

The key point is that each of the three models is an attempt to capture a *different part* of this causal web. It is perfectly possible that each could give a fine account of their own segment of it while simultaneously offering no account at all of the others' segments. That is, they might all focus on their own particular areas within the web and not even attempt 'complete' explanations of their subject. This is arguably a common state of affairs in social science. So a fundamental part of any account of context-specific approximate truth must be a definition of just which segment of the causal web we are interested in. For example, the 'fine weather' model is perhaps a good explanation of why the bomb was dropped on the precise day that it was at the precise place that it was. But it is not a good explanation of why it was dropped that year rather than another, or on that country rather than another. So it is vital to distinguish the many subtly different focuses of interest all consistent with the same explanandum of 'the decision to drop the bomb'. In other words, for a full treatment of approximate truth we are going to need to incorporate pragmatics as well as semantics.

Of course, these are familiar points. (Even [Miller 1994] accepts that judgments of approximate truth must vary depending on which question a theory is taken to be trying to answer.) The question is: can we deal with them fully in our scheme? I think the answer is 'yes'.

Choice of causal strength

Let E be the effect of deciding to drop the Hiroshima bomb, and specify two causes: let C be the state of the weather, noting that it was sunny today but cloudy yesterday. And let D be the state of technology, noting that this incorporated nuclear bombs this year but no such bombs 100 years ago.

To start with, suppose we are concerned with why the decision was made today rather than yesterday. In particular, we are interested in comparing the causal strengths (with respect to this effect) of the weather and of the state of technology. In the case of the

weather, we note whether the bomb would have been dropped today if it had been yesterday's cloudy weather (no), and whether it was dropped with today's actual sunny weather (yes). We then *compare* the two answers, and ask whether the change from yesterday's to today's weather made a difference? Clearly here it did, and therefore we adjudge that the weather has a high causal strength with respect to the short-run explanandum. A key point is that these causal strengths must be calculated holding background conditions fixed, to avoid confounding the effects of the cause of interest with those of other factors that may be varying (section 3-2).

Turn now to an analogous calculation for the long-term cause D, the state of nuclear technology. Comparing that state today and yesterday, it is clearly the same. Therefore, in contrast to the weather case, the switch between yesterday's and today's nuclear technology would have made *no* difference. That is, holding constant the background condition of today's sunny weather, the bomb would have been dropped just as much given yesterday's nuclear technology as given today's. Hence with respect to this short-term explanandum the level of technology has zero causal strength.

What about the second, longer-term explanandum – which factor, weather or technology, has the higher causal strength with respect to the effect of the bomb being dropped this year rather than 100 years ago? Again, we compare the effects of changing the relevant cause while holding others fixed. Thus, given today's state of technology, the weather of 100 years ago would presumably have the same significance as the weather of today – namely sometimes sunny and therefore allowing the dropping of the bomb at some point during the year. Thus switching between the two states of the weather now makes no difference and so has zero causal strength. (Note that if the weather *had* so altered, for instance if 100 years ago it was always so cloudy that not even one bombing mission could have been attempted, then the change in climate since then might indeed legitimately be seen as causally important even with respect to the long-term explanandum.) By contrast, holding the weather constant, the change from the technology level of 100 years ago to that of today *would* be significant. Thus the results are, as desired, the reverse of those for the short-term explanandum – now it is the level

of technology, not the weather, which has the high causal strength.

The key underlying suggestion here is this: that interest-relativity boils down to the *choice* of causal strengths on which our definition of approximate truth should operate. In particular, in this example the short-term explanandum implied the calculation of one particular set of causal strengths, while the long-term explanandum implied the calculation of a different set. Moreover, technically the causal strengths we were defining above were *relative* strengths, defined as the difference of two absolute ones (see section 3-2 for more on these). So while Nature, as it were, yields the same objective causal strengths regardless of our subjective interests, still we can choose to focus on any of an infinite number of different *relative* strengths according to our fancy. These relative strengths, once specified, would in turn be determined objectively, as were the ones above for instance. Thus we have two dimensions of freedom, as it were: we can select between different absolute strengths, or choose any particular relative strength. Summing up, interest-relativity boils down to the subjective choice of which causal strengths to focus on. But once that choice is made, the actual values of those strengths (relative or absolute) are determined objectively and hence so would be the scores for approximate truth calculated from those values. (For more on subjective and objective, see section 2-10.)

(Again, refer to chapter 3 for a more rigorous discussion of how to define causal strengths. Some of the subtleties of picking out different absolute and relative causal strengths go beyond the scope of this thesis. For a more fundamental treatment, see [Northcott 2004].)

Completeness versus accuracy again

Suppose we are interested in the causes of lung cancer and we have two models of this. The first cites smoking and gives a very accurate estimate of smoking's causal strength. However, it makes no mention of asbestos. The second model mentions both smoking and asbestos and gives estimates for both their causal strengths, although these estimates

are not as accurate as the first model's for smoking. Which of the two models should be preferred? On the one hand, the first one is, as far as it goes, the more accurate of the two. On the other hand, the second one has captured more of the factors at play and so although less accurate is also more complete. It seems that there are two distinct desiderata here, accuracy and comprehensiveness, and our scheme does not tell us how to balance them.

This issue was of course precisely one of the reasons why we criticised attempts to define approximate truths of general theories, arguing instead for our context-specific model-centred approach. But can we really avoid it now even in our scheme? I think that our treatment of interest-relativity shows how we can. We may be *interested* in one of two different things – either the strengths of all the causes of lung cancer, or else the causal strength just of smoking only. If the former, then it may well be (depending on the exact accuracies and choice of definition of similarity) that the less accurate but more complete second model is preferred to the more accurate but narrower first one. If some model only gets part of the picture, so to speak, then even if it gets that particular part very accurately there is no offence against the meaning of approximate truth to judge it only a poor approximation of the picture *as a whole*. (Imagine if the first model had furnished an accurate causal strength just of asbestos; given the relative lack of statistical importance of that cause of lung cancer, such a model would hardly have qualified as an approximately true picture of the overall causation of lung cancer.) Thus the smoking-only model may make a true causal claim, and may even calibrate that claim exactly correctly, but it does not follow that it should therefore always score best for approximate truth.

If, however, we were interested instead in the causal strength only of smoking, then now all that would matter would be the accuracy of the estimation of that particular variable only, and so by assumption our narrow but accurate first model would score the better. Because of the now narrowed range of interest, the only part of the second model that would even be relevant would be its estimation of smoking's causal strength, and from the point of view of this particular calculation of approximate truth the model's furnishing

also of a causal strength for asbestos would no longer garner it any credit.

Thus which of the two models would score best for approximate truth *varies*, on our definition, according to what we are trying to explain. Mere accuracy is not itself enough; we need also *relevant* accuracy. This seems to me a highly desirable feature for any definition of approximate truth. A model may be useful in some contexts but not in others. In terms of our earlier evocation of vectors in a hypothetical cause-space, we can imagine that interest-relativity means that the target 'truth-vector' can now vary. As it were, even though there exists only one physical world, still our definition is able to give due accommodation to the stupendous variety of human scientific investigations of it.

A corollary of this is that it is possible for many different models of the same event each to score highly for approximate truth – with respect to different focuses of interest. The decision to drop the Hiroshima bomb had many different causes; interest-relativity specifies in which particular ones we are interested. Thus a 'complete' account of an event, in the sense of one covering all that event's causal antecedents, is not always necessary for achieving a high score for approximate truth; and neither, if too narrow, is a merely accurate account sufficient. (Similarly, a satisfactory causal explanation need not necessarily cite the causes behind the immediate causes, the causes behind those ones, and so on *ad infinitum* back to some hypothetical first cause. Rather, interest-relativity would specify exactly which bit of the sequential chain is of interest, and a good model of this bit of the chain could then achieve a high score without any need to produce a first cause.)

Interest-relativity and the literature

Of course, in one sense allowance for interest-relativity must presumably be obvious and uncontroversial. Nevertheless, I think there is value in highlighting the issue so explicitly. It has usually been ignored, although an exception is the characteristically thorough Niiniluoto. He essentially agrees with the thrust of this section, when acknowledging that: 'the choice of the target ... is not a matter of logic, but depends on

our cognitive interests' [Niiniluoto 1998, p14]. In other words, an element of interest-relativity is unavoidable. (And therefore so is an element of subjectivity, although like us in the next subsection, Niiniluoto is keen to emphasise that once relativised to some particular interest a definition of similarity can still be objective thereafter.)

As noted, I think our treatment of interest-relativity also does, for any one context, just the job that Popper originally earmarked generally for comprehensiveness. A general theory is useful only if it has wide range as well as accuracy, which was of course Popper's starting point. But for any specific context (which was not Popper's focus), this just boils down to saying that the general theory is useful if its range happens to include what we are interested in here and now. As it were, general range is irrelevant, what matters in any particular case is only whether that range covers that case. The gain from incorporating Popper's notion of content boils down to exactly the same gain we achieve from our incorporation of interest-relativity – namely, that a theory or model does indeed cover the particular factors of interest to us. Throughout, our whole approach to approximate truth has concentrated on Popper's accuracy rather than comprehensiveness. In a sense, the motivation for including Popper's comprehensiveness at all has now been satisfied instead by our incorporation here of interest-relativity. As it were, once – but only once – we have specified the question, we are left free to focus purely on accuracy.

Finally, is it perhaps a mistake to try to analyse at all something so seemingly vague and subjective as interest-relativity? Yet similar difficulties have not stopped other philosophers profitably analysing, for instance, subjective degrees of belief. At heart, our scheme is conceptual, and I do not see a problem with conceiving of *some* specification of relevant causal strengths as being the reflection of our subjective interest. Moreover, in many cases that focus of interest may be pretty uncontroversial. (Vagueness is separately discussed in more detail in section 2-11.) Consequently, although I accept that approximate truth is indeed inextricably bound up with interest-relativity, I do not see any necessity to abandon ambitions for a general definition of it just on this account – contrary to the pessimism of several authors, for instance [Smith 1998] and [Psillos 1999].

Divergence between theory and consequences

One particular conundrum is highly troublesome for most of the mainstream literature, and turns out to be relevant now. This is the fact that sometimes the consequences of an approximately true theory may not themselves necessarily be approximately true. For example, the 'butterfly effect' in dynamical systems means that a very good approximation of the underlying theoretical equations may yield predictions that quickly become very bad approximations of the actual outcomes, even assuming perfect knowledge of the initial conditions. As Miller points out, if it is these *outcomes* that we are interested in, then approximately true knowledge of the theoretical generating equations no longer seems particularly desirable or important [Miller 1994, pp198-202]. That is, there seems to be a divergence between achieving approximate truth with respect to one part of the problem and achieving it with respect to another part. Should we put more weight on the underlying equations or the final predictions?

While any approach emphasising the approximate truth of general theories has no real solution to this dilemma, I think that our treatment of interest-relativity does provide one for us. The key is that we have only even *defined* approximate truth with respect to particular outcomes, not with respect to general theories. Thus in this example, we need feel no obligation to pronounce on the 'overall' success of some theory that got the equations almost right but the outcomes wildly wrong. Rather, we would first specify whether we were *interested* primarily in the equations or in the outcomes, or – alternatively put – in which particular outcomes. (See also our discussion of [Smith 1998] in section 1-6, and our discussion of time in section 2-11 below.) Thus there is no paradox whereby it seems that the same theory must be adjudged simultaneously both approximately true and not approximately true. Rather, it may be adjudged approximately true with respect to some focuses of interest but not with respect to others. By definition, there *is* no general calculation to be made. (The point that a single theory is likely to be more accurate in some of its descriptions than in others, recalls our original interest (section 1-1) in making sense of the variable usefulness of economic theories.)

2-10) Subjective and objective

We have seen that, although the choice of which causal strengths are relevant is inevitably subjective, once we have specified them they are then determined objectively. Nevertheless, this does still leave a subjective element in our definition. Should we be alarmed? No – I think that this amount of subjectivity is unavoidable anyway, and that we are better off for analysing it explicitly.

One worry might be that this could boil down to '*all* models are OK once some sufficiently sympathetic specification of relevant causal strengths has been found for them'. But some models may not be giving an accurate description of *any* causal strengths. And in practice, often rival models do agree on what the target causal strengths are, merely disagreeing as to their evaluations of them.

Take our Hiroshima example again, agreeing to focus now on what motivated the decision to use the bomb rather than more conventional military means. This is actually a lively historical controversy, with several competing theories as to what was the *main* cause of the US action. Thus it is argued that America wanted to avoid the need for a bloody invasion of mainland Japan, and so dropping the bomb on balance actually saved lives. Or maybe merely saved American servicemen's lives. Alternatively, perhaps America wanted to end the war quickly in order to pre-empt further Soviet territorial advance in Manchuria. Or it may have wanted to intimidate other governments in advance of negotiations for a post-war settlement. Although all of these motives might have been present, clearly not all could simultaneously be the *main* one. The key point is that the controversy concerns the relative strengths of these causes; no one disputes that they are indeed the causes that are relevant. As it were, the controversy is over which

model most accurately captures the relevant part of the causal web, *not* over what that relevant part of the web is in the first place. But the subjective element only concerns this latter aspect. The former aspect – the actual focus of debate – is presumably determined objectively by the actual state of the world (at that point in time). Accordingly, in all important respects the determination of our scores for approximate truth is satisfactorily objective.

Perhaps surprisingly, we can enlist Popper himself as a supporter of roughly this point. Miller reports Popper's opinion that: 'We can relativize any theory, in a perfectly objective manner, to its historical setting, and thus we may be able to construct an objective ... mode of comparison of two theories' [Miller 1975, p186]. (In this way, Popper speculated, we might after all be able to overcome Miller's objection, by making our judgments of approximate truth relative only to the parameters of the particular problem a theory was designed to solve. But as Miller quickly argues, when it comes to evaluating whole theories, it is unclear why we should in this way conveniently focus our attention exclusively on some rather than other problems. An obvious response in turn to Miller's point is to analyse approximate truth not at the level of general theories, but rather – like our own scheme – at that of context-specific models.)

When scientific debate really does centre around which causal strengths we should be trying to capture in the first place, then it means that the subjective element has become important *in the actual science itself.* Suppose for instance that our question is: 'was the aristocracy a good influence on eighteenth century English society?' One answer might be 'yes, it helped preserve social order without which society could not have functioned'. A second might be 'no, it protected its own and repressed the rest'. Now here the first modeller understands by 'good influence' one particular phenomenon, namely preservation of social stability. The second modeller understands by 'good influence' a quite different thing, namely ensuring social opportunity for all. Consequently it is possible that both modellers do indeed describe causal strengths correctly – but *different* causal strengths. In that case, the controversy is not then about accuracy of description but rather about agreeing what the problem is in the first place. In effect the two

modellers have specified different causal strengths as being the relevant ones, and the disagreement now really lies in that initial choice rather than in the subsequent historical analysis. So this would be an instance of the subjective element indeed causing trouble, but the point now is that that trouble would be stemming from the science itself rather than from our way of defining approximate truth. Indeed, it is surely to the credit of our scheme if it captures this prior confusion by considering subjective aspects explicitly.

Some element of subjectivity is unavoidable, so it is not a question of trying misguidedly to squeeze all subjectivity out of our definition of approximate truth. Rather it is (or should be) a question of trying to incorporate it in the best possible way. When the subjective aspect is unimportant in a controversy our ranking of models by approximate truth turns only on objective factors, but when the subjective aspect is dominating the actual science then it also dominates our scheme in its turn. We only reflect proportionately the significance of subjectivity in the underlying science itself. No extra subjectivity is being unduly imported – and none unduly denied either.

2-11) Other discussions

Vagueness

In this subsection we are concerned with vagueness – not lack – of reference. That is, we are concerned with the case where it is not clear to which ontological entity a scientist is referring, rather than the case where the scientist is referring to nothing at all. Thus 'the Establishment' might be a vague reference but does seem to exist in some way, whereas 'phlogiston' presumably simply does not exist all and so does not have any reference. The latter case is not considered until the following sections.

Although descriptive vagueness is common in natural-language descriptions of the world

and hence in large parts (perhaps most) of science, still this need not mean that all such science is useless. On the contrary, the cost of vagueness may very often (perhaps usually) be small enough still to permit sufficient clarity for the purpose at hand. Useful science is still possible despite ontologically imprecise description; as it were, even without his glasses a short-sighted man can still see something.

The important point is that where our terms are sufficiently precise for useful science, so they will be sufficiently precise to license judgments of approximate truth. This is because under our scheme the latter depend only on obtaining sufficiently precise values for causal strengths, which in turn should in general presumably be neither more nor less reliable than the rest of science. Any vague term has more than one possible reference, and so in the terms of our hypothetical vector space of causes this would presumably cash out as meaning we needed to consider more than one possible target point and model point. But if all such target-model pairs generate high similarity scores in one real-world case, for instance, and only low scores in another, then despite the vagueness we would still be justified in claiming that a greater degree of approximate truth had been achieved in that first case.

It is also evident that the impact of a term's vagueness on our ability to furnish authoritative scores for approximate truth will vary by context. For example, the term 'red' clearly refers at best to a range of the colour spectrum rather than to any specific point on it. This vagueness will sometimes matter and sometimes not. Thus the *exact* frequency of the red supposed to have been invoked may be significant if discussing the manufacture of a laser, but insignificant if describing a traffic light. Similar remarks apply to other pernickety cases – was the object really *exactly* 5cm long or in fact merely somewhere in the range between 4.9 and 5.1 cm? Was it really *exactly* circular? And so on. The fact that the cost of vagueness varies by context in this way also provides another reason why approximate truth is best worked out context-specifically.

Of course, the issue of vagueness has long been addressed by logicians, for instance partly inspiring Zadeh's apparatus of fuzzy logic, but unfortunately this work does not

seem to help us here particularly. There still appears to be some variation of opinion on the best way to proceed. An obvious initial possibility is simply to introduce a scheme of multi-valued logic, that is to say to allow new truth values in addition to True and False, perhaps a third one 'Indeterminate' for instance, or perhaps an infinity of new values on some continuum between True and False. But none of this is easily applied to the approximate truth of theories or models in science, or to the concerns of section 1-3 for instance. (It is also unclear that we should wish to rule out, like fuzzy logic does, the possibility of a model's distance from the truth being exactly zero.)

A big problem for any new logics, such as fuzzy logic, incorporating such ideas is that standard operations will no longer preserve truth values. One response, the so-called 'super-truth' approach, is to say that for any vague predicate there exists a number of possible 'precisifications' [Fine 1975], or clarifications. Each of these individual precisifications is itself true or false. Roughly speaking, we then pronounce a vague sentence true if and only if it is true for *all* of these ways of making it completely precise, otherwise we pronounce it false. The precisifications themselves are taken as primitives. But notwithstanding its logical merits, for our purposes such an approach does not help much (nor of course was it particularly designed to). The concentration on whether any particular vague statement should be classified as true or false is too crudely put for our concerns – as it were, we are more interested instead in tracing degrees of *approximate* truth within the 'false' category. Similar remarks apply to other work within this literature too.

Time

Some problems will explicitly incorporate time as an important factor. Yet our scheme consists essentially of comparing two (weighted) *static* snapshots of the world for similarity, which would seem to make incorporation of time difficult. The solution may be to treat time in the same way as we have throughout already implicitly been treating space. Thus, for instance, the sun's gravity has a much stronger causal influence on Earth than that of another star, simply because the sun is closer. The causal strength assigned

to the sun's gravity is therefore correspondingly more, and by this means we have implicitly incorporated spatial distance. If temporal separation is a causal variable, we can incorporate that similarly. For example, the causal strength of a snowfall with respect to disrupting traffic might be very high initially, but as time passes become less and less as the snow melts away. Thus the value of this causal strength in effect incorporates the passage of time automatically.

It might seem that the following kind of case would present a problem for this strategy though: suppose we have two dynamic theories that postulate particular trajectories through time for some variable. Presumably there is some true trajectory for this variable. How would our scheme enable us to compare for similarity the postulated trajectories with the true trajectory? ([Smith 1998] is motivated by exactly this kind of example.) The answer is that it would not, but also that it need not. Recall our discussion (section 2-9) of a similar example, and in particular of the case when a theory is almost correct with respect to the underlying mathematical equations but is very wrong with respect to the actual outcome later on. Such a theory will have many different values for approximate truth, depending on exactly when we measure it. To attempt some 'general' evaluation of approximate truth would be to attempt the mistake of evaluating general theories rather than context-specific models. Here, by 'model' we mean postulated values for the variables at any particular time. Thus we would not compare for similarity trajectories as a whole; rather, we would only compare for similarity the values ascribed to the variable by those trajectories at any particular moment. In this way, and interpreting or re-expressing the variable at hand as a causal strength, our definition of approximate truth could proceed unproblematically.

