# Lewis's Philosophical Method

Daniel Nolan

Forthcoming in Loewer, B. and Schaffer, J. (eds) *Blackwell Companion to Lewis*. Wiley-Blackwell, Oxford.

Lewis is famous as a contemporary philosophical system-builder. The most obvious way his philosophy exhibited a system was in its content: Lewis's metaphysics, for example, provided answers to many metaphysical puzzles in an integrated way, and there are illuminating connections to be drawn between his general metaphysical views and, for example, his various views about the mind and its place in nature.

A case can be made that Lewis's philosophy also exhibited a systematic methodological approach. I doubt that much of this was self-conscious on Lewis's part, at least at first: my conjecture is that his famous claim "I would have liked to have been a piecemeal, unsystematic philosopher, offering independent proposals on a variety of topics" (Lewis 1983 p ix) would have applied to what he would have liked in the matter of philosophical method: for each question, adopt the method best suited to make progress with it, without requiring that the method necessarily be the same for different philosophical topics.

Even in his earliest philosophical writing, Lewis would not have been an anarchist about appropriate philosophical method. Clarity in presentation seems to have always been something he valued. Lewis, like most analytic philosophers (and probably most philosophers in general) preferred his deductive arguments to be valid rather than invalid. Lewis's work, even from the beginning, tended to proceed through "armchair" methods: by and large, Lewis did not aim to establish conclusions on the basis of detailed empirical investigations using, for example, statistical methods, or complex scientific equipment. It is not that he was averse to citing results of empirical studies in the natural and social sciences: Lewis 1989 cites several discussions of US and Soviet weapons and intentions, for example, even though in the same paper he says "[a]s a philosopher, my business is with the coherence of positions and the range of logical possibilities – not with the truth of empirical hypotheses." (p 53). Lewis 2004 is a paper about "many worlds" interpretations of quantum mechanics, which, while theoretical, is not obviously more so than many papers in theoretical physics, and like papers in theoretical physics is indirectly constrained by scientific discovery. Lewis did not himself publish the results of empirical investigations (Lewis 1995 is an exception), but despite remarks like the one quoted above, he did not see philosophy as concerned purely with a realm of facts inaccessible to empirical investigation.

As Lewis's work developed, his methodological views were articulated more explicitly as well. In the introduction to Lewis 1983 where he provides an overview of some of the strands running through his philosophy to that time, he presents (pp x-xi) an overview of his philosophical method (as of 1981, at any rate). Topics raised in that overview will be the basis of sections 1 and 3 of this chapter, though I will also discuss his distinctive method of philosophical analysis, which is one of the most striking methodological features of his work despite not being mentioned in Lewis 1983 x-xi.

My intention in this chapter is not just to present what seem to me some of the more important methodological themes in Lewis's work, but to offer some of my own suggestions about how Lewis's own methodological stances can be improved. While doing so might risk arriving at methodological recommendations worse than the recommendation to do as Lewis did, it does follow his methodological practice in at least one respect. Lewis would never uncritically recommend a philosopher's philosophical views just on the basis that it was what so-and-so believed: and even though I depart from Lewis's methods in a number of ways, paradoxically this lack of *uncritical* respect seems to me one aspect of Lewis's technique to follow.

This chapter will be divided into three sections. I will begin with "starting points": places where Lewis looked for data for philosophical inquiries. This will be followed by a section on one of the distinctive features of Lewis's technique of philosophical analysis: the project of specifying theoretical roles and then identifying deservers of those roles. (This is one of the best-known pieces of Lewisian method, despite an absence of any discussion of it in Lewis's 1983 remarks on method.) Finally, I will pay attention to the project of weighing costs and benefits, and philosophical theory-selection in this way. I have also written about Lewis's philosophical method in chapter 9 of Nolan 2005: while I will touch on a number of the same themes, I have tried not to repeat material from there more than necessary.