Omitted causes

Frequently, it will be the case that a model ignores some of the causes that are in fact pertinent, indeed we have seen this in several of our examples already. Often, it will be natural in such cases to assume that the model assigns a weighting of zero to the causes it ignores. For instance, our Earth-model ignored the gravity from the mountain, which

amounts to it having assigned zero causal strength to the mountain's gravity.

But sometimes it will be better to assume that the model is implicitly assigning a *non-*zero strength to ignored causes. For example, a model of the flow of a glacier may implicitly assume some constant 'normal' air temperature, ignoring small variations. This hardly assumes 'zero' influence from air temperature, since the influence of the temperature on the mechanics of the glacier would be very different at –273 °C than at 5 °C, or 200 °C. In general, typically any model is likely to be carrying plenty of implicit background assumptions concerning ambient temperature and pressure, strength of gravitational and electrical fields, and so on. All these are likely to correspond to non-zero weightings on particular causes. More generally, any model typically assigns its ignored causes some *implicit functional form*. A famous example is the assumption by pre-relativistic cosmology that space-time possessed a Euclidean geometry. (The point resembles our need when defining a cause's strength to set a 'neutral level' not always equal to zero – section 3-2.)

In practice, what matters of course is whether the omitted cause is salient to the problem at hand, or – to use our terminology – whether the omitted cause had a high causal strength. Obviously, models are usually designed to capture precisely what seem to be the most salient features of a problem. Perhaps a more common difficulty is the systematic blind spots, so to speak, that a general theoretical approach might suffer from. In particular, because the implicit functional form assigned to the omitted cause will not have been explicitly considered it is likely in such cases to be unresponsive to new data, making the cost of the omitted cause likely to rise as the theory is applied to new areas. For example, Newtonian theory ignores relativistic effects. For many problems this does not matter significantly, and neither does the exact form of the implied error. But it does carry the latent weakness that Newtonian theory itself gives us no way of knowing in advance in which problems those ignored causes will become significant after all. So in the case of, for instance, the precession of the perihelion of Mercury's orbit, Newtonian theory gave us no way of adjusting our assumptions appropriately.

In plainer language, this is the familiar observation that models often go wrong when applied to new domains because in these new domains the important factors may turn out to be just the ones that the old model has not catered for. This point is particularly pertinent to some arguably overambitious applications of economic theory, which was one of my original motivations for starting this thesis (section 1-1). The important thing here is that our scheme for approximate truth caters for this, and via its system of context-specific weighting functions does so to precisely the appropriate extent.

2-12) When ontologies differ

When does it matter?

We have emphasised how in practice most scientific disputes do not concern choice of ontology or vocabulary. However, occasionally they do, and in any case the issue has been a major focus of past philosophical interest. So even though it may be of, as it were, only second-order importance methodologically, still the topic needs to be addressed. Note that, to my knowledge, no previous work in the literature has discussed this issue from the same angle as shall we. As a result, I believe the next three sections to be original, in that it is impossible to know just what previous opinion would have made of some of the following examples and questions.

The first thing to note is that our own definition assumes prior agreement with regard to both ontology and vocabulary. If there is no such agreement, therefore, then our definition is not directly applicable. Moreover, it in fact seems to need to assume something stronger than that: not only must the ontology and vocabulary be agreed on, but they must pick out real causes in the world, that is they must be true. For example, suppose two competing models talk about the causal strengths of phlogiston or of dragons. The problem is that our definition requires us to compare those causal strengths

with the true causal strengths but that it is hard to make sense of the latter here. What could be the true causal strengths of fictional entities like phlogiston and dragons?

The conditions for our definition to be applicable would therefore seem to be onerous indeed – we need prior agreement not only on both ontology and vocabulary, but also on a *true* ontology and vocabulary. Fortunately, however, in practice these conditions can be considerably softened. Take vocabulary first. Clearly, to insist on prior agreement on literally the same vocabulary would be unreasonable. Suppose one model talks about the forces on the 'rock', a second the forces on the 'boulder', a third the forces on the 'obstruction' and a fourth the forces on the 'target'. In context, it may be clear that these all refer to the same physical object. In the example from section 2-6, for instance, we talked about 'the Sundarbans' and 'the forest' similarly equivalently. All that is needed is sameness of reference. Moreover, if reference is not identical it its true we lay ourselves open to potential Miller-type problems, still in practice usually no such problems will actually arise even then. For instance, suppose one commentator speaks of 'the allies' advance in Iraq' and another of 'the Americans' advance in Iraq'. These two phrases do not have (exactly) the same reference. The causes they pick out will therefore likely have slightly different strengths. Suppose the claim was made that the allies' advance caused the Iraqis to flee more than did the hot weather. This claim will likely be equally as true (or false) as the same claim with respect to the advance of the Americans rather than the allies. The causal strengths of either will be similar here, and in particular much greater than the causal strength of the hot weather. Essentially, in practice what matters when assessing approximate truth is the size of a causal strength, and if small differences in reference lead only to small differences in causal strength then many qualitative judgments of approximate truth will be correspondingly unaffected. And recall (section 2-7), we should not in any case over-interpret the *exact* quantitative scores for approximate truth that our definition delivers.

Remarks similar in spirit can be made regarding ontology. Suppose one model says that kicking a football moves it further than does throwing it, while a second model claims the opposite. Each model will assign some causal strength to each of the factors, and we can

compare them to the true causal strengths (for some particular context) to see which model is (more nearly) right, i.e. to see which is more approximately true. This is a perfectly mundane example, and exactly the kind we would want our definition of approximate truth to be applicable to. But look a little more closely and we would likely see (very often) that our models were Newtonian and hence arguably, in the light of relativity theory, in fact ontologically mistaken due to their assumptions of constant masses, Euclidean absolute space and so forth. Are we therefore disallowed from speaking of the 'true causal strengths' for kicking and throwing here? Strictly speaking, perhaps yes indeed we are. When we speak of a kick in Newtonian terms, we may need to 'interpret' this as being a particular cause in the terms of some true ontology. The point here is that the objective authority of the true causal strengths we assign to the Newtonian causes will be entirely a function of how much objective authority we accord to the 'interpretation' between the true causal ontology and the erroneous Newtonian one.

In the particular case of kicking footballs, it seems that we normally do look upon such interpretations favourably. It follows that, in practice, our definition does not demand absolutely the right ontology in order to be able to proceed. Rather, what is required is a *sufficiently* correct ontology (and vocabulary) for 'true causal strengths' to be ascribed with sufficient objective authority. In particular, what is required is that actual scientific disputes do not concern what the true causes are or whether true strengths for them exist, but rather agree (positively) on those things and concern instead whether or not a model has got an accurate estimate of these strengths. So long as this condition holds, then so long is our definition of approximate truth applicable. Fortunately, as already argued at length, in practice this condition nearly always does hold. In sections 2-13 and 2-14 we shall discuss in more detail the tangled issue of what might govern just when such inter-ontological 'translations' are or are not admissible. (From now on, we shall use the word 'translation' even though what we have in mind is sameness of reference rather than the more usual sameness of meaning.) In particular, the issue can appear more fraught in the unusual cases when two models differ in their ontological commitments.

Note that this is an issue implicitly facing all suggested definitions of approximate truth,

not just our one. Measures of logical distance such as Niiniluoto's and Oddie's, for instance, assume from the beginning that the truth can be represented in the particular language concerned, or else concede that their verisimilitude measures are relativised just to that language's version of the truth. How to understand the approximate truth of a theory expressed in a different language (or ontology) to the truth itself, or how to compare two theories of different ontologies, are nowhere addressed satisfactorily. Similar remarks apply to all other previous approaches too.

The four cases

So far, we have seen that our own definition can apply smoothly when there are no controversies over ontology or vocabulary, but that otherwise the situation becomes more complicated, dependent on whether or not translation is deemed acceptable. Recall now our original notions of OAT and EAT. Perhaps they can be of help in these new awkward situations when ontological differences between models crop up? It is useful to proceed systematically through the full inventory, so to speak, of possible cases. Along the way, this will also prove helpful for elucidating our approach generally.

For each case, we shall consider the task of comparing for approximate truth two different models. Sometimes they will differ ontologically, or empirically, or in their vocabularies; sometimes they will not. Let us explore all the possible combinations, examining each time the prospects for our own definition and also those for OAT and EAT. (A note on the upcoming terminology: by 'ontologically equal' we shall mean here that the two competing models cite the same ontology, by 'vocabularies equal' that they employ the same vocabulary, and by 'empirically equal' we shall mean that they are empirically equivalent with respect to the explanandum variable. The opposite of 'equal' here will be denoted by 'separated'.)

Case 1) Ontologically equal and empirically equal.

1-a) Vocabularies equal. Suppose our two models both have the same ontology. If that ontology is the true one then we are in our ideal case and our scheme works as desired. If

that ontology is not the true one, however, then our scheme can only work if there is an acceptable translation into the true one (as just discussed above). What of OAT and EAT? As noted before, in these circumstances neither will be able to express a preference for one model over the other, therefore being forced to declare the two models equal. Assuming translation is possible, our scheme is therefore much the best here. As repeatedly argued, since such circumstances are much the most common in actual science, this result is the main motivation for adopting our scheme.

If translation were not possible, then none of OAT, EAT or our scheme could express a preference. Other definitions, such as the mainstream logical-similarity and structuralist ones, could – but only if they were prepared to take seriously the language of the models. This would certainly be formally possible but it is hard to imagine why we should want to do this. Remember, two such models would be in an ontology and vocabulary not only wrong, but so wrong as to defy translation into anything correct. Perhaps they would be referring to dragons and angels to explain some physical phenomenon, one citing two dragons and one angel, the other three dragons and four angels. Which of these models is more approximately true? I do not see it as an important desideratum of a definition of approximate truth that it be able to adjudicate cases such as these.

1-b) Vocabulary separated – the Miller problem. If the two models use different vocabularies, then we require translatability between the two. If they each have the same ontology, as here, this should always be possible. However, there would still be the spectre of potential Miller-reversals. For instance, in the canonical weather example one model might be expressed in the hot-rainy-windy language and the other in the hot-minnesotan-arizonan one, while each sharing the same ontology. To compare them for approximate truth, we would have either to convert them both into the former language or both into the latter one. This choice, Miller shows, can have a decisive influence on our approximate truth rankings. In general, therefore, our scheme will only work smoothly if the translation between vocabularies does not result in a Miller-reversal. But as we saw (chapter 1) this will likely be true in almost all actual cases, given the unintuitive nature of the predicates needed for Miller-reversals. I know of no case of a Miller-reversal from

actual scientific practice. Notwithstanding [Miller 1994]'s claims, in the thought-examples that do exhibit Miller-reversals a preference between vocabularies is always immediately suggested, and so there are no cases where it actually does seem the most appropriate option just to accept that no judgment can be made.

Case 2) Ontologically equal and empirically separated.
2-a) Vocabularies equal – translation difficulties even for EAT. Once again, our own definition will work as desired here, assuming that (in the case of the ontology being false) there exists an acceptable translation into the true ontology. As before, OAT would not be able to express a preference; however, now that the two models disagree empirically, EAT would.

Note though that the applicability of EAT is *also* dependent on an acceptable translation being available. In EAT's case, there needs to be an accepted translation of the parameter we wish actually to measure, else there would be no agreement on what the relevant empirical observable even is. ([Miller 1994] also makes this point.) It is true that this translation requirement is less strenuous than the one on our own scheme, since it requires the translatability only of the final effect whereas in our scheme we require in addition the translatability of the relevant component causes. In cases where translation difficulties affect only the causes and not the final effect, this difference could therefore become significant. For example, if one model cites two dragons kicking a ball three metres with particular contributions from each, whereas the second model claimed it was one dragon and two unicorns kicking it four metres with particular contributions from each, EAT would be able to compare the two since (we take it) the distance the ball was kicked is a well-defined observable all can agree on. Our definition, on the other hand, would struggle since we cannot easily make sense of the true values for the causal strengths of the dragon and unicorn kicks. Thus, in this circumstance, EAT would be able to offer an opinion and our definition would not.

I do not think, though, that this represents an advantage for EAT, and not just because such cases are presumably rather rare. The deeper point is: would we necessarily even

want our definition to deliver judgments in such cases? If no translation of the posited causes is available, this means in effect that we consider the models irretrievably mistaken, and so any empirical success of theirs with respect to the final effect we might want to put in the 'fluke' category. Accordingly, it should then be considered a *weakness* of EAT that it rewards such fluke successes, and a *virtue* of our scheme that it does not. Thus it is a mistake to imagine that EAT, by ignoring underlying causes and focusing only on measurement of simple observables, will therefore be applicable even in the difficult cases where our own scheme is not. Or rather, it could be so only by rewarding fluke correlations. In sum, EAT is not really compensated for its inapplicability to cases of empirical equivalence by any desirably greater range elsewhere than that already available to our own scheme anyway.

2-b) Vocabularies separated. Again, just as in case 1-b, the extra issue raised now would be the possibility of Miller-reversals, but the same analysis as before applies equally now. Note also that, in so far as the Miller issue affected our empirical observable, that analysis would apply to EAT too.

Case 3) Ontologically separated and empirically equal.

3-a) Vocabularies equal. It is not clear to me whether two models' vocabularies even can be thought the same if their ontologies are different, in which case this category is empty. But I suppose that, literally speaking, classical and relativistic mechanics each share a vocabulary of 'mass', 'acceleration', 'angle' and so on. In which case, refer to the analysis in case 3-b below, save with the amendment that we would not now need to worry about the Miller problem.

3-b) Vocabularies separated. Our own definition will still be applicable here if there are acceptable translations for both models, both of their ontologies and their vocabularies. If only one model has such acceptable translations, then that one will presumably be preferred, on the grounds that the very acceptability of the translation reflects a preference for that model's ontology compared to its rival's – but it is necessary to warn here that the question of when a translation actually is acceptable becomes considerably

more complicated when the competing models are ontologically separated (section 2-14), although we sidestep that issue for now. If neither model has an acceptable translation, then our definition must remain silent. (Of course, if a model is expressed in a true ontology, then it would be automatically suitable and no translation would be required.) EAT will clearly be of no help in this category, given the empirical equivalence. However, OAT may come into its own. In particular, if one model has the true ontology and the other does not, then (and only then, as we saw) OAT can express a preference between them. Whether or not this enables it sometimes to pronounce on cases on which our own definition remains silent, depends again on the details of how we analyse translatability (section 2-14). On the other hand, if both models have incorrect ontologies, albeit different incorrect ones, then while OAT must now remain silent, our own definition will still sometimes be able to express a preference – again, subject to translatability. With respect to the possibility of Miller-reversals, similar remarks apply as before.

Case 4) Ontologically separated and empirically separated.
4-a) Vocabularies equal. See the comments above for case 3-a, and hence also those below for case 4-b.

4-b) Vocabularies separated. The situation will be similar to case 3-b with regard to our own definition, to OAT, and to Miller-reversals. The only difference concerns EAT, since with empirical separation this now becomes potentially relevant – provided that we can agree on which observable to measure. In other words, EAT requires that the two models' final effect variables be translatable into an accepted common currency. As in case 2-b, this raises the possibility of EAT being able to deliver a verdict when our own definition (and OAT too) is unable to, but again only in cases where it is doubtful that we would *want* to be able to. Waters are further muddied now by a potential trade-off between ontological success and empirical success (section 2-14). Finally, it bears repeating again that, as argued in section 2-6, in the great majority of scientific disputes there is in fact no ontological disagreement between the competing models anyway, so all the discussion of cases 3 and 4 really concerns only rather unusual possibilities.

In summary:

1) *Ontologically equal and empirically equal.*

-- Our own definition works if the competing models are in a true ontology already, or if there is an acceptable translation into the true ontology.

-- Neither EAT nor OAT would be applicable under any circumstances.

2) *Ontologically equal and empirically separated.*

-- Again, our own definition works if the competing models are in a true ontology already, or if there is an acceptable translation into the true ontology.

-- Again, OAT is inapplicable.

-- EAT is now applicable, but in effect subject to the same translation constraints as our own definition.

3) *Ontologically separated and empirically equal.*

-- Our own definition is again at the mercy of translatability, although that latter issue may now become more complicated.

-- OAT may be applicable here sometimes, although whether ever when our own definition is not depends on the details of translatability.

-- EAT definitely will not be applicable.

4) *Ontologically separated and empirically separated.*

-- Our own definition is once again at the mercy of translatability, with the extra proviso that translatability may now become more complicated still.

-- OAT again may be applicable here sometimes, although whether ever when our own definition is not again depends on the details of translatability.

-- EAT is now applicable, but in effect again subject to the same translation constraints as our own definition, save that it is now more complicated whether it is ever usefully applicable when our own definition is not.

Thus note: first, we do not often if at all find that either OAT or EAT offers superior

coverage to our own definition. This is particularly noticeable in the cases 1 and 2 that in practice account for most real-life examples of science and scientific dispute, and arguably still holds even in the arcane cases 3 and 4 as well. Second, our own scheme's applicability depends critically on how we handle the issue of translation. So also does EAT's, as well as the extent to which OAT and EAT could ever exhibit superior range to our definition. Hence we must now turn explicitly to that issue of translation.

2-13) More on translation I: when is it acceptable?

(Note again that, throughout the next two sections, by 'translation' we shall mean sameness of reference rather than sameness of meaning.)

Three conclusions

As noted in the previous section, the critical requirement for our own definition to work is that we can make sense of there being true values for the relevant causal strengths. As we saw, even when two models agree on an ontology, often that ontology will nevertheless still be incorrect, as for instance with two competing Newtonian models. In such cases, it is necessary to interpret (or 'translate') the causes cited by the models into true causes that actually exist, in order that their strengths may be assigned true values. When is such translation acceptable? Or rather, as we argued, the key question is when the issue is sufficiently pressing that scientific argument concerns not what the values of the causal strengths are, but rather whether they can even be said to have objective values in the first place. In most everyday Newtonian cases such as kicking a ball, it seems that translation is indeed deemed acceptable. Accordingly, objective values for the relevant causal strengths are assumed to exist, and thus our definition can work smoothly.

1) Context-specificity

This clarifies what translatability means here. What principles, if any, govern when it obtains? To begin with, consider the simplest case, where two competing models agree on their ontology. Then the only question is whether that ontology is sufficiently 'correct' for a translation to be licensed. In the particular case of kicking footballs, it seems that we normally do view such translations favourably. We typically seem to be willing, that is, to accept that a Newtonian force may, in this context, indeed be ascribed an objectively true magnitude. Or at least that the translation into a true ontology is sufficiently clear and precise that for practical purposes there is 'near enough' a single true value.

But is this always true of Newtonian models? For consider instead if we were comparing instead two different Newtonian models of some interaction between quarks within the atomic nucleus. In this new context, the ontology of quantum mechanics or quark theory – assuming this to be the correct one – is now so different from the Newtonian picture that it might well seem impossible to furnish any good translation between the two. Thus if different Newtonian models assigned different strengths to two quarks 'pushing' a third one, it might well seem impossible to translate this into the exotic ontology of the particles' 'charm', 'spin' and so forth we take now to be the actual ontology. In which case, true values for the strengths of the quarks' Newtonian 'pushes' might not be forthcoming. The first conclusion then is that the acceptability of an ontology may vary by context. Thus it may be (sufficiently) possible to assign true values to Newtonian causal strengths in the everyday context of kicking a ball, even though it may *not* be so possible in the different context of quark-quark interaction.