## 1. Starting Points: Science and Common Sense

From Lewis's earliest published articles, Lewis expressed his conviction that science, particularly physics, is a very good way of finding out about the world. In 1966 he said, writing as a philosophical materialist "[a] confidence in the explanatory adequacy of physics is a vital part, but not the whole, of any full-blooded materialism" (Lewis 1983 p 105). In 1994 he expressed a willingness to "optimistically extrapolate the triumph of physics hitherto" to embrace a materialism according to which all the fundamental features of our world are physical (Lewis 1999 p 292). Lewis also employed scientific premises in a number of his philosophical explorations: the existence of single-case objective chances is demonstrated by radioactive decay (Lewis 1986a pp xvi, 83). One proof that not all explanation cites a cause of the phenomenon is provided by the explanation of why some collapsing stars (e.g. those that become white dwarfs) stop collapsing: Pauli's exclusion principle ensures "it's gone as far as it can go" (Lewis 1986a p 222).[1] Lewis 2004 is an extended discussion of what lessons we should draw from Everettian interpretations of quantum mechanics, should they turn out to be best supported by our evidence and theoretical considerations. Lewis was not a philosopher who saw his conclusions as insulated from the findings of science.

Despite this, Lewis has gained somewhat of a reputation for ignoring science, particularly physics, in his metaphysics. (One recent representative quote: (French and McKenzie forthcoming) describe Lewis as "a philosopher who is often pilloried for his lack of engagement with science" (p <2>).) There are a number of reasons some philosophers of physics in particular have an animus against Lewis, but one of the main ones must be his notorious remarks about what quantum mechanics can teach metaphysics:

[1] Lewis did defend a causal theory of explanation: but cases like the collapsing star meant that his theory was in terms of causal information more generally, not just the citing of causes.

> I am not ready to take lessons in ontology from quantum mechanics as it now is. First I must see how it looks when it is purified of instrumentalist frivolity, and dares to say something not just about pointer readings but about the constitution of the world; and when it is purified of doublethinking deviant logic; and—most of all—when it is purified of supernatural tales about the power of the observant mind to make things jump. If, after all that, it still teaches nonlocality, I shall submit willingly to the best of authority. (Lewis 1986a xi)

Lewis's description of physics (appropriately purified) as "the best of authority" has not mollified all physics-friendly readers of this passage. It can be read as dismissive of physics, but in my view a better reading of it is as a (polemic) statement of a preferred philosophical interpretation of quantum mechanics; or at least the view that the best interpretations are not instrumentalist, quantum-logical, or Copenhagen-esque. Partisans of those interpretations might object, but I do not think they should object to this as a case of anti-scientism: rather, it is a case of Lewis taking sides in a live dispute in theoretical physics (a dispute that may seem a little dated now, but those remarks were written in 1984, albeit published in 1986). Respecting the deliverances of science need not require not taking any position in scientific debates themselves: it would be untenable to believe everything scientists say when scientists themselves disagree. (Nor is it clear that we should always trust scientists as having the last word about what is in fact shown by scientific evidence, though of course specialists will often be best placed to make those judgements.)

The other important starting point for Lewis is what I will label "common sense". Lewis thought that our ordinary starting opinions were important constraints on our philosophical theorising, and that straying too far from common sense is methodologically out of bounds. A theory "cannot gain, and it cannot deserve, credence if it disagrees with too much of what we thought before. And much of what we thought before was just common sense." (Lewis 1986b p 134). Common sense starting points are all over Lewis's philosophy: counterexamples to rival analyses are often common sense ones, folk psychology is an important starting point for his philosophy of mind, his theories of causation rely on common sense observations as well as more unusual thought-experiment cases, and so on.

Lewis's respect for common sense and starting opinions takes two major forms. One is that we cannot diverge too far *overall* from "what we thought before": a generalised theoretical conservatism. The other is a more specific attitude to some pieces of common sense: the "Moorean" attitude about some *particular* claims that they are more certain than the premises of any argument to the contrary could be (or perhaps any *philosophical* argument to the contrary).

Lewis uses the expression "Moorean facts" to characterise these particular undeniable claims, and at one point says we "lose our Moorings" if we deny the existence of sensations, simultaneity and values (Lewis 1999 p 418). In doing so, Lewis is alluding to G.E. Moore's famous claim that certain obvious truths (such as the fact there were two hands in front of him) were more obvious than any philosophical premises that yielded conclusions to the contrary. Furthermore, Moore thought, our knowledge of such obvious truths was sufficiently sure to refute scepticism: a theory of knowledge that entailed Moore could not know there were hands in front of us is refuted, since he knew there were hands in front of him (Moore 1939).