2) Empirical not ontological

Two supplementary remarks arise. First, it might seem as if one of the reasons Newtonian models are particularly favoured, at least in macro-level contexts, is because it seems there already exists a ready-made translation into the true ontology. In particular, if we assume the true ontology to be relativistic there is a clear sense in which Newtonian mechanics comes out as a limiting case if we let the speed of light tend to infinity and so forth. However, I do not think that such ontological *near*-compatibility, so to speak, is

actually the driving force behind the Newtonian models' acceptability here. Consider fluid dynamics, for instance. Suppose water flowing down a pipe is split into two channels, one narrow and one wide. Now compare the causal strengths of what the theory would term the 'water pressure' in each channel. It seems to be uncontroversial that there would be a true fact of the matter as to which was the stronger here. But this is despite the fact that models from fluid dynamics all assume that the fluid concerned, i.e. water, is a perfect continuum, whereas of course real water is in fact granular and consists of hard molecules. Thus classical fluid mechanics is a stubbornly false theory (to use the terminology of section 1-3), is *not* 'near-compatible' with the true ontology (to use the terminology of this paragraph), yet still suffers from no translation problems. It seems that this is because the *empirical* error or uncertainty (with respect to the variables we are concerned with) due to its false ontological assumptions is very small. Thus, we reach a second conclusion: what determines a false ontology's translatability is whether the error or uncertainty (with respect to the particular causal strengths of interest) due to having that false ontology is small *empirically*. The degree of 'ontological error' is irrelevant.

3) Holistic

Our second remark concerns what we might mean by terming an error to be 'small empirically', and highlights a holistic aspect. Loosely speaking, when translating here we are reconstructing what a false-ontology might 'really mean' in terms of the true ontology. In order to do this, it will typically be helpful to incorporate information – if available – from contexts other than the one immediately to hand. For instance, it will likely be from our knowledge of many different applications that we confidently translate particular Newtonian terms into relativistic equivalents, in order to assign objective causal strengths to them. And equating (in some contexts) 'dephlogisticated air' with oxygen will likewise likely be from knowledge of many different instances of that term's usage. Choosing to declare 'dragon' as having no reference in a true ontology will also likely be a similarly holistic decision. One might appeal vaguely to Quinean remarks in order to motivate doing our best to reconcile all the different theoretical ideas and empirical data when determining the reference of a mistaken ontology. Any translations will emerge from such efforts. Can particular false-ontology terms reasonably be translated into particular

true-ontology ones? This *a posteriori* judgment we take typically to be holistic in nature – which is our third conclusion.

All three of these conclusions will now be illustrated in the context of a thought-example.

A fourth conclusion: partial objectivity

Note that whereas we take the strengths of true causes to be objective facts of the universe, whether or not some false-ontology cause is translatable into a true-ontology one will typically be a much more arguable matter. Nevertheless, I do not think that translatability is thereby rendered unacceptably arbitrary. Consider the following example: two stereotypically medieval drunkards argue over the cause of a tree falling down. One assigns the felling primarily to the action of a malevolent dragon, the other primarily to that of a malevolent unicorn. Which model is the more approximately true? Suppose that the actual cause of the tree falling was a lightning strike. We should likely say that neither model is translatable into the relevant true ontology of low air pressure, electrical potentials and so forth, hence that no true causal weightings exist within the drunkards' quoted ontology, hence that our scheme cannot judge between the models, and hence that approximate truth cannot even be made sense of in this case. (Neither OAT nor EAT could distinguish between the two models either.)

Now suppose that the drunkards found a soothsayer friend, who announced that in fact there was a true causal weighting for each of the dragon and unicorn after all, because the gods had revealed it to him in a dream. Suppose that each drunkard was convinced by this and accordingly agreed that the causal weightings quoted by the soothsayer were indeed the true ones, and so did enable us to compute scores for the approximate truth of each of their models after all. But just because the two drunkards agree on this, it does not follow that *we* have to. Rather, we are free to find the authority of a soothsayer's dream somewhat insufficient.

Next imagine a further twist to the story. Suppose that in the mythology of the two

drunkards' village, thunderstorms were attributed to angry dragons breathing down lightning strikes, while damage caused by unknown large animals was attributed to angry unicorns righting some ancient wrong. Suppose further that, notwithstanding the supernatural elements to these beliefs, much practical knowledge and behaviour was expressed in the same terms, for example that being exposed on the top of a hill although normally safe was likely to provoke the breath of a dragon when it was angry and so should be avoided at those times. Would it not be natural to interpret this to be advice that exposure on a hill during a thunderstorm leaves one vulnerable to being struck by lightning? And would not such advice reflect some genuine knowledge of the world?

Suppose now that we wished, like our drunkards, to know the cause of some tree falling down. If we saw that it had been sliced down the middle this would be evidence favouring a struck-by-lightning hypothesis, whereas if it had been lifted from its roots this would favour an uprooted-by-animal hypothesis. If the soothsayer reported that, having seen the tree sliced down the middle, he had 'heard from the gods' that the tree was felled by an angry dragon rather than unicorn, this might be yet further supportive evidence for the following conclusion: that, in this particular context, it is acceptable to translate the first drunkard's dragon model to be that the tree was felled by lightning, and the second drunkard's unicorn model to be that it was felled by (real) animals. Accordingly, assuming the lightning-animal ontology indeed to be a true one, we could now after all assign true weights to each of the dragon and unicorn models, and therefore could now after all meaningfully compare them for approximate truth.

This example illustrates several points. First, whether or not models actually are working with the true ontology is presumably an objective matter – we take it that dragons and unicorns do not exist. Second, whether or not the researchers themselves realise that their models have the right or wrong ontology is irrelevant to the definition of those models' approximate truths; what matters is whether the ontology is indeed sufficiently correct to be translatable. Third, for this reason the researchers may not themselves know whether their models are or are not translatable. Fourth (and this is our fourth conclusion generally in this section), whether or not models are translatable is at least partly

objective. The elements in the last part of the story that persuaded us the models might after all be translatable were neither arbitrary nor purely subjective. And it is these elements that decide the matter. Thus, in the earlier absence of them, even though the drunkards themselves were happy to accept the values for the causal strengths offered by the soothsayer, we felt licensed to reject translatability on objective grounds.

The example also illustrates nicely our three earlier conclusions too. First, that the assessment of translatability is holistic: the ontology of the drunkards' models seemed, in isolation, clearly too fantastic to permit translatability. It was only in the light of the *wider* usages we then postulated for the terms 'dragon' and 'unicorn' that their possible reference to a true ontology became sufficiently established. Second, context-dependence: although clearly a lightning strike is not the same thing as a dragon breathing fire, this ontological error is not important in the particular context of deciding between the two possible causes of the tree falling. For instance, if we were designing protective measures for the tree, with both the dragon and lightning ontologies we would correctly perceive that a protective fence around the tree ('to guard against unicorns') would be ineffective. But in other contexts the ontological error might become more important. For instance, an anti-dragon amulet would be a rather less effective protective measure for the tree than a lightning rod, but these two new causal strengths the old dragon ontology would presumably get seriously wrong, meaning that the choice between the two ontologies would have become significant again. Third, note once more that the reason it would have become significant is because of the large *empirical* difference between the relevant causal strengths of the amulet and lightning rod. That is, the ontological error now leads to a seriously incorrect empirical reading for the causal strength of the amulet 'protective measure', and it is this empirical fact that leads to translatability being denied.

Causal invariances

Loosely speaking, we feel licensed to infer translatability just when there seems to be some implicit reference by a false ontology to entities in a true ontology, enough to

enable us to assign – sufficiently non-arbitrarily – objective causal strengths to the entities of the false ontology. It seems that such inference depends on capturing a certain kind of empirical regularity. For instance, in our medieval drunkard example we postulated that the use of 'dragon' (in particular contexts) tracked the real-ontology entity of lightning. Thus 'dragon' was invoked whenever there were bright flashes in the sky during thunderstorms, evidence of trees split down the middle, and so on, but was not invoked (in these contexts) otherwise. In this sense, the term matched certain empirical regularities we associate with lightning, hence justifying the translation here.

It is also possible, indeed convenient, to state this point in terms of *causal invariances.* Thus, the argument might go: all else being kept equal, the cause 'dragon' invariantly makes certain effects more likely, and if those effects match (sufficiently) those that lightning invariantly make more likely, then we are justified (in this context) in our translation of 'dragon' as lightning. This then is perhaps the key to defining translatability. Although there are pragmatic qualifiers to it ('sufficiently' and 'in this context'), nevertheless I do not think the criterion is hopelessly arbitrary or subjective. And although the adjudication of borderline cases may of course be left disputable, in practice it may often be relatively uncontroversial whether or not a particular translation is acceptable.

Much, perhaps most, old science worked with incorrect ontologies, for instance Newtonian physics and phlogiston chemistry. With our translatability criterion in mind, we can specify more clearly now just when disputes conducted in the terms of these false ontologies can and cannot be meaningfully analysed for approximate truth. In the case of our medieval drunkards, the dragon and unicorn ontology seemed too fantastical to allow any such judgments until it turned out that (in certain contexts) they did match up (sufficiently) onto the invariances associated with real causal entities such as lightning and animals, at which point an approximate truth analysis did become possible after all. And we saw that Newtonian models were similarly acceptable for translation, except in contexts (for instance nuclear physics) where the Newtonian causal entities no longer matched up sufficiently well with the invariances associated with any of the relevant

causal entities from the real ontology.

As a final illustration, consider a hypothetical case from phlogiston chemistry. The ontology involved is now known to be incorrect, of course. Nevertheless, rather in the manner of Newtonian models, sometimes it can still seem sensible to speak of one model being more approximately true than another. Suppose we were concerned with producing water, and a first model cites a particular causal strength for a particular quantity of 'flammable air' (i.e. what we would now call hydrogen) in the presence of a particular quantity of 'dephlogisticated air' (i.e. oxygen). The second model cites a different causal strength. Say the first model implies that to synthesise water two parts flammable to one part dephlogisticated air is required (i.e. H_2O as it were), and the second model implies instead a ratio of one-to-one. Can our definition say which model is more approximately true? If we can make sense of what the true causal strength of the 'flammable air' is here, then yes we can. In our terms, the acceptability of the suggested translations is established by holistic consideration of which real-ontology causal invariances the phlogiston terms seem to track, and in particular whether (in this context) this match is sufficiently close. Here, even though the ontology of phlogiston chemistry is known to be wrong, still sometimes there seems to be a sufficiently clear interpretation of its terms into modern vocabulary that we can legitimately prefer one phlogiston model to another – in particular, prefer the two-to-one over the one-to-one model. Of course, it may be that in other examples the phlogiston models have no such easy interpretation. The acceptability of such interpretations will vary context by context. Still, I think there will be some occasions where disagreements between phlogiston-era scientists are scientifically meaningful rather than merely ontologically confused, and this will be so just in those cases where the models involved are translatable.

[Aronson, Harre and Way 1994] state that a theory's ontological adequacy can be determined by the ability it affords us to manipulate individuals of the kinds that it treats. That is, if it posits certain entities that we can then manipulate by intervention, this is evidence for those entities' existence. I take this point to be similar to our one here concerning causal interventions – namely, that a causal strength is (sufficiently) well-

established if and only if it licenses an associated causal intervention, and hence that a successful such intervention confirms the existence of the causal strength.

Summary

We have concluded that translation is acceptable if and only if the causal terms of the false ontology track sufficiently well invariances associated with causes from a real ontology. This criterion incorporates our four earlier conclusions about translatability:

1) The acceptability of a false ontology will vary by context. Thus the terms from the false ontology will in general match the invariances associated with particular real-ontology causes only in some but not other contexts.

2) What governs a false ontology's translatability is that the error or uncertainty (with respect to the particular causal strengths of interest) due to having the wrong ontology is small *empirically*. The match with invariances is defined with respect to the quantity of effect, which is an empirical entity. The intuitive closeness or otherwise of the false ontology to any real one is irrelevant. For instance, dragons are presumably far removed ontologically from reality; nevertheless, if they track empirical invariances associated with lightning, still they may be good candidates for translatability.

3) Judgments of translatability are holistic. Thus (in practice) we cannot assess well whether a particular invariance has been satisfactorily captured except by examining a plurality of instances. Note though that these instances must of course be relevant, that is must concern the same effects in the same or (sufficiently) similar context – this follows from point 1 above. So to that degree the holism is circumscribed.

4) Whether or not models are translatable is at least partly objective. Thus when we speak of an invariance being captured 'sufficiently', although this is partly a pragmatic decision, clearly it is also informed by objective factors.

There is no denying that the whole issue of translatability seems to be unattractively messy, but the discussion of this section has aimed to show that it can nonetheless still be analysed with profit. More importantly, the issue is anyway *unavoidable*. In particular, even though it may seem unattractively messy, translatability is necessary for

underpinning clearly useful notions such as deeming one Newtonian model more approximately true than another. Therefore a hardline anti-translation approach would seem unreasonably strict. For instance, we can generate and calibrate useful Newtonian interventions, so to rule those out just because Newtonian ontology is incorrect seems unnecessarily restrictive. True, the causal strengths assigned to Newtonian causes can only ever be as precise as the associated translation, but as remarked repeatedly, this may be plenty precise enough for all practical purposes. To insist on a perfectly objective and perfectly precise score for approximate truth for such models, would seem just to be making philosophy irrelevant to scientific practice. My position is that we should, even at the expense of some conceptual complexity if necessary, seek if possible to give a rigorous account that is relevant to actual science, and in particular relevant to actual judgments of approximate truth.

2-14) More on translation II: further discussions

When two models' ontologies differ: degrees of translatability

The previous section was concerned exclusively with cases where the competing models, although framed in a false ontology, still were framed in the *same* false ontology. What now if the two competing models are framed in *differing* ontologies? The issue of translatability will still be relevant, since we shall be interested here in cases where at least one of the models has an ontology that is false. Can we just use the same analysis of translatability as in the previous section? We shall indeed again make use of the notion of a false ontology sometimes capturing invariances of real causes, but it turns out that now the issue is subject to some fresh twists. In particular, the biggest flaw in our previous treatment of translatability has been interpreting it as an all-or-nothing thing, with no concept of *degrees* of translatability.

Suppose that, as in our first presentation of it, the Ptolemaic model performs empirically better than the Newtonian one. If the Newtonian model is deemed translatable and the Ptolemaic one not, then our definition would unambiguously prefer the Newtonian model. If both were deemed translatable, then it would likely prefer the Ptolemaic one. Problems would arise only once we started to think that the Newtonian model was *to some degree* preferable ontologically to the Ptolemaic one, perhaps as it were that the Ptolemaic model was only 'half-translatable' but the Newtonian one fully so. In such a case, we would seem to be forced to make a *trade-off* between the ontological and empirical aspects of approximate truth – the Newtonian model would be to some degree preferable ontologically, the Ptolemaic one to some degree preferable empirically.

Perhaps this point can be sharpened if we imagine comparing a Newtonian model now with a relativistic one, and stipulating that the Newtonian one was empirically superior. In other words, now we are comparing for approximate truth one model with a false ontology against one with a true one. If we deem the Newtonian model fully translatable, then in this example our definition would prefer it. But it seems to me that this is likely to seem intuitively unacceptable if the empirical difference between the two models is very small. For instance, suppose the two differed in their prediction only in the eighteenth decimal place – in that case, the relativistic model's empirical deficit, so to speak, would seem negligible, and certainly less than the benefit that should accrue from its ontological superiority. (This horn of the dilemma might be made sharper still if we substituted for the Newtonian model a Ptolemaic one that we had deemed translatable.) On the other hand, suppose that the Newtonian model was empirically accurate whereas the relativistic model was empirically hopelessly wrong – that the Newtonian one could (say) successfully land a spacecraft on the moon, whereas the relativistic one could not even get it off planet Earth. In such a case, there would be a strong intuition that the Newtonian model had captured more of how the universe actually is, that is was more approximately true, notwithstanding its ontological inferiority. In other words, neither 'extreme' solution is satisfactory (as we noted in our original discussion of OAT and EAT) – both the ontological and empirical aspects of approximate truth matter, and neither can be ignored completely. It does not seem that *either* the relativistic or the

Newtonian model should *always* be preferred. Rather, it would all depend on the extent of the Newtonian model's empirical superiority, compared to the extent of the relativistic model's ontological superiority. A trade-off is unavoidable. It follows that our intuitions about approximate truth do implicitly support working in terms of degrees of translatability, as this seems to be the only way (within our framework) of licensing that trade-off between ontological and empirical success. (The ontological half of these trade-offs would presumably be dictated by much the same criteria as dictated our translatability before, save now rephrased appropriately – thus, 'the *degree* to which a model's ontology captures real causal invariances', and so forth.)

Have we moved beyond simple EAT and intuition?

It may be worth returning to the question of whether our own definition even actually adds much to a simple EAT one. Recall first our arguments from section 2-9 on this matter: namely that, in order to sort out correct from incorrect explanations we must examine the value not just of the final explanandum variable (which is all EAT would consider) – since this is bound to be the same for all explanations good and bad – but also those of the cited component causes. This method also enables us to rule out fluke correlations too.

Now, our discussion of translatability here might seem to raise the following worry: whether or not something is translatable seems to boil down just to tracking when an EAT result does or does not feel intuitively satisfactory, and adds nothing to this intuition. Thus if a model yields the right final answer empirically but still seems to be unacceptably wrong ontologically, for instance if it is couched in terms of dragons, then that model being deemed correct by EAT seems unsatisfactory. In just such cases we would deem the model 'not translatable', but on what grounds beyond EAT's intuitive unsatisfactoriness? Perhaps translatability is just an *ad hoc* label for our intuitive reactions to an EAT result? Another way of putting this point is to say that what matters when assessing the satisfactoriness of an EAT result is whether or not the 'ontological gap', so to speak, is important. That is, when has an ontologically mistaken model

nonetheless still captured successfully some relevant patterns in nature, and when instead has it achieved a high EAT score only through luck or through *ex post* worthless storytelling? When, as it were, are we dealing with Newtonian celestial mechanics, and when with magic dragons? These ontological considerations become still more pointed, as we saw, in examples where we are forced in effect to trade off ontological success against empirical success. So, does our notion of translatability do anything other than track our extra-EAT intuitions here?

First, EAT itself cannot escape the need to consider translatability in any case (section 2-12). Admittedly, tackling translatability indeed turns out to be messy, but unfortunately just because there is no easy treatment does not make the *need* to attempt one therefore any less pressing. What we have dubbed 'translatability' is just a systematic exposition of what factors are relevant to the task, such as tracking causal invariances and so on, which perhaps may serve at least to clarify just what our intuitions are. Such reflection may also clarify the motivations behind them, and perhaps thereby eventually *lead* intuition. It is true that in exotic mixed-ontology examples the weaknesses of translatability – namely its imprecision and its pragmatic elements – are emphasised. But in most cases of actual scientific dispute there is little controversy over ontology, and in such – more typical – circumstances it is now the strengths of our approach over EAT that are highlighted. Thus, I think, overall our emphasis on translatability does enable us to construct a much better account of approximate truth than would have been possible armed solely with EAT and intuition alone.

Some other questions

Lexicographic OAT/EAT?

Perhaps another way to avoid the troublesome talk of translatability would be to construct a simple two-part definition, utilising only the more clearly defined notions of OAT and EAT. The idea would be: if two models have the same ontological status, EAT should be used as the tie-breaker; otherwise, the ontologically superior model always wins. The

advantage of this simple system, so the argument goes, would be that we could ignore translatability.

Unfortunately, the weaknesses of such a scheme are clear and indeed by now familiar. Thus: sometimes we might want to prefer an ontologically inferior model if it is sufficiently empirically superior. Or sometimes one wrong ontology seems much preferable to another. Plus, in cases where the models' ontologies differ, only crude qualitative judgments would be available. And in cases where the models' ontologies are the same, the suggested definition would inherit the same weaknesses as plague EAT. Hence fluke correlations could not be ruled out, nor explanations that while correct ontologically were otherwise obviously ridiculous or mistaken. And as noted, EAT anyway in fact does require some consideration of translatability in any case. Thus again, I do not think that there exists any satisfactory short-cut enabling us to evade a discussion of translatability.

Given the inadequacy as definitions of OAT and EAT on their own, and now of any combination exclusively in terms of them, it seems that there is little alternative left but to move beyond them. In particular, it seems there is little alternative but to tackle head on the question of when a false ontology is or is not acceptable. This of course is precisely the focus of our discussion of translatability.

When do we have normative warrant?