There are a number of different things that Lewis counts as "Moorean facts" in different places in his work, and many of these are parts of apparently secure common-sense. That there is "apparent sameness of type", for example that things sometimes are the same shape as other things, is a Moorean fact that any theory of properties and relations should respect. (Lewis 1999 p 20). That many of our ordinary beliefs about colours are close to true (Lewis 1999 p 333) is Moorean. That our language has a fairly determinate interpretation (Lewis 1999 p 47) is Moorean. That we know a lot (Lewis 1999 p 418). Many of Lewis's philosophical investigations start from starting points like these: not negotiable pieces of our ordinary picture of the world.

Why think any of our beliefs have this special status of not being up for grabs? Lewis never says explicitly, and a number of his characterisations of Moorean facts are ones that rightly apply to *any* known facts. (See Nolan 2005 p 208). Perhaps it is because of a limited faith in the powers of philosophy: Lewis at one point makes fun of those who would use philosophy to overturn apparently secure knowledge: in this case, those who would rely on philosophical arguments that standard mathematics is full of falsehoods.

> If they challenge your credentials, will you boast of philosophy's other great discoveries: that motion is impossible, that a Being than which no greater can be conceived cannot be conceived not to exist, that it is unthinkable that anything exists outside the mind, that time is unreal, that no theory has ever been made probable by the evidence (but on the other hand that an empirically ideal theory cannot possibly be false), that it is a wide-open scientific question whether anyone has ever believed anything, and so on, and on, *ad nauseum*? Not me! (Lewis 1991 pp 58-9)

Lewis did not think these "discoveries" were genuine, of course, but his point presumably is that philosophical argument has often led us astray, and badly astray at that. A look at the track record of philosophical argument should make us dubious that it can be used to overthrow apparently very secure common-sense opinions. (Or secure deliverances from sciences like mathematics, for that matter.) Or so Lewis's suggestion seems to go.

Some general confidence in our epistemological abilities seems reasonable, especially when beliefs seem to be part of generally successful epistemic projects. I am a reasonably good detector of colour, of tables and chairs, of whether the Earth is round, or whether plants are alive, among a host of other questions. To the extent I think I know many such things, I should believe that it will not turn out that I was wrong. Perhaps I even should have an expectation, about each thing I think I know, that I will not turn out to be wrong about *that*. But Moorean confidence seems stronger than this. It seems to be strong enough to put one is a position to dismiss any arguments or evidence to the contrary. Many of the things I know are not like this: I know that none of my colleagues are Kierkegaard specialists, for example, but I could be convinced otherwise (perhaps merely on people's say-so, perhaps by a bit of internet searching that discovers my colleague's seminal books on Kierkegaard).

I doubt that there is anything very interesting that is so secure that *any* argument to the contrary is rightly dismissed by mere modus tollens. Those that think that philosophy is a particularly weak source of reasons to believe might think that, at any rate, there is a body of ordinary knowledge that should be immune to *philosophical* challenge (it would be "presumptuous", to use Lewis's expression for the envisaged philosopher who disagrees with mathematicians). But philosophers do not seem particularly limited in the methods we use: philosophers who wants to challenge our ordinary colour judgements in part by appeal to colour science, or philosophers who wants to challenge our ordinary temporal judgements by appeal to cutting-edge physics, do not seem to be doing anything outside the bounds of philosophy. So if a challenge can be launched at all against apparently well-grounded doctrines in common sense or science, then it is hard to see how philosophers could not legitimately launch such challenges.