Because our causal strengths are defined in terms of empirical effect (chapter 3), if a model gets a causal strength correct this can be verified through empirical intervention. That is, a particular intervention can be accurately predicted to have a particular effect if and only if a model has the relevant causal strength correct. In this sense, the degree to which a model gets the relevant causal strengths correct can be verified empirically. Thus the normative force of a ranking for approximate truth on our definition can be backed up by such empirical warrant. However, if it is disputed not whether a model has gauged the true causal strengths correctly, but rather whether a model's terms have any objective causal strengths at all – that is, if what is disputed is not a model's degree of

approximate truth but rather its translatability – then this normative warrant no longer applies. The warrant comes from intervention but that requires agreement on what the intervention variable is, and this latter is of course precisely what is at stake in disputes over translatability. For example, if we could agree that 'phlogisticated air' indeed referred to nitrogen then we could check its causal weighting via an intervention with nitrogen, but such an intervention could *not* tell us whether 'phlogisticated air' indeed *did* refer to nitrogen in the first place. The normative warrants for our judgments of approximate truth thus only kick in *after* decisions about translatability have already been made.

The inference from this is that our judgments regarding translatability are therefore rather less authoritative than our subsequent ones regarding degrees of approximate truth. I agree. The point is then what further inference we should draw as a result. One would be that, while fine in the context of normal science, we should abandon all talk of approximate truth in the more difficult cases where translatability is controversial. But we have seen that neither always allowing nor always disallowing translatability seems satisfactory, and that it is still possible to analyse the issue with some profit. Certainly, we still feel strong intuitions about approximate truth even in the wrong-ontology contexts where translatability is required, and I would claim that some philosophical account here is better than none.

A plurality of accounts?

As noted (section 2-5), our scheme can be seen as a way of combining the 'best of both worlds', incorporating into one definition both the ontological and empirical aspects of approximate truth. We have seen that complications arise both when extending this scheme beyond normal science but also when, in contrast, the issue of translatability is instead evaded. Perhaps then a better solution would be just to accept that the univocality of the notion of approximate truth comes apart in these problematic contexts? Thus neither OAT nor EAT are fully satisfactory, and may yield contradictory results, but this should be seen merely as symptoms of the fact that 'approximate truth' actually conflates two separate notions which we are erroneously used to thinking about in a unified way.

Such an attitude might seem tempting, but I do not see that it actually confers any advantages. Recall that, upon closer inspection, even EAT on its own must rely on some notion of translatability, so that it is not really any better off than our own definition in that respect. Moreover, OAT too must embrace translatability on pain of disallowing all false-ontology models – that is, potentially all of science past and present – from positive consideration, even though it was examples of just such models (e.g. the presumed ontological superiority of Newtonian over Ptolemaic astronomy) that provided the motivation for OAT in the first place. That is, even if we accept that approximate truth comes apart into two separate notions, still we cannot avoid the issue of translatability. This therefore removes the motivation for the pluralist approach. In trying to provide a univocal notion of approximate truth – as we have been doing – we are no worse off conceptually, and better off in so far as we do at least capture the intuition of approximate truth's univocality. Hence, even in awkward cases, there seems to be no philosophical profit here to adopting pluralism.

Conclusions about translatability

First, in practice disagreement about translatability is very rarely at the heart of any actual scientific disputes. Accordingly, it is mainly for the sake of philosophical completeness rather than practical applicability that we need address the issue at all.

Second, translatability is the label we have used for the issue of ontological acceptability. In particular, what is the cost a model incurs by citing an incorrect ontology, and how do we accommodate the fact that some wrong ontologies are judged worse than others?

Third, the issue is unfortunately unavoidable. No serious definition of approximate truth can evade it. Or at least, any attempt to do so turns out either to be deluded or else to come at the cost that we could only rarely pronounce on approximate truth at all, even in many of the commonest and apparently simplest cases of it.

Fourth, the issue is messy, in the sense that there seem to be no neat philosophical criteria that fully determine to what degree any particular translation is acceptable. Pragmatic influences are to a degree inescapable. Thus our criteria for translatability should perhaps be seen merely as a summary of the clearest thinking we can manage on this tricky issue.

Fifth, although messy, the issue is not so messy that we cannot still get some philosophical purchase on it. While pragmatic factors do inescapably enter our criteria for translatability, so also do perfectly objective factors too, meaning that these criteria are not wholly arbitrary. Thus, rather than just leave the assessment of a model's ontological acceptability entirely to intuition or extra-philosophy, it is possible with profit to analyse it more rigorously.

Appendix – An exact definition

Choosing a measure of similarity

We have introduced a framework in which to assess approximate truth, but not yet given an exact definition of it. What we have are a set of true values for the relevant causal strengths, a set of values ascribed to them by a model, and the conclusion that what matters is these two sets' similarity. It is useful to re-describe this situation in terms of a hypothetical vector space with causal strengths along each axis. Then if the truth of any particular context is some particular set of values for the relevant causal strengths, this will correspond to one particular point in this vector space ('truth-point'). And competing models' own estimates of these causal strengths will likewise correspond to their own particular points ('model-points'). The degree of approximate truth of any given model is then seen as the closeness of its point to the truth-point. But what we have so far left unspecified is exactly how to define this 'closeness' between two points. We turn to that task now.

Smith concludes in a similar context: 'and if problems remain here, they are problems in geometry, of specifying suitable metric approximation relations, not conceptual problems' [Smith 1998, p253]. I think it is true that no remaining conceptual problems remain, which is why we tackle this issue only in an appendix. Moreover, to concentrate excessively on the question of the precise measure to use would be to repeat an emphasis that we previously criticised much of the existing literature for – namely, an emphasis on the exact syntactic form of the measure rather than on the philosophical interpretation of the distance that was being measured. In practice, it may be only rarely that the exact choice of measure is really decisive in adjudging which of two models scores best. Nevertheless, even this claim already assumes that obviously ridiculous measures have been eliminated, without explaining how. And in any case, I think it does still turn out to prove philosophically profitable to proffer a specific definition and then test it against various difficulties. We shall see that there are good reasons to prefer some measures over others; the issues are not mere 'problems in geometry'.

For simplicity, we shall assume our vector space to be Euclidean. Since it is only an abstract space, this does not commit us to any physical implications. Nevertheless, there is certainly a philosophical interpretation attached to this vector space and this enables us to reject some candidates for a measure of similarity immediately. For example, one broad principle we can immediately endorse is that of each axis being equally important. Of course, a foundation of our approach is the allowance for asymmetric weightings on each cause, but this just reflects their different strengths in reality and is in turn already incorporated in our vector space by how far the truth-point lies *along* each axis. But to emphasise one *axis* itself more than another, beyond this allowance for causal strength, has no licence from objective reality. Thus a measure which emphasised the error in the y-coordinate (say) over the errors in other dimensions would not be acceptable, since such special treatment would have no justification from the underlying physical situation we are trying to represent and so would be unacceptably arbitrary. Otherwise, for instance, we might ask why we could not simply *re-label* the axes so that the 'y' now represented a different physical cause, and by this means alone change our result for

approximate truth. The arbitrariness of any such re-labelling reflects the underlying arbitrariness of treating the axes of our vector space in any way asymmetrically.

Another broad principle we should want to endorse is the need for absolute, and not just relative, weightings. For instance, suppose that there are two equally massive planets in space, and that we release a ball near them. (As usual in Newtonian cases, for simplicity take the planets and ball to be point masses.) Next assume that when we release this ball it is equidistant from each planet, so that the ball and two planets form the three points of an isosceles triangle. Now consider a second situation exactly the same as this first one, but with the single exception that this time the ball is twice as far away from the planets as before, although still equidistant. Define two problems, one for each of these two set-ups, in each case asking: what is the gravitational pull on the ball from the two planets?

Suppose that at first we were still considering only relative weightings. Then we would be unable to distinguish between these two cases, even though they are clearly physically distinct. This is because a relative weighting of the planets' gravitational pulls on the ball is the same in both cases – half comes from each planet. Yet clearly the absolute pull is *not* the same in both cases, as it is much stronger in the first case when the ball is nearer the two planets. A concentration solely on relative weights misses this difference, and so fails to distinguish between two substantively different physical states of the world. It would follow that we could define two substantively different models (e.g. that the pull is $10ms^{-2}$ or that it is $20ms^{-2}$), and a purely relative weighting function would be unable to tell them apart. This cannot be satisfactory. (A similar problem would arise if the ball was each time the same distance from the planets and it was the planets' masses that were varied instead.) In order to get satisfactory ontological descriptions for every case, relative weightings alone are not reliably sufficient.

One example of a measure that would fall foul of this principle is the inner product. (The inner product of two n-vectors **X** and **Y** is defined by: $\mathbf{X}.\mathbf{Y} = (\sum x_i y_i) / (|\mathbf{X}| \, |\mathbf{Y}|)$, where x_i and y_i are **X** and **Y**'s respective components in the direction of the *i*th axis, and the sum is for i = 1 to n.) We can think of the truth-point and model-point in our vector

space as defining two vectors. One way of measuring the two points' similarity is by measuring the similarity of their *directions* relative to the origin, which is what the inner product does. But **X.Y** = k**X.jY** for any constants k and j, that is absolute weightings have no influence on the inner product if applied uniformly to one of the vectors. (Intuitively, if we travel north-east we remain the same bearing relative to the centre no matter how *far* we travel.)

It might seem that this need to incorporate absolute weightings will cause problems, since often the absolute units we use will be arbitrary. For instance, if the effect term is distance travelled, then our causal strengths could equally well be expressed in kilometres or centimetres (or inches). But all that is required is that we are consistent within any one context. If the true causal strength is measured in centimetres, say, then so should be the models' postulated ones; if in kilometres, then likewise the models'. The requirement for allowing absolute weightings only applies once this initial scale has been agreed on, so to speak. So if the true distance was 3km east and 3km north, then we need to be able to prefer a model that postulates 4km each way over one postulating 6km each way.

Of course, there remain many possible measures satisfying both the above principles, namely treating each axis equally and allowing absolute weightings. How do we choose between these? In particular, what of perhaps the most obvious measure of all, namely Euclidean distance? (The Euclidean distance between two points $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$ is defined by: $(\sum (x_i - y_i)^2)^{\frac{1}{2}}$, the sum being for i = 1 to n. Intuitively, it is just the distance of the straight line drawn between them.) Can we not simply use this as our measure of closeness to the truth? But it turns out that philosophical consideration yields another desideratum, and that this one proves enough to rule out Euclidean distance too.

Why Euclidean distance does not work: the causal subdivision problem

Return to our original example (section 2-1) of dropping a ball and then estimating the relative gravitational pulls on it of a nearby mountain and the rest of the Earth. Suppose,

in some units of force, that the correct causal strengths are (3, 100) respectively. (For simplicity, ignore the fact that in reality the ratio would obviously be much greater than 100 to 3.) So in our two-dimensional vector space, the truth-point is (3, 100). Suppose that our first model ignores the mountain, but slightly underestimates the strength due to the rest of the Earth, putting it at 96 instead of 100 units. Then its model-point in our vector space would be (0, 96). And suppose that our second, mountain-only, model is accurate as far as it goes, so that its model-point is (3, 0). Then the Euclidean distances between each model-point and the truth-point are:

1) for the Earth-model: $[(3 - 0)^2 + (100 - 96)^2]^{½} = (25)^{½} = 5$

2) for the mountain-model: $[0^2 + (100 - 0)^2]^{½} = 100$

Thus the Earth-model would be judged much closer to the truth, as desired.

So far so good, but suppose now that we re-described the situation. Suppose in particular that we now considered the Earth as two separate halves, say the western and eastern hemisphere, each contributing their own gravitational pull. There would therefore now be *three* different causes at work, with strengths (3, 50, 50). Likewise, our two models' points would presumably now be (0, 48, 48) and (3, 0, 0) respectively. Of course, this is *exactly the same physical situation* as before; all that has changed is our description of it, in particular our arbitrary division of the Earth's gravity into two causes instead of one.

A third desideratum for any measure now presents itself: it should be invariant with respect to such arbitrary re-descriptions. So long as the physical reality and our focus of interest are the same, so should be our judgments of approximate truth. Such a desideratum recalls the Miller problem, but unlike that it is vital in practice that this one be satisfied. All that is varying is how we subdivide causes, and frequently there may be no obvious or 'natural' way to do this. For instance, it is not obvious whether we should or should not subdivide the Earth in this example, or even whether we should separate the mountain from the rest of the Earth. In other contexts, surely judgments of approximate truth should not be dependent on, for instance: whether we describe the causal strength of an army as a single entity or as the sum of the strengths of the individual soldiers; the causal strength of gas pressure as a single entity or the sum of the strengths of individual

molecules; or the causal strength of smoking as a single entity or as the sum of the strengths of cigarettes smoked in the morning and those smoked in the afternoon, or as the sum of all the cigarettes individually. All such choices of description are not only arbitrary, but also likely to vary *in practice*. Hence it is a practical as well as theoretical necessity that our judgments of approximate truth be invariant with respect to them.

It might be thought that our scheme of causal strengths achieves precisely this – after all, if a cause divided in two such as the Earth in our example, now each sub-cause is awarded only half the strength of the composite entity. This is true and of course one of the clinching merits of our general approach, but it turns out we now have to be careful that our chosen measure of similarity does not, as it were, squander this hard-won advantage. Return to our Earth example. Under the new two-hemisphere description, recall, the truth-point is now (3, 50, 50) and the two model-points respectively (0, 48, 48) and (3, 0, 0). This yields new Euclidean distances of:
1) for the Earth-model: $(3^2 + 2^2 + 2^2)^{½} = (17)^{½} = 4.1$ (approximately)
2) for the mountain-model: $(0^2 + 50^2 + 50^2)^{½} = (5000)^{½} = 70.7$ (approximately)
Previously, the results were 5 and 100 respectively, thus they have now *changed*. This is the reason why simple Euclidean distance is an unsatisfactory measure for our purposes: if using it, a purely arbitrary re-division of agreed causes leads to a change in scores for approximate truth.

Note that not only the gross scores have changed, but also the ratio between the two. This hints at why the following counterargument does not work either: that although the scores have changed, still there is no need to worry since the ranking has not – the Earth-model remains much the preferred one of the two. If this happy outcome always transpired it might indeed be some comfort, especially given our earlier emphasis on relative rather than absolute scores (section 2-7). But unfortunately we can construct examples where even the qualitative ranking alters too. For instance, suppose the numbers in the Earth example had been: truth-point (12, 12), the first model M_1's point (8, 8) and the second model M_2's point (9, 7). Next suppose we arbitrarily sub-divided the second cause into two, yielding a new truth-point of (12, 6, 6) and new model-points

of (8, 4, 4) for M_1 and (9, 3.5, 3.5) for M_2. These would yield the following Euclidean distances:

1) *Before the subdivision.* For M_1: $(4^2 + 4^2)^{½} = (32)^{½} = 5.7$

For M_2: $(3^2 + 5^2)^{½} = (34)^{½} = 5.8$

Thus M_1 is adjudged closer to the truth.

2) *After the subdivision.* For M_1: $(4^2 + 2^2 + 2^2)^{½} = (24)^{½} = 4.9$

For M_2: $(3^2 + 2.5^2 + 2.5^2)^{½} = (21.5)^{½} = 4.6$

Now it is M_2 that is adjudged closer to the truth. Thus Euclidean distance does indeed yield rankings for approximate truth that are not invariant with respect to arbitrary subdivisions of agreed causes. (Intuitively we might diagnose the effect as resulting from the Euclidean distance giving undue weight to a cause once it is subdivided into two, that is to a cause awarded, so to speak, two axes rather than one in our abstract vector space.)

This particular difficulty does not affect only Euclidean distance; other possible measures are similarly affected. For example, we might consider the errors in each model's ascribed causal strengths and then take the *variance* of those errors, the model with the lowest variance scoring the best. (Thus a true model's errors would all be zero, and hence it would score a minimum variance also of zero.) This would be analogous to the least-squares procedure common in statistics, but also falls foul of the causal subdivision problem. Or take another example: it is possible to adjust our inner product measure by adding a normalising coefficient reflecting the relevant vectors' absolute strengths. This enables it to meet the desideratum of incorporating absolute weights, but it too would still fall foul of the causal subdivision problem. Moreover, it seems to me that all previously suggested definitions of approximate truth are here either potentially vulnerable because they are silent on their exact measure of similarity (most previous ontological approaches), or else actually vulnerable because they are prone generally to the seriousness-of-error problem of which this can be seen as a special case (all other approaches). Fortunately though, there does seem to remain one possible measure that can overcome it, as well as meeting the two earlier desiderata. Moreover, as we shall see, it can also meet various other desiderata we shall impose. Accordingly, we shall select this measure as this thesis's 'official' definition of approximate truth. We outline it now.

Our preferred measure

Instead of the Euclidean distance, consider instead the so-called *Manhattan* or 'city-block' distance. This is defined by travelling from one point to another only in the directions of the orthogonal axes, and never taking the direct diagonal, so to speak, of the Euclidean distance. ([Niiniluoto 1987, p4] notes that both the Manhattan and Euclidean distances are special cases of the more general Minkowski metric. He though nowhere discusses the relative merits of the two, perhaps not surprisingly given that he also nowhere mentions the causal subdivision problem.) Return to our original Earth example, with a truth-point of (3, 100) and model-points of (0, 96) and (3, 0). In this case, the Manhattan distance between (3, 100) and (0, 96) is 7: we must travel 3 units on the first axis, from 0 to 3, and then 4 units on the second, from 96 to 100, making 7 in total. The Manhattan distance between (3, 100) and (3, 0) is the same as the Euclidean one, namely 100 – this is because the straight line between the two points happens to be parallel to the second axis, thus we can go straight up an avenue, so to speak, rather than take a diagonal.

The interesting point is to see what happens when we re-express our vectors using the alternative subdivided description. As we saw, now the truth-point will be (3, 50, 50) and the two model-points (0, 48, 48) and (3, 0, 0). The Manhattan distances will now be:
1) for the Earth-model: 3 + 2 + 2 = 7, in other words exactly the same as before.
2) for the mountain-model: 50 + 50 = 100, again, as desired, exactly as before.

In general, the Manhattan distance is not subject to the causal subdivision problem, which represents its decisive advantage. Clearly, it also meets our two previous desiderata as well, namely treating all axes equally and incorporating absolute as well as relative weightings. We therefore select it as our preferred measure.

Our definition of approximate truth

We can now state our full definition of approximate truth:

1) We must specify our context of interest. This means specifying our ontology and vocabulary, and also exactly which causal strengths are relevant.
2) Thereafter the definition is fully objective. It will apply only to competing context-specific models rather than general theories. The causal strengths specified in step 1 will have some particular true values, and the competing models attempt to match those values (assuming translatability in the case of mistaken ontology). The degree of approximate truth for each model is then just the degree of similarity between its ascribed values and the true ones.
3) This degree of similarity is defined by the Manhattan distance between the model-point and truth-point in the relevant abstract vector space.

More formally, for a given set of causes with true strengths (s1, s2, ...) and a model postulating strengths (m_1, m_2, ...), the model's distance from the truth is:

$$\sum | s_i - m_i |$$, the sum being over i.

Further technical notes

Numerical value

The numerical value of our measure is unrestricted, in that a model-point may be an indefinite Manhattan distance away from a truth-point. The important thing is that for any given context, that is for any given truth-point, we are able to make comparative judgments for different model-points. As discussed earlier (section 2-7), it is questionable anyway how much sense can be made of comparing approximate truth scores from different contexts. Nevertheless, if this was thought desirable then our scores could be normalised by dividing them by the Manhattan distance of the truth-point from the origin. Thus in our example above the truth-point was (3, 100) and our model-points (0, 96) and (3, 0). The Manhattan distance of the truth-point from the origin is 103, so our previous scores for approximate truth could be normalised by dividing through by 103, yielding scores of 7/103 and 100/103 respectively.

Units

If the scores were normalised in this way, then their units would be pure scalars. Un-normalised, they come in units of the effect term. In our system all the causal strengths are defined with respect to the same effect term, so within any one context all models' scores for approximate truth will also be in these same units. Only when comparing between contexts could problems over incompatible units arise, which is why we would then have to normalise as above in order to transform the units into pure unit-free scalars.

Negative causes

It is also straightforward under our scheme to incorporate negative causes. For instance, suppose a chemical reduces the probability of getting cancer, while smoking and asbestos increase it. Then it would be desirable to award that chemical a negative causal strength, compared to the positive strengths of the other two. This is naturally incorporable into our scheme simply by giving the truth-point a negative entry along the axis representing the chemical, after which the rest of our procedure could run as normal. Thus hindering causes are treated just the same as enabling ones.

Infinity of causes

What if there is an infinity of causes to be considered in a problem? To be sure, most of these may be of very little relevance but it would seem cavalier to assume that they are all of literally zero strength. And in any case, perhaps there will be ways of subdividing a causal influence into an infinite number of parts. Either way, it is desirable that our definition still be able to yield finite answers.

Happily, it turns out that this should always be possible whenever the effect term is itself finite. For if the total effect is indeed finite, then by our definition of causal strength the total Manhattan distance of the truth-point from the origin in our vector space must also be finite, and this will be true no matter how we subdivide the causes of that total effect – that is, no matter how many different dimensions there are in our space.

Multiplicative or additive?

Suppose that the true value of a causal strength is 2. On our definition, a model postulating a value of 1 for it would be a distance of 1 from the truth, whereas another model postulating a value of 4 for it would be a distance of 2. Thus the first model would score better for approximate truth. But, it might be argued, should this be so? Perhaps estimating a variable to be twice as high as the true value should be considered the same degree of error as estimating it to be twice as low, in which case the two models should have scored equally? The issue boils down to whether we assess errors multiplicatively or additively. Our definition is committed to the latter.

It seems to me that the intuitions can be argued either way on this issue, and perhaps that neither position is obviously preferable. In any case, it also seems to me that no measure could simultaneously satisfy both, and so we are forced to make a choice. One pertinent point is that I do not know of any multiplicative measure that does not fall foul of our earlier problem of causal subdivision. If this is true generally, and given that the multiplicative approach is not obviously preferable for any other reason either, it would be a strong argument in favour of our additive measure.

Chapter Three – Defining causal strength

3-1) Desiderata

3-2) A unified account

Two kinds of causal strength? Preliminary intuitive definitions. Some technical wrinkles. Final definitions. PM and DM unified. Two examples. Causal strengths in group problems. Desiderata revisited.

3-3) Causal interaction

Introductory example. Causal composition and black boxes. The full credit strategy. An independence requirement. More on background conditions. Explanation. Causal overdetermination. Conclusion.

3-4) Commensurability

Introduction. Commensurability versus separability. Two further examples. Conclusion.

3-5) Bayes nets

3-1) Desiderata

Our definition of approximate truth requires that we also have a good definition of causal strength. Which of two causes of an event is the more important? This question might appear fundamental, but is in fact relatively neglected in the philosophical literature. (Exceptions include [Good 1961], [Sober 1984], [Miller 1987], [Sober 1988], [Sober et al 1992], [Spirtes et al 2000] and [Pearl 2000].) Whereas the general metaphysics of causation has received vast coverage, this subsidiary question of comparing two causes' strengths has not. Yet answering it turns out to be a surprisingly delicate task. Except where mentioned in the text, the details of the analysis of this chapter are not prefigured in any of the literature.

From the preceding chapters we saw that, for our purposes, an account of causal strength must satisfy the following desiderata:
1) The concept should be *univocal*, that is deliver unambiguous results. It turns out that there can seem to exist at least two distinct notions of causal strength, so satisfying this requirement needs more work than might initially have been suspected.
2) Results should be (sufficiently) *objective*.
3) Ideally, results should also be *quantitative*.
4) The definition should be widely *applicable*, and in particular applicable to all situations where we might be judging degrees of approximate truth. For example, the definition should be able to deliver results even in cases of causal interaction.
5) Results should carry some *normative* force.

Note from the start three aspects of our investigation. For our purposes we need be concerned only with calculating the strengths of causes that are already *given*. Therefore: first, we say nothing about the epistemological question of how those causes might best be discovered; and second, nor anything about the venerable issue of how to define causation in the first place. Third, note also that by 'causal strength' we shall have in mind always the strength of a particular (instantiation of a) cause in a particular context with respect to a particular effect. For the purpose of defining approximate truth, there is

no need (on our context-specific account) to attempt to define general causal strengths independent of specific context.

.

3-2) A unified scheme

Two kinds of causal strength?

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

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

Notice immediately that the PM sense of causal strength is unable to distinguish between the case in which Watson is present and the case in which he is not, whereas of course DM can. Hence the two senses are indeed distinct and so may diverge, as in this example. Accordingly, we must investigate whether causal strength really can be given a

univocal understanding after all.

Preliminary intuitive definitions

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

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

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

In our Holmes-Moriarty example, if Watson would have shot Moriarty anyway then Moriarty dies whether or not Holmes fires. So, taking Moriarty's death to be E = 1 and his survival to be E = 0, the DM of Holmes's shot is given by the formula as: E(C & W) –

E(D & W) = 1 – 1 = 0. That is, Holmes's shot indeed made no difference.

The formula is clearly readily extendable to cases where there is more than one counterfactual. For example, suppose that if Holmes had not shot then either Watson would have shot or else Inspector Lestrade would have entered and shot instead (cause 'L'). Each of these possibilities could be given some weighting in the formula and the DM then calculated. For instance, if we used weightings of k_1 and k_2 for Watson and Lestrade respectively, corresponding, say, to the probabilities of them being the ones to fire the alternative shot, and if Watson would have hit and killed Moriarty whereas Lestrade would have missed him, then the DM of Holmes's shot would now be: E(C&W) – k_1.[E(D&W)] – k_2.[E(L&W)] = 1 – k_1. (k_1 and k_2 in this formula are constants, and serve as multiplying coefficients of the effect functions E.) Thus because of the possibility of Moriarty otherwise surviving, Holmes's shot now did make some difference after all, in proportion to the chances of his back-up shot being fired by the errant Lestrade rather than the reliable Watson.

Move on next to our other kind of causal strength, PM. We have just seen how the values our formula yields for DM depend in part on our choice of counterfactual. By contrast, the concept of causal potency, i.e. PM, intuitively seems to be intrinsic and local. Nevertheless, I propose that for our purposes PM is also adequately definable by using this same counterfactual technique. In particular, the potency of a causal input can be defined by reference to *the specific counterfactual of that input being totally absent*, with no other input taking its place. (What exactly could be meant by 'absent' we discuss shortly.) So the 'choice' of counterfactual here is no real choice at all - it is always the possible world exactly the same as the actual one in all respects except that the particular cause in question is absent. For E = effect, C = cause, ~C = absence of that cause, and W = the rest of the world in addition to C, preliminarily define the causal potency of C as follows:

PM of C = E(C & W) – E(~C & W).

This is really quite intuitive when applied to everyday examples. For example, to determine the PM of throwing a brick at a window, we would compare the window with the brick thrown at it ('E(C & W)') with the window with no brick thrown at it ('E(~C & W)'). In words, a cause's PM is just the quantity of its impact on the effect, compared to its absence, and holding all other causes constant. (For the purposes of this thesis, we shall ignore the issue of other possible formulations like using the quotient of the effect terms instead of their difference.)

Some technical wrinkles

What exactly do we mean when citing the 'absence' of a cause C in the right-hand side of our PM formula? It is true that in many cases the interpretation of such an absence will be natural and unproblematic. For instance, if C is throwing a brick at a window then the absence of C would presumably be simply keeping the brick in hand. However, just like effects, causes too may be either an event or a variable, and in the latter case problems arise. For example, suppose the cause of interest is a hot air temperature, say 35 degrees celsius. What would be the 'absence' of such a cause? We could hardly speak of the absence of *any* air temperature, but at the same time there is no immediately obvious fallback point to adopt as our baseline reference temperature. Choosing freezing point, for instance, may sometimes seem odd – if our question was 'how much did the hot day cause me to sweat?', this would seem if anything to imply a contrast with average rather than freezing ambient temperature. Yet on other occasions, such as the query 'how strong an effect does air temperature have on the speed of evaporation of a puddle?', the reference temperature might be freezing point after all. There seems to be no obvious general answer.

In this thesis I propose to follow [Humphreys 1990] on the issue, and to appeal in general to what he terms the *neutral level* of causal input. This he defines (p38), in the case of a variable, as 'the level of the variable at which the property corresponding to that variable is completely absent.' For instance, a neutral level of temperature in our example above would be any temperature at which there was no sweating. A key point is that this

neutral level is likely to depend on the exact effect of interest and on the exact context.

For example, it may be that our focus of interest is the *change* in the level of our variable, in which case the neutral level would of course just be the original level before the change. Or suppose C is ambient air pressure at sea level and we were interested in its effect on the optimal volume of a lung. We might interpret the 'absence' of this C to be a vacuum, but of course a vacuum would definitely not be neutral with respect to this particular effect, since presumably a vacuum would leave no role for a lung at all. Depending on our interest, we might instead want to use for comparison the air pressure at higher or lower altitudes, or perhaps the rate of change at sea level of air pressure with respect to altitude. Again, the point is that it is not enough simply to cite in the formula the 'absence' of the cause C, since the interpretation of this alone will be unclear.

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

A key point which Humphreys stresses is that this neutral level is *objective*. By this he means that, once *given* the specification of our cause, effect and context of interest, the neutral level is then defined objectively. The pragmatics only enter, as it were, in setting the background parameters; after that, the definition of the neutral level follows automatically and unambiguously. In this respect, our definition of causal strength – which will include reference to the neutral level – is objective in exactly the same way as was our definition of approximate truth in chapter 2. So to speak, there is no extra loss of objectivity here.

Finally, turn to two other issues. First, in general the effect term may be an event or a variable. So far, we have been assuming the latter and hence expressing E as a

continuous function. This is fine if, for example, E is air temperature. But suppose that the effect is more naturally thought of as an event that dichotomously either does or does not happen – for instance, if E is getting cancer. In that case, we might want instead to speak in terms of *probabilities* and adjust our notation accordingly. For example, the PM of some carcinogen C with respect to the effect E of getting cancer could be written:

$$p(E|C\&W) - p(E|W)$$

where p denotes a probability function, the probabilities concerned being conditional ones. The formula for DM would be adjusted similarly. An alternative is to keep the same notation as before but to interpret now the effect to be *expected* effect. For notational convenience, we shall adopt this latter policy. Either way, such cases will require the assumption that we are able to make sense of single-case probabilities.

Second, we should be aware of the danger of phenomena such as Simpson's paradox. In particular, if we do not hold fixed all other causal parents of an effect while varying our cause of interest, any results obtained for causal strength may be misleading. This is really just the logic of controlled experiment. (Note though that the introduction of the cause of interest may itself alter the levels of other causes – see the discussion of background conditions W below.) For us, it boils down to a reminder that our definition of causal strength can only be as good as the prior specification of causes on which it is applied. For example, if C is smoking and E is lung cancer, it may still be that C's causal strength is different with respect to one person than with respect to another, perhaps because the two individuals differ in their genetic predispositions or diets. Thus smoking per se will have a great many different causal strengths with respect to lung cancer depending on the exact state of all other causally relevant factors, and it is important that these different strengths are not conflated. (Again, note that the analysis of such cases requires that we be able to make sense of single-case probabilities.) It follows that a causal strength must always be relativised to background conditions, and in particular to the levels of other causes.

Final definitions

Putting together the resolutions of all these wrinkles, we can give final versions of our two definitions of causal strength. Start with PM. Let C be the cause, E be the effect, and for our context of interest let C_0 be the neutral level of C. Let W_1 be the background conditions given C_1, and W_0 be the background conditions given C_0. W_1 includes the levels of all other causes of E. Note that W_1 differs from W_0 since changing from C_0 to C may impact on other aspects of the world apart from E. Given that our definition of PM is comparing the cases of C and C_0, it follows that it must be relativised to the levels of background causes in each case, i.e. to W_1 and W_0. However, we can simplify a little. Recall that – once given a specification of our context of interest – the choice of neutral levels is objective and automatic, so there is no need explicitly to relativise PM to C_0. Furthermore W_0 may be defined in terms of W_1, in particular as just being W_1 save substituting C_0 for C together with all the consequences of that. Therefore, given C, our PM need only be relativised to W_1 – the other necessary relativisations are implicit in these ones. In conclusion, for the case where C is an event and E is a variable, then:

The PM of C with respect to W_1 = $E(C \,\&\, W_1) - E(C_0 \,\&\, W_0)$

Turn next to DM. For this purpose, we define a second cause D, with neutral level D_0. Let W_1 be the background conditions given C and D_0, and W_2 be the background conditions given C_0 and D. W_2 may be defined in terms of W_1, C, D, C_0 and D_0. Then by similar reasoning as with PM above, we arrive at:

The DM of C relative to a second cause D, with respect to W_1
= $E(C \,\&\, D_0 \,\&\, W_1) - E(D \,\&\, C_0 \,\&\, W_2)$

Analogous definitions hold as we vary between events and variables. For instance, if E is still a variable and now so is C, then we should re-label so that C_1 is the actual level of causal variable and C_0 the neutral level. Re-label D as D_1 likewise, and redefine W_1 and W_2 appropriately, i.e. substituting in their definitions C_1 for C and D_1 for D. Finally, let

W_3 be the background conditions given a second non-neutral level C_2. Then we arrive at the following useful definitions:

1) the PM of C_1 with respect to W_1 = $E(C_1 \& W_1) - E(C_0 \& W_0)$

2a) the DM of C_1 relative to a different level C_2 of C, with respect to W_1

= $E(C_1 \& W_1) - E(C_2 \& W_3)$

2b) the DM of C_1 relative to a level D_1 of a second cause D, with respect to W_1

= $E(C_1 \& D_0 \& W_1) - E(D_1 \& C_0 \& W_2)$

If E is an event, we can reinterpret these formulas as expected values of effect in the way described earlier. Other variations, such as DM defined relative to weighted averages over a range of counterfactuals, or cases where some of the causes are events and some variables in different combinations, can be handled in a similar way.

It follows from our definitions above that causal strengths are, so to speak, highly sensitive. They will vary with choice of cause C, of course. They will also vary with choice of effect term E - the same thing may be a strong cause of one effect but a weak one of another. Again, this is obvious. In addition, they will also vary with the values of other causes. For instance, striking a match will cause light if the atmosphere contains sufficient oxygen, but not otherwise. And as mentioned, the PM of C is also relative to the neutral level of both C and the other causes, which themselves may vary with our context of interest. Finally, in the case of DMs, causal strength will of course be relative to our choice of counterfactual, the latter presumably being interest-relative.

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

Clearly, our definitions can yield negative as well as positive values for the causal strengths, but I do not see this as being particularly problematic. In a similar way, there

is no objection to allowing 'negative causation' generally, that is - in [Humphreys 1990]'s terminology - to acknowledging counteracting as well as contributing causes.

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

Strictly speaking, however, this unit-independence will not quite hold always. In particular, if one choice of unit is *non-linear* with respect to the other, problems can arise. For example, suppose our competing units of mass were grams and logarithm-of-grams. Then if one cause was assessed to be twice as potent as another using grams displaced, it would in general *not* be assessed twice as potent using logarithm-of-grams displaced. It may well be that in practice such cases are rare, that is to say that controversies about assignments of relative causal strength only rarely if ever hinge on choice of unit in this way. Nevertheless, there seems no way to rule out the possibility in principle and accordingly if such a case did arise we would indeed be forced to concede that the comparison of causal strengths was choice-of-unit-dependent. Recall though our earlier remarks on interpreting scores for approximate truth absolutely versus relatively (section 2-7). This issue would become rather more serious if choice of unit determined also the *qualitative* ranking of causal strengths. This recalls the classic Miller problem concerning the language-dependence of approximate truth. But as with that before, so long as no units actually in scientific use generate such ranking reversals then so long can the issue legitimately be disregarded (section 1-4).

PM and DM unified

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

variable):

1) the PM of C with respect to W_1 = $E(C \,\&\, W_1) - E(C_0 \,\&\, W_0)$

2) The DM of C relative to a second cause D, with respect to W_1

= $E(C \,\&\, D_0 \,\&\, W_1) - E(D \,\&\, C_0 \,\&\, W_2)$

Notice that, although there is in it no ambiguity or choice about which counterfactual to consider, nevertheless the definition of PM is still in form similar to that of DM – how much difference does the cause make compared to its not being there at all? (Strictly speaking, the only difference between the two is that with PM we set the choice of counterfactual 'D' to be D_0.) This enables us to see now how our two senses of causal strength can be *unified.* In particular, the key insight will be to see a DM as always just the difference between two PMs. Alternatively put, if we take PM to be our core definition of causal strength, then a DM can always be seen as a *relative* causal strength. It would follow that the two senses are analytically unified. Analogously, the existence of both relative and absolute *speed* does not imply that there are really two distinct senses of 'speed'. In the same way as we need only one definition of speed so, I shall argue, we need only one of causal strength.

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

DM of C relative to D

$= E(C \,\&\, D_0 \,\&\, W_1) - E(D \,\&\, C_0 \,\&\, W_2)$

$= [E(C \,\&\, D_0 \,\&\, W_1) - E(C_0 \,\&\, D_0 \,\&\, W_0)] - [E(D \,\&\, C_0 \,\&\, W_2) - E(C_0 \,\&\, D_0 \,\&\, W_0)]$

= [PM of C] – [PM of D]

(Note that these PMs and DM are each defined with respect to W_1 (or W_2 in the case of the PM of D), and moreover that our derivation assumes that the appropriate choices of the neutral levels C_0 and D_0 do not vary. Given that all the causal strengths are taken to refer to the same context here, this latter assumption is justified.)

Thus any DM can always be re-expressed in terms of two PMs. For example, the DM of

Holmes's shot relative to Watson's is equal just to the PM of Holmes's shot minus the PM of Watson's. Notice that, as also in our original definition of DM earlier, the two PMs are each defined with respect to the *absence* of the other cause of interest. This is because the alternative DM – of Holmes's shot given that Watson had fired relative to Watson's shot given that Holmes had fired – would be trivially zero since we would just be comparing $E(C \& D \& W_4)$ with itself (for W_4 = background conditions given C and D). Hence the only DM we would be interested in here is that of Holmes's shot given that Watson had *not* fired relative to Watson's given that Holmes had *not* fired. That means that when defining the DM of C relative to D, we want to compare the case of C *without* D with that of D *without* C, which implies the particular comparison of PMs as per our calculation above.

Note that our formula specifically picks out C and D. Why pick out them rather than any of the other causes of E? Purely pragmatic reasons suffice - we happen to be interested in the PMs of C and D, and in particular in the DM of C with respect to D. The latter implies that we are interested in certain counterfactuals defined in terms of C and D and not any other causes. That is, our choice of DM of interest automatically picks out particular counterfactuals and the interesting thing is that these counterfactuals also happen to be the ones involved when defining particular PMs of those causes.

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

PM of C

$= E(C \& D_0 \& W_1) - E(C_0 \& D_0 \& W_0)$

$= [E(C \& D_0 \& W_1) - E(C_0 \& D_0 \& W_0)] - [E(C_0 \& D_0 \& W_0) - E(C_0 \& D_0 \& W_0)]$

= [PM of C] – [PM of no cause]

= DM of C relative to no cause, by our previous result

('No cause' here is shorthand for 'no cause beyond the neutral level'. Again, these causal strengths are all relativised to W_1, and we assume unvarying neutral levels C_0 and D_0.)

Thus the PM of some C can be described as the PM of C minus the PM of no cause at all, in other words described as a DM. This immediately suggests that perhaps, instead of speaking of DM being cashed out in terms of PM, we could equally well think of things the other way round – namely, of it always being possible to cash out PM in terms of DM. Why should it be PM that is taken to be the more fundamental? For many purposes it will not matter whether we talk in terms of PM or of DM. The important point is that, expressed whichever way, there is only one independent notion in play here. Despite initial appearances to the contrary in the Holmes-Moriarty example, there do not exist two kinds of causal strength; rather, there is really only the one.