While Lewis's view that there are near-unchallengeable Moorean facts is distinctive and clearly relevant to his method, my view is that not much would need to be done differently in pursuing Lewis's own projects, or Lewis-style philosophical projects, if this commitment to Moorean facts was watered down. Even if no claim was immune to philosophical challenge, still we might be reluctant to challenge beliefs that we held strongly and which were apparently part of successful common-sense pictures of the world. Even if we thought that in principle we could give up the view that there are colours in the world and they are sometimes properties of surfaces of objects, it would certainly make sense to *first* see if there was a satisfactory theory of colour that preserved that opinion. One could think that it counts very heavily against a theory of knowledge that it cannot allow e.g. that I know that have hands, without needing to insist that "I know that I have hands" is a non-negotiable constraint on any theory of knowledge. Giving common sense significant weight but not maximum weight will make it secure enough to allow us to choose between philosophical theories on the basis of how well they respect key pieces of common sense, whether or not we count any of common sense as Moorean.

The methodological question of whether we should give any part of commonsense a lot of weight in our philosophising remains, even if we do not go as far as Lewisian Mooreanism. I am inclined to think that we should, at least in the early stages of our inquiry, and at least for parts of commonsense that appear to be well supported. (We should be more confident in the existence of cats than that extraterrestrials will seem weird to us, even if both are "commonsense"). Not to do so would be as strange as starting geology without the assumption that there are rock formations, or chemistry without the assumption that fire can change substances put in it. (The call to ignore ordinary opinion in philosophy often seems to come from those who want to make philosophy more like the sciences, but I do not think a reconstruction of the epistemic position of the sciences is adequate if we ignore the common-sense bases of those sciences' starting points.) That is not to say that we might not eventually give up much of (current) common sense; but it is not clear that specifically philosophical investigation is in a good enough position to do anything like that yet. Still, the role of common sense in philosophy and inquiry generally is too large an issue to try to resolve here.

One reason not to privilege common sense *per se* is that many of my other beliefs also seem very well supported and resistant to change on philosophical grounds: why should it matter very much whether such beliefs are *common*? Perhaps we had better take our ordinary opinions as constraints on our theorising, not so much because of any special status they have in themselves, but simply because they are among our initial beliefs: perhaps it is *our initial beliefs* that ought to have a special role to play in philosophical inquiry. "Theoretical conservatism" is an expression that is used for a variety of approaches, but they share the view that one ought to continue to hold one's starting beliefs until one has good reason to change them. (Whether this is just because they are the starting beliefs, or whether it is because there is reason to suppose one's starting beliefs are in the main warranted, is one matter that defenders of conservatism may disagree on.) Lewis thought that we had no reasonable choice but to respect our starting beliefs:

> It's not that the folk know in their blood what the high falutin' philosophers may forget. And it's not that common sense speaks with some infallible faculty of 'intuition'. It's just that theoretical conservatism is the only sensible policy for theorists of limited powers, who are duly modest about what they could accomplish after a fresh start. Part of this conservatism is a reluctance to accept theories that fly in the face of common sense. But it's a matter of balance and judgement. (Lewis 1986b 134)

Despite the reference to common sense, the "only sensible policy" here seems to be to not lightly tamper with one's starting point: to be modest about what could be accomplished after a fresh start. Lewis never explicitly discusses what one ought to do if one were to come to believe that one's own starting beliefs were idiosyncratic and not "common sense", but theoretical conservatism would likely take as dim a view as a wholesale revision to match the beliefs of others as any other wholesale revision.

The version of theoretical conservatism that Lewis advocates appears to be one that does not allow us to move beyond a certain point from our starting opinions. He says:

> What credence [a theory] cannot earn, it must inherit. It is far beyond our power to weave a brand new fabric of adequate theory *ex nihilo*, so we must perforce conserve the one we've got. A worthwhile theory must be credible, and a credible theory must be conservative. It cannot gain, and it cannot deserve, credence if it disagrees with too much of what we thought before. And much of what we thought before was just commonsense.

It is not entirely obvious when "before" is here, but it seems to me to be something like "before we started theorising". That is when we can be most certain that most of what we thought was common sense. Notice again that it seems to be common sense's role as our starting opinion that makes it fit for belief here: this passage again suggests that it is common sense's role as our initial opinion, plus the imperative to conserve what we have, that explains why philosophical views should not depart too far from common sense.