Two examples

Imagine a Newtonian particle with a gravitational force on it. In this context, the DM/PM distinction can sometimes seem to collapse; let us see why. What is the potency of gravity here? Let C be gravity, C_0 be the neutral (in this case zero) level of gravity, W_1 be the background conditions given C, and W_0 the background conditions given C_0. Then:

the PM of C with respect to W_1

$= E(C \,\&\, W_1) - E(C_0 \,\&\, W_0)$

= (the particle's motion with gravity) – (the particle's motion with no gravity)

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

DM of C relative to the counterfactual C_0, with respect to W_1

$= E(C \,\&\, W_1) - E(C_0 \,\&\, W_0)$

= (the particle's motion with gravity) – (the particle's motion with no gravity)

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

The two could have diverged if in the DM calculation we had adopted a different choice of counterfactual. Suppose we were comparing the strength of gravity on Earth with that

on the moon. Let the Earth's gravity be C_1 and the moon's gravity C_2, so that the counterfactual C_2 was now some lesser but non-zero alternative level of gravity, corresponding to its strength on the moon. Redefine W_1 now substituting C_1 for C, and let W_2 be the background conditions given C_2. Now the calculation would run:

DM of Earth's gravity C_1 relative to the moon's gravity C_2, with respect to W_1

$= E(C_1 \& W_1) - E(C_2 \& W_2)$

= (the particle's motion with Earth gravity) – (the particle's motion with moon gravity)

$= [E(C_1 \& W_1) - E(C_0 \& W_0)] - [E(C_2 \& W_2) - E(C_0 \& W_0)]$

= [PM of Earth's gravity] – [PM of moon's gravity]

(Again, these PMs would be relativised appropriately, namely to W_1, and we would be assuming unvarying C_0 – in this case zero gravity in all cases.) There are two different questions here: 'how much difference does the Earth's gravity make compared to some other level of gravity?', and 'how much difference does the Earth's gravity make compared to no gravity at all?' The difference between the questions is entirely down to choice of counterfactual - either moon or zero. Often, as in the way we originally set the example up, the implicit choice of counterfactual will be zero anyway, in which case DM and PM will coincide and there will not be even the appearance of ambiguity. Perhaps this is why the issue of causal strength seems so unproblematic in the Newtonian particle case, and indeed in many everyday cases too.

Turn now to a second example (like the first one, adapted from [Sober 1988]) – do genes or environment have the most causal impact on the height of an individual corn plant? On one view, the situation is more problematic here because both genes and environment are necessary inputs for a plant to achieve any height at all, which makes it seem impossible to assign either factor a greater importance than the other. But there is a second way of looking at it too. Suppose that we have a traditional genetic input and the option of a new one (a new plant breed, say), and similarly a traditional and new environmental input too (a new fertiliser, say). Suppose further that switching to the new plant breed increases average plant height by 2cm, but that switching to the new fertiliser

increases it by 5cm. There is now a clear sense in which, with respect to these particular options, switching the environmental input has more causal impact than does switching the genetic one. So overall we have something of a paradox: on the one hand, the causal strengths of genes and environment seem inextricably intertwined and therefore necessarily equal; on the other, it seems that one of them can be seen as more important than the other after all.

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

What of the first sense of causal strength here, according to which genes and environment must be adjudged equal? I claim that this sense is the one captured by our other notion, PM. Now biologists standardly say that to speak of causal potencies here is meaningless, and [Sober 1988] agrees with them. But let us work through our definition. Let C = genes and D = environment, with neutral levels C_0 and D_0. Let W_1 be the background conditions given C and D_0, W_2 the background conditions given C_0 and D, and W_3 the background conditions given C and D. Then:

PM of C with respect to W_3 = $E(C \,\&\, W_3) - E(C_0 \,\&\, W_2)$

= (corn plant's height with both environment and genes) – (corn plant's height with environment but no genetic input)

= (corn plant's actual height) – 0

= corn plant's actual height

And for environment, we get an exactly analogous calculation:

PM of D with respect to W_3 = $E(D \,\&\, W_3) - E(D_0 \,\&\, W_1)$

= (corn plant's height with both genes and environment) – (corn plant's height with genes but no environmental input)

= (corn plant's actual height) – 0

= corn plant's actual height

(We assume in the calculation that each PM is calculated with respect to the presence of the other input. Thus with the PM of C, for instance, the background level is D rather than D_0. If the other input was instead absent, then each of genes and environment would have been awarded zero potency. But either way, our basic point here would still hold.) So our definition implies that: first, genes and environment *each* has maximum causal potency here – the plant's height goes from zero to full with their (individual) presence compared to their absence. Second, therefore each has the *same* degree of PM, as desired.

Intuitively, the second conclusion seems fine. I would also defend the first conclusion – intuitions that each of genes and environment could only have perhaps a causal potency of 'a half' reflect, I suspect, an intuition that the total potencies of two inputs should not add up to more than the total effect. But such an intuition would be misplaced here. These potencies are being calculated individually, i.e. for each input while assuming the other input is already in place. Were we to calculate the 'joint potency', i.e. where C = (genes & environment) in our PM formula, then the joint potency would again just be the plant's actual height, and no more than that. So under no circumstances is any PM ever calculated to be more than the total effect. If there are many jointly necessary causes, then it is surely no weakness of our scheme if any one of those causes when taken individually is found to have maximal potency – *given* that all the other causes are already present. (See also section 3-3 for discussion of this point.)

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

framework. Therefore, given that DM and PM in fact boil down just to the single underlying notion, our suggestion is that a unified understanding across science of the notion of causal strength *is* possible after all.

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

Causal strengths in group problems

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

Start with the case of one smoker, who smokes at a certain level, say 10 cigarettes per day for 30 years. Label this level of causal input C_1, the neutral (here zero) level of smoking C_0, and the effect of cancer E. As before, let the levels of other causes of cancer be represented be incorporated in the background conditions, with W_0 being the background conditions given C_0, and W_1 the background conditions given C_1. Then, as usual, the PM of C with respect to $W_1 = E(C_1 \,\&\, W_1) - E(C_0 \,\&\, W_0)$.

Now imagine a group of two smokers, each member of which smokes at this same level C_1. What will be the total PM of smoking for this group? Clearly, the total expected number of extra cancers in the group is given by the PM for the first smoker plus the PM for the second. But, despite initial intuitions perhaps, this will *not* necessarily come out to be just twice the first individual potency calculated above. For example, it may be that the second smoker has a much better diet than the first one and hence a lower PM of smoking. In that case, the group-PM would be *less* than twice the first smoker's individual-PM. In terms of our formula, although C_1 might be the same for the second smoker, W_1 would not be. What if, on the other hand, the level of all other carcinogenically relevant factors *was* the same for the two smokers after all? In that case, the two would indeed have identical individual-PMs, and the group-PM would indeed just be twice that individual one. Label here such a convenient group 'causally homogeneous'. For our purposes we can, so to speak, treat such groups in the same way as individuals. They thus simplify our calculation, since for any causally homogeneous group we can obtain the group-PM simply by multiplying the individual-PM by the number of group members. On the other hand, if a group is not causally homogeneous – as with the first case of different diets – then the group-PM can after all only be calculated by laboriously adding up all the individual-PMs one by one.

In any real population, of course, we are likely to see many different levels of smoking. If two smokers smoke at different levels, it automatically follows that they cannot be part of the same homogeneous group. In practice, therefore, we are likely more to be interested in the number of causally homogeneous *sub*groups within a population. For each such subgroup, we can calculate an individual-PM. In order to calculate the group-PM, we would then multiply each such individual-PM by the number of members of the subgroup to reach a 'subgroup-PM', and then sum over all the subgroup-PMs in the total population. Thus the score for each subgroup would be appropriately weighted for its relative preponderance in the population as a whole. If no causally homogeneous subgroup had more than one member, then the calculation would just collapse back to summing over all the individual-PMs one by one again.

Either way, we would reach the same final figure for the group-PM. It represents the total effect across a group of a cause C, compared to the neutral level of that cause C_0. In our example, it would correspond to the total number of cancers in the population caused by smoking, that is to the total number given the actual distribution of smoking minus the total number there would have been given zero smoking. ([Sober et al 1992] calls this group-PM 'distribution-dependent' causation and contrasts it with 'potency', by which it means what we have been calling individual-PM.)

In formal terms, suppose that for each individual j within a group we have: an effect E, a level of cause C_{1j} with neutral level C_0, background conditions W_{1j} given C_{1j}, and background conditions W_0 given C_0. Then, for C = this particular *distribution* of the cause in the population (i.e. this particular set of C_{1j}'s):

The group-PM of C with respect to all the W_{1j}
= SUM OVER J OF: $[E(C_{1j} \& W_{1j}) - E(C_0 \& W_0)]$

And equivalently, for each causally homogeneous subgroups j, each representing a proportion K_j of the total population, then with our notation suitably adjusted to refer to the subgroups rather than individuals, we would have:

The group-PM of C with respect to all the W_{1j}
= SUM OVER J OF: $K_j[E(C_{1j} \& W_{1j}) - E(C_0 \& W_0)]$

Note that an exactly analogous analysis applies also to DM, not surprisingly given that as we saw earlier it is just relative PM. Thus suppose that for each individual j within a group we have: an effect E, a level C_{1j} of the first cause with neutral level C_0, and a level D_{1j} of a second cause with neutral level D_0. Let W_{1j} be the background conditions given C_{1j} and D_0, and let W_{2j} be the background conditions given C_0 and D_{1j}. Then we arrive at:

The group-DM of C_1 relative to a second cause D_1, with respect to all the W_{1j}
= SUM OVER J OF: $[E(C_{1j} \& D_0 \& W_{1j}) - E(D_{1j} \& C_0 \& W_{2j})]$

(A definition of group-DM in terms of causally homogeneous subgroups could be formulated in a similar way to before.) For example, if we took C_1 to be the actual distribution of (all kinds of) smoking in a population, and D_1 to be the distribution if all cigarette smokers smoked pipes instead, then this DM would yield the total number of extra cancers in the population that occur due to the smoking of cigarettes instead of pipes. In other words, it would tell us the potential 'benefit' to society of a mass switching from cigarettes to pipes.

In one sense, the group definition is the true general definition of causal strength. We can think now of our individual-PMs before as being merely special cases of group-PM where the total number of individuals (or number of homogeneous subgroups) is equal to one, enabling us exceptionally to calculate a PM in a single go, so to speak. In this way, our approach again yields us a unified analysis – this time of singleton and group cases.

Desiderata revisited

1) *Univocal*. This was the whole thrust of our demonstration that DM and PM boil down to one and the same underlying notion.

2) *Objective*. As stated earlier, once given a specification of our focus of interest, to whit a specification of our cause, effect and background conditions, our definition of causal strength proceeds objectively.

3) *Quantitative*. Our definition is quantitative.

4) *Applicable*. For causal interaction, see section 3-3 in a moment. Already we have shown that the same approach can be applied to singleton and group cases, and also to cases from biology often claimed to be awkward. We show later (section 3-4) that using it we can compare non-trivially the strengths of two causes of any given effect, no matter how apparently incommensurable those causes be.

5) *Normative*. Since we define causal strength to be the quantity of effect caused, it follows that our scores for causal strength have instrumental normative force. For example, to say that (in a particular circumstance) a fertiliser has a causal strength of 5cm

with respect to the height of a plant, implies that the addition of the fertiliser actually *would* increase that plant's height by 5cm.

Our definition of causal strength therefore promises indeed to complete satisfactorily our overall definition of approximate truth. In the rest of this chapter we address the remainder of desideratum 4, plus review briefly how our account of causal strength relates to causation generally, and to work on Bayes nets.

3-3) Causal interaction

Introductory example

Imagine that adding one bag of Green fertiliser increases a plant's height by 2 inches, that adding instead a bag of Blue fertiliser increases it by 4 inches, but that adding both the fertilisers together does not increase the plant's height by 6 inches, as we might expect, but rather by 14 inches. That is, there is a positive interactive effect between the two of an extra 8 inches. In such a case, what is the causal strength of, say, one bag of the Green fertiliser? Intuitively, the issue seems confusing because it is not clear how – or whether – to include the big interactive effect with Blue. Is Green's causal strength 2, or 10, or half of 14, or 2 plus some share of an interaction of 8, or perhaps exactly half as much as Blue's, or maybe the question is just meaningless or under-specified?

We may represent the fertiliser puzzle in the form of a table:

Table – Plant heights and interacting fertilisers

	Blue fertiliser	Nothing

Green fertiliser	14	4
Nothing	2	0

Application of our formula yields the following immediate solution: Green has *two* causal magnitudes, depending on whether Blue is also present. In particular, Green has a magnitude of 10 units when Blue is present and of 2 units when it is not. There is no need to declare the issue intractably confused.

There is also no need to be embarrassed by Green possessing more than one causal strength. We defined causal strength to be a particular and so naturally two different contexts means two different strengths. In a similar way, striking a match leads to light given sufficient atmospheric oxygen but not otherwise. In this example, the notion that there is a single potency for Green is therefore in effect a misplaced attempt to define a single potency across several relevantly different circumstances. I am sceptical whether we can even make sense of a cause having an unchanging 'general' strength or power across contexts (on which see [Northcott 2003a]). But in any case the issue is irrelevant to the points of this section, which all concern causal influence in singleton instances.

Suppose we had had a population of ten plants, five treated with Blue fertiliser and five not. In half the cases therefore, Green would have been awarded a potency of 12, and in the other half only 4. The total potency across this population would have been (5 x 12) + (5 x 4) = 80, or an average of 8. As we saw in section 3-2, it makes perfect sense to define a group-potency like this for a given population. Of course, if we altered the population so that now more than half the plants were treated with Blue, then we would reach some new average potency figure for Green. Hence the result for the group-potency is specific to choice of population, exactly analogously to how individual potency was specific to choice of Blue or non-Blue background conditions. We may define Green's causal strength with respect to a population of plants treated with Blue, with respect to one of plants not treated with Blue, or with respect to a specified mixture of the two. What we cannot meaningfully do is define it independent of any specification

of context at all. For this reason, our table of results above cannot be captured fully by just a single causal strength for each of Green and Blue.

Causal composition and black boxes

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

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

Once given the background conditions, for any given cause C all we need to know is the value of the effect E in its presence compared with the value of E when it is absent.

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

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

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

our potency scores do not reflect this. Second, our formula incorporates the full interactive effects into the potency scores for individual inputs – is this acceptable?

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

But I think the issue in fact boils down merely to confusion about the explanandum (a danger also emphasised in [Sober et al 1992]). Our effect E here was the final output of widgets. For that effect, the worker on his own was indeed unable to produce many widgets, whereas – once the worker was in place – the introduction of the form-signing manager did indeed have a dramatic impact on final widget output. Accordingly, it is desirable that our formula captures this dramatic impact. I suspect the objection here is really more a moral one, and is motivated by the sense that the worker is putting in a lot more labour and physical effort than is the manager, and that this should be recognised. Maybe so, but in that case we should re-specify our effect E to be something like hours of effort or litres of sweat, rather than final output of widgets. Or maybe instead widgets that are built but unauthorised should be awarded a score of 0.8 of a unit or some such. All these modifications of E would indeed tend to yield higher potency scores for the worker and lower ones for the manager. The point here is that our controversy turns out only to lie in the specification of E, and none of the foregoing constitutes a criticism of our definition of causal strength itself. (Perhaps our contrast here between the different specifications of E is an example of the classic one in economics between the labour and exchange theories of value, or more generally is an example of the divergence between moral and economic accounting.)

Of course, none of this is meant to deny that knowledge of underlying causal mechanisms

may often be extremely useful, perhaps essential, methodologically, for instance for working out the likely impact of a cause in *different* circumstances. Moreover, a necessary condition for applying our definition of causal strength in the first place was that we agree on the causal ontology, and presumably knowledge of causal mechanisms is likely to be more than useful for achieving that. So we are not making any claims here about instrumentalism in general. Rather, our claim is much narrower – merely that, once we are agreed on our ontology, then *for the specific purpose of defining causal strength* we need not worry about the underlying mechanism.

The full credit strategy

Turn now to the second objection. A more contentious aspect of the solution is what we might call the decision to award each cause *full credit*. When Blue is already present, the addition of Green indeed yields an extra 10 inches, but intuitively it seems we can decompose this into the 2-inch 'pure' effect of Green acting alone, plus a further 8 inches due to the interaction between Green and Blue. Why, it might be objected, should Green receive full credit for all of these extra 8 inches, rather than 'share' that extra credit with Blue? Similarly Blue is awarded a causal strength of 12 inches in the presence of Green, likewise receiving full credit for the positive interactive effect. The unease is heightened by the following consequence: when acting on the plant together, Blue and Green are now awarded strengths of 12 and 10 respectively – even though this adds up to more than the total effect of 14. How can this be right? How can two components *each* have a strength of more than half the total?

But I suspect that our intuitions here are, so to speak, informed by a naïve additive sensibility. For why take it as *a priori* that component causal strengths *should* always add up to the total? This will be correct of any additive system, true enough, but our starting point was precisely that real-world systems are often *not* additive but instead interactive. In such cases, by definition, the component strengths will therefore not add up to the joint strength.

The fact that many laws of physics, such as conservation of energy, are additive has no bearing on the matter. Energy, for instance, is typically dissipated from a given system, in which case with respect to any *particular* effect it may not be conserved. For example, friction implies that the energy of a moving rock will after a time be less than the energy given it initially by my push. Remember that our formula encompasses the possibility of the change from C_0 to C leading also to a change in background conditions W. In particular, the latter may include any changes resulting from the energy dissipated. Thus conservation of energy is preserved globally even though not in our system of interest. Similarly, no contradiction of the laws of physics was implied by the positive interaction of our fertilisers.

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

In any case, raw intuition is a fickle master. It might be argued intuitively that since Green is necessary for the interactive effect to occur at all, so it should receive full credit for it. Then again, since Green on its own is insufficient for that effect, maybe it actually deserves no credit. All in all, I think the moral is that arguing purely at the intuitive level does not get us very far here. We must appeal instead to the wider logic of the issue.

To that end, note immediately two preliminary points in defence of our full credit strategy. First, the *joint* causal strength of Green and Blue is just equal to the effect with both present compared to that with both absent, in other words is equal to the total effect of 14. Thus no causal strength is ever calculated to be more than the total effect. Second, our definition of causal strength in section 3-2 automatically yields the full credit

strategy, so to the extent that the ideas behind that definition are persuasive so this lends us support here.

But we need to make a more detailed case. For this purpose, I shall take the only feasible alternatives to the full credit strategy to be:

1) The 'pure' strategy. None of an interactive surplus should be added to a cause's individual strength. In our example, this would mean that Green scores 2.
2) The 'sharing' strategy. On this view, a cause should be credited with only a partial share of the interactive surplus. In our example, this would lead to Green receiving (in the interactive case) a score somewhere in between 2 and 10. One advantage of this manoeuvre, and perhaps the motivation for it in the first place, is that – given an appropriate split – now Green's and Blue's individual causal strengths could add up to their joint strength of 14 after all.

I shall argue that neither of these other strategies is defensible, and that for this and other reasons the full credit solution is correct. Green's causal strength in the interactive case is indeed 10.

An independence requirement

Following [Davidson 1980] and many others, in this thesis take causal relations to hold between two events and in particular therefore to be a feature of the world rather than of our descriptions of it. That is, a causal strength is fixed by physical reality. Accordingly, once we are agreed on which two events are cause and effect, I claim that the following *Independence Requirement* must hold:

> Causal strengths are independent of arbitrary re-descriptions of physically identical situations

The idea here is straightforward. For example, the event of my kicking the ball has a certain causal strength, given by the ball's subsequent acceleration. This strength will

remain the same whether we choose then to describe the weather as 'cloudy' or as 'gloomy', or to describe the kick itself as 'strong' or as 'spirited'. Likewise, it will stay the same regardless of pragmatic or external considerations like whether my kick was stronger than my sister's or than my brother's. It may or may not have been stronger than either of these other kicks, but of course that has no bearing on the causal strength of *my* kick.

Turn now to the issue of foreground versus background causes. So far we have been considering only the Green and Blue fertilisers, but of course there are many other causes of the plant's growth too: sunlight, warmth, nutritious soil, water, and so on. When considering the causal strength of Green, say, we have up to now been concerned only with whether it is interacting with Blue and implicitly ignoring the issue of whether it is also interacting with these other causes. That is, Blue has been foregrounded for special attention and these other causes left in the background. Much has been written on the distinction between foreground and background causes. But I think it is uncontroversial that from the point of view of physical reality there is nothing special about the Blue fertiliser over these other causes that mandates us to privilege it in this way. So for our purposes grant all the causes equal status, as it were. The only one that is privileged in our example is Green itself, since by definition when assessing Green's causal strength we are comparing the plant's height with it to that without it. But beyond Green, all the other causes are on a par.