Lewis had another particular reason to embrace conservatism, given his other philosophical commitments. Lewis's favourite story about belief revision was a Bayesian one, according to which an ideally rational agent starts with a distribution of probabilities over all the possible contents of belief: these are the agent's rational degrees of belief in each of those options (and in the case of the entirely rational agent, her rational degrees of belief will match the degrees of belief she has as a psychological matter). Equipped with these degrees of belief, the agent then *updates* on new information that comes in, for example by the senses. Lewis's preferred account of updating, for the entirely rational agent, is by *conditionalising* (Lewis 1999 403-407). The agent learns some perceptual information with certainty (probability 1), and then her other degrees of belief change in light of the information learned: when she learns A, the new subjective degree of belief she assigns to a proposition B is equal to the old value her belief system applied to the conditional probability of B on A (P(B/A)).

Lewis was prepared to concede that agents of more limited powers than the ideally rational agent might legitimately update with a less demanding rule, e.g. that of Jeffrey conditionalising (Jeffrey 1983): in effect, limited agents like us may be legitimately uncertain of what we have learned, and so not give any direct deliverance of our senses a probability of 1 on a given occasion. And presumably limited agents like us depart from ideal rationality in other ways as well: perhaps our beliefs are sometimes probabilistically inconsistent from time to time (e.g. when we believe a number of propositions separately that are jointly inconsistent: see Lewis 1998 97-110). Across time, our degrees of belief might sometimes shift without relevant evidence, as when we become aware of new hypotheses or are affected by wishful thinking or suffer imperfect recollection of our past opinions. Even when we accept that the theory of the rational Bayesian agent is an imperfect model of what beings with our powers can realistically aspire to, one feature of it in particular suggests that even beings like us ought to be theoretically conservative.

It is part of the standard Bayesian model that there is a fair amount of leeway allowed in the starting rational degrees of belief of an agent (their "prior credences"). Agents will start out almost certain of some contingent propositions, for example (especially when a model has infinitely many propositions in it). And typical models of Bayesian agents over time explain the rationality of later states in terms of how the agent got there from earlier states: for a defender of conditonalisation like Lewis, the agent's credences *now* have to be produced by conditionalisation from her states *before*, if there is a change. Being automatically rational in having initial opinionated states not derived from anything, together with the rationality of later states being largely derived from how they relate to states had before, suggests a picture where the mere having of the earlier states, together with the evidence that has come in as input, justify the later states. (The core of Bayesianism talks of rational degrees of belief rather than "justification", but this is one obvious way to try to map talk of justification into a Bayesian framework.)

Admittedly, a Bayesian formal framework is compatible with shifts in credence very unlike what we would expect conservatism to endorse. An iconoclast who radically changes most of her unconditional degrees of belief each time she receives a piece of evidence can be modelled without violating any formal constraints (the conditional probability of B on A can be wildly different from the unconditional probability of B). But the spirit that motivates Bayesianism can motivate more conservatism than the letter of Bayesianism entails. *Plausible* conditional probabilities of ordinary sensory input will often leave much of what is believed largely intact: I see my cup on the desk, and that should normally not radically change my opinion about the Mongol empire or the existence of isobutane. If this hunch about the rational conditional degrees of belief of creatures like us is correct, then we get a picture where my degrees of belief now are not only rationally determined by my degrees of belief before, but that, rationally, my unconditional beliefs will tend to remain largely unchanged until evidence impacts on them relatively directly. If this is the right picture of rational agents, it suggests the rational agent will be theoretically conservative.

My suspicion is that this is not the only point of contact between Lewis's broadly Bayesian views about rational belief revision, on the one hand, and his views about philosophical method (and appropriate belief revision in philosophy), on the other. Another obvious point of contact is between, on the one hand, the Bayesian sympathy for equally rational prior probabilities yielding different but equally rational reactions to evidence, and on the other hand, Lewis's conviction that rational philosophers presented with the same philosophical considerations can rationally disagree about their conclusions (see below in section 3). Since Lewis never explicitly connected the topics of Bayesianism and philosophical method, drawing connections requires a certain amount of reconstruction, and so the task of drawing out other links is best left to another occasion.