Accordingly, I interpret the Independence Requirement to imply in our fertiliser example that the causal strength of Green is independent of how we arbitrarily demarcate those other causes – that is, independent of which of the background causes we choose to highlight for special attention. We shall compare the case where Blue is foregrounded, as it has been up to now, with the case where by contrast Blue is consigned merely to be one of the background causes and it is the plant's need for nutritious soil that is foregrounded instead. The actual physical reality is identical either way; all that is changing is our labelling. But we shall see that only on the full credit account is causal strength invariant with respect to such arbitrary foreground/background designations. Following either of

the pure or sharing strategies, by contrast, renders Green's causal strength unacceptably mutable.

To demonstrate the point clearly, we shall in our formula have to be precise about the definition of the background conditions W. To begin with, suppose that we are concerned with the causal strength of Green in the presence of Blue. Let W_1 be the background conditions just excluding Blue so as to foreground it, just as we have always been doing up to now. Let W_0 be the background conditions just excluding Blue, only now given the absence of Green. For notational ease, label the addition of one bag of Green fertiliser to be cause G, and of one bag of Blue to be B. Then the causal strength of Green according to our formula is (assuming the neutral level of Green to be just the simple case of its absence):

> E(Green & Blue & background conditions) – E(not-Green & Blue & background conditions)
> = $E(G \,\&\, B \,\&\, W_1) - E(\sim G \,\&\, B \,\&\, W_0)$

Next turn to the case where instead of Blue, the foregrounded cause is now the soil nutrients. Label those nutrients N. Assume that they actually are present, and so were part of W_1 above. We now define W_2 to be the background conditions just excluding these nutrients, and instead including Blue. For simplicity, label still by W_0 the corresponding background conditions given the absence of Green, even though strictly speaking these will not be the same as for the W_0 above. (Assume for simplicity also that the levels of B and N are independent of whether G is present or not.) Then the causal strength of Green according to our formula will then be:

> E(Green & nutrients & background conditions) – E(not-Green & nutrients & background conditions)
> = $E(G \,\&\, N \,\&\, W_2) - E(\sim G \,\&\, N \,\&\, W_0)$

The important point is that these two expressions for Green's causal strength are the *same*. They both compare the effect with and without Green, given that all the other causes are in place. Thus (for our chosen figures) they give the same numerical answer

of 10. If we defined W_3 to be the background conditions just excluding Blue *and* the nutrients (and again disregard irrelevant technical niceties in our specification of W_0), then both formulas alike could be re-written:

$E(G \& B \& N \& W_3) - E(\sim G \& B \& N \& W_0)$

= E(Green & Blue & nutrients & background conditions) – E(not-Green & Blue & nutrients & background conditions)

Therefore, as desired, causal strength as given by the full credit strategy does *not* vary depending simply on which of the background causes we arbitrarily promote to the foreground. But if, by contrast, we *exclude* interactive effects in the manner of the pure strategy, we shall see now that we no longer get this desirable constancy result. The pure strategy, recall, says that Green's causal strength is 2, in other words that it should not include any of the interactive surplus with Blue. Formally, we may write this:

E(Green & not-Blue & background conditions) – E(not-Green & not-Blue & background conditions)

$= E(G \& \sim B \& W_1) - E(\sim G \& \sim B \& W_0)$

= only 2, not 10, on our figures

(Again we shall gloss over non-germane technicalities, thus enabling us to keep the same notation as before for the different background conditions.)

But now turn to the case where it is the soil nutrients, not Blue, that are foregrounded. By exactly parallel reasoning to above, we shall want to exclude from Green's causal strength any interactive effects with the soil nutrients. This will yield:

E(Green & not-nutrients & background conditions) – E(not-Green & not-nutrients & background conditions)

$= E(G \& \sim N \& W_2) - E(\sim G \& \sim N \& W_0)$

The sting in the tail is that this expression for Green's causal strength will in general be *different* to the one of the previous paragraph. With Blue foregrounded, we were considering the impact of adding Green given that Blue was not present but that soil nutrients (as part of the extant background W_1) were. Now with the nutrients foregrounded, we are considering the impact of adding Green given that Blue *is* present

(as part of the extant background W_2) but that soil nutrients are *not*. The difference is illustrated by the fact that, assuming plausibly that soil nutrients are necessary conditions for any plant growth at all, the latter value for Green's causal strength is 0, not 2.

This difference comes out clearly in the formulas if, as before, we let W_3 be the background conditions excluding both Blue and soil nutrients. Under the pure strategy, with Blue foregrounded the causal strength of Green was in effect deemed to be:

$E(G \& \sim B \& N \& W_3) - E(\sim G \& \sim B \& N \& W_0)$

= E(Green & not-Blue & nutrients & background conditions) – E(not-Green & not-Blue & nutrients & background conditions)

But with soil nutrients foregrounded, the causal strength of Green was instead calculated as:

$E(G \& B \& \sim N \& W_3) - E(\sim G \& B \& \sim N \& W_0)$

= E(Green & Blue & not-nutrients & background conditions) – E(not-Green & Blue & not-nutrients & background conditions)

Where in the first formula we have ~B&N, in the second we have instead B&~N. I conclude that the pure strategy does not satisfy the Independence Requirement.

The root of the problem is that the pure strategy disallows the interactive effect with one background cause (i.e. the foregrounded one) while implicitly letting through those with all the others. This suggests that there might be a remedy for such arbitrariness, namely just to disallow interactive effects with all background causes equally. Thus, the argument would run, the pure strategy should stipulate that all causes other than Green be set to their 'normal' non-interacting levels. Here, for instance, this would presumably mean to assume in all calculations that the Blue fertiliser was not present. That way, the Independence Requirement would indeed be satisfied. But we would also be left with a problem, namely the obligation to set *all* causes of E to their 'normal' levels. In our example for instance, we would need to identify the 'normal' levels of soil nutrients, of water, of the farmer's psychology, of planetary distance from the sun... and so on. Such a task would seem daunting indeed, a nightmare of arbitrariness, yet our value for Green's causal strength would be critically dependent upon getting it right.

Does the sharing strategy fare any better? Recall, this stipulated that Green's causal strength incorporate some but not all of its interactive surplus with Blue. But why pick out just Blue here? If the rule is that Green should share only with whatever background cause we choose to put in the foreground, then once again Green's causal strength will be inconstant with respect to that arbitrary choice and so the Independence Requirement be breached again. For instance, if we had foregrounded the soil nutrients instead of Blue, Green would now have been sharing an interactive surplus of 10 rather than 8 units and so in general would have had a different value. (Assume plausibly that soil nutrients are necessary for any plant growth at all. Then, given the presence of Blue in the background, Green without nutrients yields a height of 0 inches, nutrients without Green one of 4 inches, and the two together one of 14 inches. Therefore Green and the nutrients generate an interactive surplus of $14 - (4 + 0) = 10$.) If alternatively the rule is specifically that Green's causal strength should include a share of its interaction with Blue but no other cause, then we must justify such an odd stipulation on objective context-independent grounds. I can think of no such justification. The only remaining solution would be, as with the pure strategy in the previous paragraph, to identify 'normal' levels for *all* the background causes and then presumably to share out their joint interactive surplus universally. But this would again provoke the nightmare of arbitrariness (as well as several other undesirable consequences that we turn to in a moment). The obvious alternative, of course, is simply to set each background cause instead to the level that actually obtains in reality. But to do this while simultaneously avoiding the problem of undesirable inconstancies, any definition of causal strength must include interactive effects fully.

More on background conditions

In many treatments of causation, there is a tendency to neglect background conditions. But here I think that would be a costly mistake – because background conditions are often causes that turn out themselves to be relevant non-additively. For instance, any necessary but insufficient cause of the plant's eventual height implies a small or zero

height given its or Green's absence and full height only if both are present. In other words, as we saw above in the case of soil nutrients, it implies such a background cause's own positive interactive effect with Green. The other background causes we mentioned earlier, namely sunlight, warmth and water, are also each plausibly necessary but insufficient conditions and thus, strictly speaking, themselves compose interactively with Green (with respect to the particular effect of the plant's height). (More generally, the same will be true even in the realm of apparently additive analyses such as those of elementary mechanics, since shielding conditions are implicitly always present in the background even then and are in turn themselves presumably necessary but insufficient for the particular mechanical effects under study. Thus there is a sense in which interactive effects are in fact ubiquitous even in contexts usually thought safe for additive composition.)

These thoughts lend further, somewhat more informal, support to the full credit approach. Return to the sharing strategy. We see now that many background conditions also compose interactively with Green and so presumably should be equally as entitled as Blue to a share of the interactive surplus. Indeed, strictly speaking, the entire total effect of the plant's height is an interactive surplus – between the two fertilisers and all these background conditions. This raises several new difficulties for the sharing strategy.

First, we become vulnerable to intractable counting controversies. Is there really a canonical categorization of every single background cause? For example, are soil nutrients to count only as one cause, entitled presumably only to one unit-share of the interactive surplus? Or ought they to be disaggregated into nitrogen, phosphorus, lack of salinity, air circulation, and so on, all of these disaggregated factors being entitled to unit-shares of the surplus in their own right? The sharing strategy requires that such questions always admit of non-arbitrary answers.

Second, even if some canonical list of causes were available, it is not clear that each of those causes should receive an equal share of the surplus. For example, while the interactive surplus between Green and Blue is 8, that between Green and the soil

nutrients is 10 (see above). Should not this objective asymmetry between Blue and the nutrients be reflected in their shares of the overall surplus? One can imagine further complications in a similar vein. For instance, suppose (varying the story) that Blue and warmth were enough to yield some plant height even without the nutrients, whereas nutrients and warmth without Blue yielded only a smaller height. This would now argue, contrary to the lesson above from their interactions with Green, in favour of the greater importance of Blue over the nutrients. It is not clear that there could exist any single scheme for sharing the overall surplus that satisfactorily resolved all such complexities.

Third, the sharing strategy also leads to unsatisfactory results in everyday examples. For instance, the causal strength of my kick on a ball would not be just the ball's resultant acceleration. Rather, it would only be some (arguably ill-defined – see above) *share* of that acceleration, the rest of the credit being distributed around the many various interacting background causes like ambient air pressure and temperature, the rules of football, the fact that I was free to play that day, and so on. In other words, we would lose our basic notion of the strength of a cause being the size of effect it leads to. (This last difficulty applies also to the pure strategy.)

Explanation

If called on to apportion explanatory responsibility between Green and Blue in the case of the 14-inch plant, we would proceed in a similar way to how we have been defining their causal strengths – how much difference, for instance, did it make that Green was present rather than absent? It is not clear what alternative way of thinking about explanatory credit would be plausible, at least with regard to *causal* explanations. Therefore this section's arguments have force not just with respect to causal influence but also with respect to our explanatory talk, so that on our account we should adopt a full credit strategy for the latter too.

A further strike against the pure strategy as applied to explanations, is that it seems to leave large numbers of events simply unexplained at all. For instance, if – following the

pure strategy in our fertiliser example – we award Green 2 units of explanatory credit and Blue 4, it is hard to see how any plant height above 6 inches, let alone one of 14, has been thereby accounted for. This problem becomes still more acute given the ubiquity of background causes that are necessary but not sufficient for the final effect. Strictly speaking, remember, such background factors are themselves composing interactively with the fertilisers. In these cases therefore, the entire final effect would be the result of interactive effects and so the pure strategy would be forced to conclude that none of the causes receive any explanatory credit at all. In our example Green, Blue, soil nutrients, water and so on would *all* be assigned exactly zero explanatory force with respect to the plant's height. This seems like a *reductio ad absurdum* for any account of apportioning explanatory responsibility. Moreover if it is shown that such background causes can be disaggregated into infinitely many sub-causes, and if shown further that there is no non-arbitrary way of disallowing this, then the sharing strategy would suffer the same fate.

Really, underlying this is a central truth – that at least in matters of explanation we in fact all the time naturally follow a full credit strategy *already*. For example, suppose that I work a pump hard while you work it only softly, and that between us we create the effect of a certain degree of air pressure inside the pump. It seems clear then that my hard pumping explains more of that air pressure than do your softer efforts. Suppose further that my and your pressings on the pump do not interact, i.e. that they compose additively. Nevertheless even here we have seen there would be one further way in which we were still incorporating interaction, since my work on the pump only produced greater air pressure because of interaction in its turn with background causes such as the rigidity and airtightness of the pump's lining, the ambient atmospheric temperature being what it was, and so on. The latter causes on their own do not create any air pressure in the pump beyond the baseline background level, but then neither on its own does my pumping; it is only the joint action of all three that yields the effect of increased air pressure inside the pump. This is therefore an interactive effect for which my pumping was all along being given full credit when, a few lines ago, we agreed that it explained more of the pressure increase than did your softer pumping. That is, even in cases of apparently additive

composition like the pump, our explanations have implicitly always been committed to the full credit strategy already.

To see this point in a different way, consider that perhaps we could have taken your working on the pump for granted, put it among our background causes and concentrated instead on the relative impacts of my work and of the ambient air temperature, contrasting the case of a summer's and winter's day. Which explains more of the increased air pressure inside the pump – my pumping, or that it was summer rather than winter? Again, when before we accepted that my hard pumping was more explanatorily important than your softer efforts, implicitly we were also accepting the inclusion of the interaction effect with the background air temperature. There is no reason this latter decision should suddenly become unacceptable now just because we have arbitrarily designated air temperature rather than your pumping to be the other cause to foreground.

Causal overdetermination

Just as necessary but insufficient background conditions are examples of positive causal interaction, so also are cases the 'other way round' – when causes are sufficient but unnecessary – examples of *negative* interaction. In particular, this is true of when we have multiple sufficient causes of an effect, in other words true of cases of (symmetric) causal overdetermination. That is, technically speaking, causal overdetermination is merely an example of interactive effects at work. To see this more clearly, construct a table for the standard story of how the multiple shots of a firing squad overdetermine a prisoner's death. (For simplicity, assume a squad of only two soldiers.)

Table – Prisoner and the firing squad

	1st soldier shoots	1st soldier does not shoot
2nd soldier shoots	1	1
2nd soldier does not shoot	1	0

Take the shot of each soldier to be a cause, and the state of the prisoner to be the effect. In particular, label the prisoner's death to be an effect of 1 and his survival to be one of 0. Then it is clear that the effect when both causes are present (i.e. 1) is less than the sum of the causes' individual effects (1 + 1), in other words that this is an example of negative interaction. None of the central arguments in this section has turned on the distinction between the interaction being positive or negative, so the full credit strategy is equally mandated for the latter case. Thus the causal strengths here incorporate the interactive effect fully.

I therefore conclude: each soldier's shot has a strength of 0 when the other shoots too and of 1 when the other does not shoot, while the two soldiers' shots together have a joint strength of 1. In particular, in the normal firing squad situation where both soldiers do indeed shoot, the causal strengths of their individual shots are 0 – and not either 1 or any fractional share of 1. If (plausibly) we identify a zero strength with not being a cause at all, then this is equivalent to saying that in the case of the firing squad neither soldier's shot is a cause of the prisoners' death individually even though the two shots together are. Put differently, each shot kills the prisoner if fired alone but not if fired in tandem with the other.

Overdetermination has long been an embarrassment for counterfactual theories of causation, since the above result has been deemed unavoidable and yet simultaneously absurd. Hence the many attempts to evade it, such as the hope that micro-examination might show that 'really' one of the bullets hit the prisoner before the other, thus turning the problem into (the admittedly scarcely more tractable) one of pre-emption. But even perhaps the most sophisticated attempt, this time applying the exciting apparatus of structural equations and causal graphs, yields the same stubbornly absurd result [Hitchcock 2001, p289]. (Admittedly, Hitchcock (following Pearl) does go on to suggest a definition by which the shots would after all be causes individually. But this relies on the somewhat gerrymandered notion of 'weakly active' causal routes, about which even

he himself comments (p290): 'I fully grant that [they are] less intuitive and less well-motivated' than the main concept of 'active' routes that he uses elsewhere.) Thus our result is in itself hardly novel; rather, the novelty lies in the attitude it is recommended we adopt towards it. In particular I do not see it as absurd, nor therefore as an embarrassment to counterfactual theories. It follows that attempts by theorists to evade it are not just unsuccessful but also *unnecessary*.

The prisoner's death does not go unexplained, since it is agreed by all that the two shots together do cause it. Therefore any sense of absurdity we diagnose to be the symptom merely of an unwarranted attachment to thinking additively. In particular, we have no good reason to demand that a cause's strength in cases of interaction be the same as it is otherwise, and plenty of reason to demand the opposite. In a world of interactive effects we must learn to adjust our thinking about the relation between individual and joint causation. Once we do, it becomes apparent that the paradox of overdetermination is really no paradox at all.

Conclusion

I do not think it is immediately clear intuitively how to trace causal influence in cases of interaction, as our initial fertiliser example was designed to illustrate. But I conclude that the correct analysis actually turns out to be straightforward: a factor takes full causal credit for any interactive effects in which it participates. New light is thereby cast on the familiar puzzle of causal overdetermination. Following any alternative strategy to ours leaves causal influence unacceptably inconstant with respect to arbitrary re-descriptions of physically identical situations, as well as raising the prospect both of intractable counting controversies and of unsatisfactory results in everyday examples. Standard explanatory practice also turns out implicitly to have endorsed the full credit strategy all along. Moreover, furnishing explanations any other way would frequently imply an inability to grant positive explanatory credit to any factor at all.

Therefore there is a clear answer as to how a definition of causal strength should handle

interactive effects, and happily all the desiderata our account satisfied earlier remain satisfied now. It follows that our account of approximate truth is safely applicable even to cases of interaction.

3-4) Commensurability

Introduction

Is it possible to compare the strengths of two causes that work in apparently incommensurable ways? To get a sense of this issue, consider the question of which is the stronger cause of cancer, smoking or plutonium [Sober et al 1992]? To be sure, smoking is responsible for more cancers in our society than is plutonium, but on the other hand it is also much more widespread. And whereas just one grain of plutonium lodged in a lung makes cancer virtually certain, even years of heavy smoking makes cancer merely more likely. So there seems to remain a sense in which, although fortunately rare, still plutonium is a stronger (or more 'potent') cause of cancer than is smoking. Is this intuition well founded?

One problem seems to be that it is very hard to get any 'absolute' means of comparison. Although one grain of plutonium seems like a small chunk of cause next to years of heavy smoking, there is no common natural unit by which each can be measured. Is 1g of plutonium equivalent to 1 cigarette, 1 pack of cigarettes, or 20,000 packs? By contrast, if comparing the causal strengths, say, of gravity and electromagnetism, this problem seems to go away since the strength of each can be compared readily using the common notion of 'force'. Likewise, if comparing the strength of my push on a rock with yours, again there seems to be a common natural unit for comparing our two efforts. This suggests that the key to being able to compare the strengths of two causes is that they be *commensurable*. But, we shall argue, this analysis is mistaken since

commensurability turns out to be neither a necessary nor sufficient condition for such comparisons.

Note from the start that since this essay's sense of commensurability is at issue even in cases where the causes themselves are already agreed on, it has nothing to do with the *ontological* sense of commensurability discussed by Kuhn and others. (Note also that, in my view, the true diagnosis of the problem in the smoking-plutonium case actually lies in a mistaken implicit appeal to the notion of a *general* causal strength independent of context [Northcott 2003a].)

To get a further sense of the issue, turn now to an example from earlier – which of genes or environment was more responsible for the height of an individual plant? As noted, biologists have been taught to regard this type of question as meaningless since both genes and environment are necessary inputs for a plant to achieve any quantity of height at all, which makes it seem impossible to assign either factor a greater importance than the other. Perhaps it would be better to describe any such comparisons as *trivial* rather than meaningless – once given the presence of the other, the addition of either genes or environment will each score 'full' causal strength in the sense of leading to a full plant rather than none at all. But still, even on this reasoning, neither factor could ever be assigned a greater strength than the other. [Sober 1988] attributes the root of the problem to the incommensurability of genes and environment. For instance, does switching from one plant breed to another represent a greater or lesser change than switching from one fertiliser to another? There is no general answer because, unlike in the case of the Newtonian particle, there is no common natural unit we can use to equate one chunk of genetic cause with one chunk of environmental cause.

The same issue crops up in many places. For example, a long-running debate in the philosophy of biology concerns the relative importance in evolution of various factors – selection, genetic drift, migration, rate of mutation, and so on. But are these different factors even commensurable? If not, their relative importance (it is claimed) would not be well defined. [Matthen and Ariew 2002, p68], for instance, complain that 'there is no

common currency in which to compare the contributions of [these] different evolutionary "forces".'