I also have some sympathies for theoretical conservatism, and agree that there is a danger in philosophy of too quickly shifting to radical or exciting views that abandon too much of what we thought before on a topic. (Even those who oppose philosophical conservatism should think this is *one* of the dangers, just as even philosophical reactionaries should concede that *one* of the dangers is not shifting one's philosophical position enough in the face of arguments and evidence.) Lewis's version seems to me unduly strong. Contrast Lewis's view, that we cannot (rationally) end up disagreeing with our starting point beyond some limit, with *stepwise* conservatism. According to stepwise conservatism, each transition we make cannot depart too far from our initial network of beliefs. But once we have made some rational transitions because of evidence or argument, when assessing whether a further transition is rational we consult our *current* opinions, rather than evaluating the envisaged transition partly on the basis of where we began our theorising. Maybe a new theory would have struck us as incomprehensibly radical twenty years ago. But if it seems like a measured response to our evidence and argument now, why should it worry us if "it disagrees with too much of what we thought" twenty years ago?

If one is a Lewisian conservative (as I have interpreted his view, at least) rather than a stepwise conservative, one faces an awkward challenge of explaining opinion change through time *across* individuals. Suppose there is some set maximum "doxastic distance" which someone can reasonably travel before one "disagrees too much" with what they thought before. Adam begins rational enquiry with beliefs $B_1$, and after a life of unexpected surprises and philosophical reflection rationally ends up with beliefs $B_2$, near the edge of the maximum distance from $B_1$. Adam raises his daughter, Belle, with a $B_2$ world-view, and by the end of her life of inquiry she (rationally) reaches $B_3$, near the maximum distance from $B_2$ and almost twice the maximum distance from $B_1$. If this is coherent, a puzzle arises.

Consider Adam's very long lived brother Carl. He also begins at $B_1$, lives the life of the mind with Adam until he dies, and then follows Belle's fascinating work thereafter. Presented with the same lines of argument and evidence as Adam and then Belle, he also ends up at $B_3$. But why should we have to judge Carl's overall trajectory irrational when Adam's and Belle's were both rational? Why would Belle, but not Carl, be allowed to take note of Belle's excellent evidence and argument to the conclusion $B_3$ from their shared starting point $B_2$? If we erased Carl's beliefs around the time of Adam's death and "restarted" him at $B_2$, then would he have been allowed to revise onwards to $B_3$? Rejecting stepwise conservatism seems to leave us having to make an invidious and undermotivated distinction between Carl and Belle once Belle starts questioning their shared $B_2$ views.

There are various ways a Lewisian conservative could resist this case if she wished to say that Belle and Carl are restricted to roughly the same theoretical options when Belle begins her independent inquiry. Perhaps I have been unreasonable in thinking that anything like $B_2$ is a fit place to *begin* inquiry. If we all started in roughly similar "pre-theoretic" states of nature, for example, Adam might already have had to move Belle a long way from her starting point to get her to $B_2$. Or perhaps Belle is only permitted to move so far from $B_1$ because she *lacks* a lot of reasons and evidence that those who have been to $B_1$ possess: maybe there must be some epistemic *advantage* Adam and Carl have that she lacks (despite holding Adam and Carl's views at the beginning of her inquiry), and it is this advantage Carl has that makes it irrational to follow Belle's course. Here is perhaps not the place to consider every feasible response a Lewisian conservative might offer, though any of these would need significant fleshing out if we are to be convinced that Belle may go where Carl cannot after they both endorse $B_2$.

There are other uncomfortable questions that face the Lewisian conservative. We unfreeze Oog from the ice she has been trapped in since the Pleistocene. How close may we bring her to contemporary opinions in physics, economics or philosophy through evidence and argument, if we want her to be able to rationally believe things we know? Abstracting from practical limitations she might have, we might imagine that eventually we can teach her the latest about fundamental physics or economics or philosophical methodology even though the starting point for her enquiry might be considerably further back than those of us brought up with contemporary assumptions about how the world works. The stepwise conservative, on the other hand, may need to concede that Oog may not be able to jump straight from flint-knapping to quantum mechanics or from tribal ceremonies to the theory of the liberal state, but can at least hold out the prospect that in principle she can learn the things we know by stages, maybe via imperfect approximations of what we take to be the truth (Newtonian physics, Locke on the state, etc.).