[Lewontin 1974, p402] illustrates the general point vividly:

> If two men lay bricks to build a wall, we may quite fairly measure their contributions by counting the number laid by each; but if one mixes the mortar and the other lays the bricks, it would be absurd to measure their relative quantitative contributions by measuring the volumes of bricks and of mortar.

Accordingly [Sober 1988, p312], speaking for many, offers the following conjecture:

> For it to make sense to ask what (or how much) a cause contributes to an effect, the various causes must be commensurable in the way they produce their effects.

But despite its apparent reasonableness this conjecture is wrong.

Commensurability versus separability

All causes are 'commensurable' in that they impact on the same effect. But the claim at issue here is that they need also to be commensurable in the *way* they produce their effects. Thus although the bricklayer and mortar-mixer each contribute to the same final effect (i.e. the wall), the reason the strengths of their contributions are not comparable is seen to be because they contribute in, as it were, incommensurable currencies. There are many causes in the world that are incommensurable in this second sense, so if the conjecture really were true it would represent a serious limitation on, for instance, our ability to compare the impacts of different causal interventions. That is, many choices of intervention are between incommensurable instruments; must we declare all such instruments' efficacies incomparable?

True enough, there does seem to be something distinguishing cases like genes-environment with trivial causal strengths from those like gravity-electricity with

interesting ones. I believe the important factor though is not commensurability; rather, it is marked by what I shall label *separability*. By this term I do not mean merely that two causes are distinguishable (although that too is necessary); rather I mean that their *effects* are (potentially) distinguishable. The key structural feature is really whether at least one of the causes is individually sufficient to produce any quantity of the effect of interest. For example, genes and environment are easily individuated but neither without the other could have produced any quantity of the final effect at all, and this property is symmetric. In our terminology the two are not separable, and it makes no sense awarding them different causal strengths. If, on the other hand – as with gravity, electricity and the Newtonian particle – each cause *is* individually sufficient to produce some effect, i.e. we do have separability, then (and only then) may each be deemed individually responsible for different particular quantities of that effect and hence their strengths indeed be deemed to differ.

[Sober 1988] gives a thought-example of genes and environment each contributing 'height particles' to a plant, and claims that this would enable non-trivial comparisons of causal strength by creating commensurability of genetic and environmental effects. But my view is that these height particles could only achieve that goal in so far as they led to separable impacts on the plant's final height. Their commensurability is irrelevant.

Note that often this whole issue applies, as it were, only to absolute, not to relative, causal strengths. For example, which is more important for producing speech, my brain for thinking of the words or my vocal chords for generating the physical sound? Clearly, both are necessary for producing any speech at all and so in our sense are inseparable. Accordingly, each must be awarded the same absolute causal strength. But comparing my vocal chords when healthy to when they are hoarse, it may well be that my power of speech is a little bit greater – thereby yielding a positive but small strength for healthy relative to hoarse vocal chords. Comparing my powers of speech before and after a major stroke, on the other hand, it may be that the difference is now enormous, indicating a much larger strength for the healthy brain relative to the stroke-damaged one. Thus sometimes relative strengths may be interestingly comparable even when the absolute

ones are not.

To summarise so far: our interest lies in what determines whether, in our terminology, a comparison of causal strengths is trivial. Now, in the Newtonian particle example we have both commensurability and separability, and we get non-trivial comparisons. In the genes-environment example, by contrast, we have neither commensurability nor separability and the comparisons *are* trivial. So neither of these cases is really decisive, since of course they are both consistent with either of commensurability or separability being the key factor. To illustrate that it is indeed separability that matters we shall present two further examples, this time with the two factors diverging.

Two further examples

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

The two lightning bolts are surely commensurable if anything is – they are, after all, two examples of exactly the same phenomenon. But their impacts, *with respect to this effect*, are nevertheless inseparable. Individually, neither triggers the chemical reaction; jointly they do. Therefore, defining their causal strengths in the usual way by how much they

produce of the effect we are interested in, we have to conclude that individually each has zero strength while together they have full strength. This of course is exactly the same situation as in our genes-environment example: on their own, neither genes nor environment can produce any plant while together they produce a full plant. The important point is that commensurability plus inseparability has yielded a case of trivial causal strengths. All the strengths here will be either zero or maximal, and there is no way of saying that the causal strength of one bolt is any different from that of the other.

Note particularly that this analysis holds even though we have specified that one bolt is bigger than the other. Intuitively of course one might assume that the bigger bolt should, as it were, be assigned more of the credit. It is this intuition, perhaps, that motivates an emphasis on commensurability in the first place – since the energies of the two lightning bolts can be directly compared (i.e. are commensurable), *therefore* differential causal strengths can be assigned. But I believe such reasoning is incorrect. Remember, the specific effect we are concerned with here is the chemical reaction, and this is dichotomous – it either occurs or it does not. To be sure, when considering how efficacious they are at producing *other* effects, for instance inducing voltage in a wire, then of course the two lightning bolts may well have different causal strengths. But when considering our particular effect of triggering the chemical reaction, because of the activation energy threshold I do not see how assigning different strengths could be justified. In our *particular* example, that is, the comparison of causal strengths must surely be trivial, even though our two causes are commensurable, and even though their comparison is not trivial in *other* examples. (Observe that the two bolts' separability, our marker for non-trivial comparability, itself varies correspondingly with choice of effect.)

Turn now to the last category of example, this time the other way round from before: namely, with separability but not commensurability. This last example will demonstrate that commensurability, as well as being insufficient for non-trivial comparability, is also unnecessary. Suppose I am taking my dog for a walk on a windy heath and he gets interested in a ball lying in the grass a long way from me. I want him to come back to me, so call out to him. Assume that hearing or seeing my call induces him indeed to

move back to me. Now suppose that at exactly the same moment an especially huge gust of wind blows up. Suppose further that, being only a small dog, this huge gust physically blows him back towards me, independently of any voluntary motion of his. So we now have two independent causes – namely the dog's reaction to my call and the physical gust of wind – each producing the same effect, namely the dog's movement closer to me. Which cause is stronger?

I think we can answer that straightforwardly. The definition of the wind's strength is how much the dog moves given the gust of wind compared to if there had been none. And similarly, the strength of my call is given by how much the dog moves compared to if I had not called. This straightforwardness is a direct result of the easy separability of the two causes' effects. The two strengths could perfectly conceivably differ from each other and in that respect the case is clearly analogous to our Newtonian particle example. But unlike electricity and gravitation in the Newtonian case, the two causes here do *not* seem to be commensurable. My call presumably stimulates some reaction in the dog's brain, and thence voluntary movement. The gust of wind, in contrast, bypasses such mechanisms completely and simply physically pushes the dog's body. How could we define one unit of wind gust and equate it to one unit of call? The two are like Lewontin's bricks and mortar, and there is no analogue to the common role of force in the Newtonian particle case. But despite this lack of commensurability, non-trivial comparisons of causal strength are clearly still possible.

Conclusion

It is perhaps easy to think – and as we have seen often *has* been thought – that the key to comparability of causal strengths lies in those causes being commensurable. But I conclude that this is a mistake, and that the critical factor is actually not commensurability at all but rather that the impacts of the causes be separable. In any case, even if some causal strengths are trivial – as with the PMs in the genes-environment case – still they are anyway definable. Thus our account of approximate truth is safely applicable also to cases of incommensurability.

In passing, we might ask just what significance then should we ever attach to commensurability? It seems to me that people already unproblematically compare the strengths of incommensurable causes all the time in everyday life. For example, which is the quickest route home – left to avoid the multiple traffic lights, or right to avoid the roadworks? Outside of physics, similar remarks surely apply to much of science too. For example, which is the most effective way to speed up a particular chemical reaction – further heating the reagents, or adding a catalyst? Perhaps therefore (my own view) commensurability merely makes comparison of causal strengths particularly obvious and easy sometimes, but is otherwise something of a red herring. Moreover, upon closer inspection some claimed instances of it seem to be more mirage than reality. Electricity and gravity were supposedly commensurable via the common unit of force, for example, but could that force in turn ever be measured *except* via a common effect such as the Newtonian particle's acceleration? If not, such cases collapse to the trivial 'commensurability' common to all two causes of the same effect, and the second sense of commensurability – referring instead to the *way* two causes bring about their effects – seems to melt away.

3-5) Bayes nets

Much sophisticated recent work has analysed causation in terms of directed acyclic graphs, and developed new formal definitions – as it were a new calculus – for it [Pearl 2000] [Spirtes et al 2000]. A characteristic feature is to incorporate from the start probabilistic causation, and once given a causal graph and the probabilities of some nodes on that graph develop techniques for inferring the probabilities of remaining nodes. The central role of conditional probabilities and Bayesian updating in this process has generated the moniker 'Bayes nets' for such graphs. A particular area of application is in

the analysis of raw statistical data from which, given certain assumptions, can be inferred quantitatively the generating causal structure. It is claimed that these methods are superior to many traditional statistical techniques for doing the same thing such as regression analysis. Contained in this program is clearly some treatment of causal strength. How does it relate to ours?

We need to distinguish between two issues: defining causal strength/causal weightings, and defining approximate truth. Start with the former. The conditional probabilities in Bayes nets give the chance of a particular effect occurring given that the related cause does, usually implicitly compared to that cause's absence. In other words, the basic conception of causal strength is the same as in our scheme. The rules for working out the quantitative details – i.e. the probability calculus, essentially – are also in effect the same as ours, notwithstanding the greater complexity of some of their examples. A lot of the effort with Bayes nets is directed instead towards the epistemological issue of inferring what (quantitative) causal relations are present in the first place. This of course is not quite the same as our issue, which is rather the conceptual question of defining what causal strength – i.e. those quantitative causal relations – amounts to in the first place. With regard to the latter task, the Bayes net approach is in fact rather more simplistic than ours since it needs to make several restrictive assumptions in order to derive its impressive epistemological theorems. Most notably, it needs to assume the so-called causal Markov and faithfulness conditions. It also has no explicit treatment of choice of counterfactual (C_0 in our formula), or of many of the other issues covered in this section. In this way, with regard to causal strength it is actually our treatment that is the more general.

Move on now to approximate truth itself. Bayes nets do not yield a natural analysis of this (of course they are not particularly designed to). True enough, if some bit of a network is missing or a particular conditional probability wrongly specified, Bayes nets do provide an apparatus for calculating the impact on subsequent probabilities. But this in itself would *not* be sufficient for judging the impact of such events on a model's approximate truth, for two main reasons – neither of which Bayes nets analyse explicitly.

First, the impact of a missing segment would vary depending on which final effect we are interested in. This is just the sensitivity of any causal strength (and hence approximate truth on our account) to specification of effect. It might be argued back in return that part of the specification of a Bayes net is precisely what final effect we are concerned with, in which case we are just reminded to choose the right Bayes net in the first place. If in addition that Bayes net was equipped with an appropriate rule for converting a set of causal weights into a final score for approximate truth, then we could say that Bayes nets have our notion of approximate truth built in already. But note that working out such a rule was not trivial (see the appendix to chapter 2), and required careful consideration of factors nowhere explicitly discussed by Bayes nets.

In any case, there is also still a second point to be taken into account. As argued in section 2-9 (about interest-relativity), in order to calculate a model's accuracy we also need always to specify exactly which causal strengths we are concerned with in the first place. For example in the Hiroshima case, the Los Alamos project, the course of the American advance on Japan, and the weather that day could all be seen as being part of the same Bayes net leading to the explosion. But different explananda will demand focusing on different areas within this net. Again, it is true that we could just focus on that subsection of the network that contained our causes of interest, and then call only this subsection our 'Bayes net'. But given that the choice of net in the first place is so critical yet is left unmodelled by Bayes net theory itself, and given that Bayes nets offer no definition of causal strength superior to ours anyway, it seems to me that the real philosophical work is being done here by our theory. For our specific purpose of defining approximate truth, that is, Bayes nets do not offer us anything substantively new.

References

-- Adams, E. [1990]. 'Review article: Ilkka Niiniluoto, *Truthlikeness*', *Synthese* 84, pp139-52

-- Aronson, J. [1990]. 'Verisimilitude and type hierarchies', *Philosophical Topics* 18, pp5-28

-- Aronson, J., R. Harre and E. Way [1994]. Realism Rescued: how scientific progress is possible

-- Barnes, E. [1995]. 'Truthlikeness, translation, and approximate causal explanation', *Philosophy of Science*, 62, pp215-26

-- Boyd, R. [1990]. 'Realism, approximate truth, and philosophical method', in C. Savage (ed.) Scientific Theories, pp355-91

-- Brink, C. [1989]. 'Verisimilitude: views and reviews', *History and Philosophy of Logic* 10, pp181-201

-- Davidson, D. [1980]. "Causal relations," in Essays on Actions and Events

-- French, S. and J. Ladyman [1998]. 'Semantic perspectives on idealization in quantum mechanics', in N. Shanks (ed.) Poznan 63 Idealization IX: idealization in contemporary physics, pp51-73

-- Giere, R. [1988]. Explaining Science: a cognitive approach

-- Good, I.J. [1961]. "A causal calculus" parts I and II, *British Journal for the Philosophy of Science* 11: 305-318 and 12: 43-51

-- Goodman, N. [1972]. 'Seven strictures on similarity', in Problems and Projects, pp437-47

-- Gordon, J. M., D. Feuermann, M. Huleihil, S. Mizrahi, R. Shaco-Levy [2003]. 'Fibre optics: surgery by sunlight on live animals' *Nature* 424, p510

-- Guala, F. [2001]. 'Building economic machines: the FCC auctions', *Studies in History and Philosophy of Science* 32(3), pp453-77

-- Hendry, R. and D. Mossley [1999]. Review of [Aronson et al 1994], *British Journal of the Philosophy of Science* 50, pp175-179

-- Hilpinen, R. [1976]. 'Approximate truth and truthlikeness', in M. Przelecki, K. Szaniawski, and R. Wojcicki (eds.), Formal Methods in the Methodology of Empirical Sciences, pp19-42
-- Humphreys, P. [1990]. The Chances of Explanation
-- Kieseppa, I. [1996]. 'On the aim of the theory of verisimilitude', *Synthese* 107, pp421-438
-- Kuhn, T. [1962]. The Structure of Scientific Revolutions
-- Kuipers, T. (ed.) [1987]. What Is Closer-to-the-Truth?
-- Kuipers, T. [1992]. 'Naïve and refined truth approximation', *Synthese* 93, pp299-342
-- Laudan, L. [1981]. 'A confutation of convergent realism', *Philosophy of Science* 48, pp19-49
-- Laudan, L. [1984]. Science and Values: the aims of science and their role in scientific debate
-- Laymon, R. [1982]. 'Scientific realism and the hierarchical counterfactual path from data to theory', PSA 1982, Vol. 1, pp107-121
-- Lewis, D. [1973]. Counterfactuals
-- Lewis, D. [1986]. On the Plurality of Worlds
-- Lewontin, R. [1974]. 'Analysis of variance and analysis of causes', *American Journal of Human Genetics* 26, pp400-411
-- Liu, C. [1999]. 'Approximation, idealization, and laws of nature', *Synthese* 118, pp229-256
-- Maddy, P. [2000]. Naturalism in Mathematics
-- Matthen, M. and A. Ariew [2002]. 'Two ways of thinking about fitness and natural selection' *Journal of Philosophy* 99: 55–83
-- McAfee, P. and J. McMillan [1996]. 'Analyzing the airwaves auction', *Journal of Economic Perspectives* 10(1), pp159-75
-- McKee, J., P. Sciulli, D. Fooce, T. Waite [2004]. 'Forecasting global biodiversity threats associated with human population growth', *Biological Conservation* 115(1), pp161-4
-- McMillan, J. [1994]. 'Selling spectrum rights', *Journal of Economic Perspectives* 8(3), pp145-62

-- McMullin, E. [1987]. 'Review of *Is Science Progressive?*', *Isis* 78, pp200-01
-- Mill, J. S. [1846]. A System of Logic
-- Miller, D. [1974]. 'Popper's qualitative theory of verisimilitude', *British Journal for the Philosophy of Science* 25, pp166-77
-- Miller, D. [1975]. 'The accuracy of predictions', *Synthese* 30, pp159-91
-- Miller, D. [1994]. Critical Rationalism: a restatement and defence
-- Miller, R. [1987]. Fact and Method
-- Morgan, M. and M. Morrison (eds.) [1999]. Models as Mediators: perspectives on natural and social science
-- Mormann, T. [1988]. 'Are all false theories equally false?', *British Journal for the Philosophy of Science* 39, pp505-19
-- Newton-Smith, W. [1981]. The Rationality of Science
-- Niiniluoto, I. [1978]. 'Truthlikeness in first-order languages', in J. Hintikka, I. Niiniluoto, and E. Saarinen (eds.), Essays on Mathematical and Philosophical Logic, pp437-58
-- Niiniluoto, I. [1987]. Truthlikeness
-- Niiniluoto, I. [1998]. 'Verisimilitude: the third period', *British Journal for the Philosophy of Science* 49, pp1-29
-- Northcott, R. [2003a]. 'Can a cause have a potency independent of context?', presentation to University of London graduate philosophy conference, May 2003
-- Northcott, R. [2003b]. 'Causal strength and the analysis of variance', presentation to LSE 'Causality: Metaphysics and Methods' group, July 2003
-- Northcott, R. [2004]. 'Causation as contrastive dependence', manuscript.
-- Oddie, G. [1986]. Likeness to Truth
-- Pearce, D. and V. Rantala [1985]. 'Approximative explanation is deductive-nomological', *Philosophy of Science* 52, pp126-40
-- Pearl, J. [2000]. Causality
-- Platek, S., S. Critton, T. Myers and G. Gallup [2003]. 'Contagious yawning: the role of self-awareness and mental state attribution', *Cognitive Brain Research* 17(2), pp223-227
-- Popper, K. [1963]. Conjectures and Refutations: the growth of scientific knowledge
-- Popper, K. [1972]. Objective Knowledge: an evolutionary approach

-- Psillos, S. [1995]. 'Review of *Realism Rescued*', *International Studies in the Philosophy of Science* 9, pp179-83

-- Psillos, S. [1999]. Scientific Realism: how science tracks truth

-- Putnam, H. [1975]. 'How not to talk about meaning', in Philosophical Papers, Vol. 1, Mathematics, Matter and Method, pp250-69

-- Santer, B., M. Wehner, T. Wigley, R. Sausen, G. Meehl, K. Taylor, C. Ammann, J. Arblaster, W. Washington, J. Boyle, and W. Brüggemann [2003]. 'Contributions of anthropogenic and natural forcing to recent tropopause height changes', *Science* 301, pp479-483

-- Schurz, G. and P. Weingartner [1987]. 'Verisimilitude defined by relevant consequence elements', in T. Kuipers (ed.) What is Closer-to-the-truth?, pp47-77

-- Smith, P. [1998]. 'Approximate truth and dynamical theories', *British Journal for the Philosophy of Science* 49, pp253-77

-- Sneed, J. [1971]. The Logical Structure of Mathematical Physics

-- Straume, T., G. Rugel, A. Marchetti, W. Rьhm, G. Korschinek, J. McAninch, K. Carroll, S. Egbert, T. Faestermann, K. Knie, R. Martinelli, A. Wallner, C. Wallner [2003]. 'Measuring fast neutrons in Hiroshima at distances relevant to atomic-bomb survivors', *Nature* 424, pp539-42

-- Sober, E. [1984]. 'Two concepts of cause', in P.Asquith and P.Kitcher (eds) PSA 1984, pp405-24

-- Sober, E. [1988]. 'Apportioning causal responsibility', *Journal of Philosophy*, pp303-318

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

-- Spirtes, P., C. Glymour and R. Scheines [2000]. Causation, Prediction, and Search (2^{nd} edn)

-- Tichy, P. [1974]. 'On Popper's definition of verisimilitude', *British Journal for the Philosophy of Science* 25, pp155-60

-- Tuomela, R. [1985]. Science, Action and Reality

-- Weston, T. [1987]. 'Approximate truth', *Journal of Philosophical Logic* 16, pp203-27

-- Weston, T. [1992]. 'Approximate truth and scientific realism', *Philosophy of Science* 59, pp53-74