Even stepwise conservatism requires defence, of course: one might even think that it is rationally permitted to go through a "conversion experience" and move all of one's views at once to something radically different. (See van Fraassen 1989 chapter 7.) Unfortunately a discussion of theoretical conservatism quickly raises many of the most fundamental issues in epistemology and belief revision. Whether supported by good arguments for theoretical conservatism or not, Lewis adopts common sense as a starting point as well as the deliverances of mature natural sciences. But the starting point is not the end point: armed with the starting points, we can turn to trying to solve philosophical problems.

## 2. After the Starting Points: Defining Theoretical Roles, Finding Deservers

A particularly influential part of Lewis's method concerns a way of approaching philosophical puzzles. Questions as diverse as "what is pain?", "what is meaning?", "what are properties?" and "what is moral value?" were all, at one time or another, treated in the same general way. One aspect of this approach is now called the "Ramsey/Carnap/Lewis treatment of theoretical terms", and another might be called a kind of "generalised functionalism". It has been the inspiration for philosophical movements like the "Canberra Plan" (see Braddon-Mitchell and Nola 2009), and its influence can be seen in scores or hundreds of papers in the current philosophical literature. Since this method has been extensively discussed elsewhere, <including in another chapter of this volume>, a brief sketch here together with some critical remarks might suffice.[2]

First, one assembles a theory featuring a term apparently referring to the phenomenon of interest. It may be as easy as taking a canonical theory from a science (as we might do if we were interested in "what are electrons?" or "what is an ecosystem?") or if the term is one in widespread unsystematic use, we might need to articulate a "folk" theory of the phenomenon of interest. For example, if we want to define pain or belief, we are to use "common sense psychology" as the relevant theory (Lewis 1999 249). Then one uses that theory to define a "role" associated with the phenomenon. If the expression in the theory we are interested in is a predicate, we first transform the theory to talk about the associated property. (Instead of "when people are in pain they tend to cry out" we might use "when people have the property *pain* they tend to cry out", for example.) Then when the theory is modified so that the expressions of interest are all nouns, we can create a "matrix" by replacing the terms of interest with variables, a different variable associated with each term of interest. ("when people have *x* they tend to cry out").

[2] Lewis discusses this method explicitly in Lewis 1983 78-95 and Lewis 1999 248-261. I discuss this aspect of Lewis's method not only in Nolan 2005 213-227, but in Nolan 2009.

One thing we can do here is use a theory to simultaneously provide a matrix for a number of different terms and so provide a role for a number of things at once: the role of belief might be specified partly in terms of the role of desire and the role of intention, if it is part of our folk theory that beliefs and desires go together to produce intentions. The aim is to provide a matrix where a range of puzzling vocabulary is replaced with variables, but enough else is said using expressions that have not been replaced to give the matrix substantial content. (We might offer a matrix replacing *all* the psychological terms in a theory of mental states, for example, leaving only causal vocabulary and specifications of non-mental inputs and outputs, such as perceived objects on the input side and behaviour on the output side.)

Armed with this matrix, we can treat it as being associated with a *role*. That *role* is satisfied by a collection of entities (properties, events, "things" or whatever) when those entities collectively satisfy the matrix. (To use the same kind of example Lewis uses in Lewis 1999 249-252, if our matrix is "X killed Y in the kitchen of the mansion using Z", and Anne killed Bertrand in the kitchen with Excalibur, then Anne, Bertrand and Excalibur collectively *satisfy* our matrix.) We call the entities that jointly satisfy the matrix the *realisers* of the role associated with the matrix.

Lewis did not think that an expression had to pick out something that satisfied *every* clause of such matrices: even canonical theories might be wrong in some detail. To count as the property *pain*, or the property *red*, or the property *morally right action* the relevant properties have to play *enough* of the relevant role, and to play more than any rival properties. In early applications, Lewis emphasised the construction of *causal* roles: roles specified largely in what the typical causes and effects of entities and states are. While that seems appropriate for some philosophical projects, it does not seem crucial to the method, and applications of the method to e.g. mathematical objects, or possibilities, or values, seems a feasible thing to try even if we doubt the roles taken from available theories will have much to do with that sets or possibilities or values *cause*.

The methodological point of extracting roles from theories and talking about potential realisers that satisfy all, or most, of those roles, can sometimes be a little obscured by the framework of variables, matrices, definitions and so on (particularly because usual presentations often talk about “Ramsey sentences”, “Carnap sentences”, strings of quantifiers, and other technical devices). One methodological purpose of this focus on roles and what realise them is to try to extract *definitions* that can provide analytic truths about the topics of interest. For example, we can attempt to define “pain” as that which best satisfies enough of the pain role, “value” as that which best satisfies enough of the value role, and so on. Providing definitions and an account of the analytic/synthetic distinction was one of Carnap’s main interests in using this sort of machinery (see Carnap 1963 pp 958-66). Using this method to extract plausible definitions is not entirely straightforward, however, especially once the issue of deciding which theory is canonical is taken seriously. (Nolan 2005 p 219-222 expresses some preliminary worries.) However, I think another aspect of employing this method is where the more important methodological implications lie.

Constructing a role from an entire theory and looking for entities that play that role tends to do two things. One is that it shifts the focus from trying to isolate a crucial core to a philosophical concept to paying attention to many theoretical connections: instead of trying to work out what the special mark of belief is, for example, we look for what it is that is connected in the right way to perception, desire, action, and so on. This style of extracting a characterisation of something from an entire theory of it is sometimes called “network analysis” (e.g. Smith 1994 pp 44-56), because of this feature, and because of the way it can be applied to a family of puzzling phenomena at once: all of the mental in terms of causal relations with the physical, or all of the semantic with linguistic behaviour, for example. Instead of trying to isolate a balance between necessary and sufficient conditions, for example, assembling many indicators and looking for what satisfies most of them allows a theorist to not have to immediately defend some particular criterion or two as all-or-nothing features of the phenomenon under discussion.

The second tendency of this way of approaching philosophical analysis is that, in the first instance, it focuses less on the *natures* of entities than the relationships between those entities and other phenomena (hopefully some of which are better understood). For example, instead of trying to reflect directly on the nature of moral value, for example, Lewis’s style of analysis encourages a theorist to articulate connections between moral values and other matters: rational behaviour, desires, moral obligation, and so on. The focus is more on what something of interest *does* than, in the first instance, on what it *is*. This might be literally true, when an analysis in terms of a causal network is offered (for beliefs, for example, or for colours), or more metaphorically true (when giving an account of propositions, or of moral values, or of numbers), when the theoretical role is unlikely to feature much about causation.

Both of these shifts have been found useful and liberating by philosophical inquirers. Another advantage, it seems to me, is that questions of which part of the theory are necessary and which contingent, and questions of which conceptual truths and which mere synthetic ones, do not need to take centre stage in this style of philosophical analysis. As long as we have a *true* theory concerning an entity of philosophical interest, our target will be among the potential realisers of the role we extract from that theory; and if it is plausible that only one thing plays the relevant role, then it will be plausible that that is, indeed, our target. If I tell you a cluster of truths about the (surface) colour scarlet, and a particular physical property of surfaces is the unique property that satisfies the role defined by those truths, then that property must be scarlet, even if the truths we used to define the role were contingent and synthetic. After all, if only one thing does what the colour scarlet does, it is the colour scarlet.[3] Many who deploy the Lewisian method of analysis hope to establish analytic or conceptual truths with it, and the method does not preclude doing so; but those inclined to leave the issues of analyticity and conceptual truth to one side can still employ the method to derive interesting philosophical conclusions.

It should be conceded that Lewis's suggestion about philosophical analysis is far from the only way to gain the benefits of looking at many theoretical connections rather than few, and of looking at what an entity is supposed to do rather than, in the first instance, what the nature of that entity is. It should also be noted that one can offer something in the form of a Ramsey-Carnap-Lewis analysis, complete with theoretical role and identification of theoretical realiser, while engaging in a project with neither of these features. "x is justified, true belief" is a matrix we could associate with knowledge, if we wished, and then we could find a state that played that role (primitive knowledge, for example, or beliefs that were both justified and true).

[3] I make this point in Nolan 2009 pp 280-282.

Unless we went on to do something more substantive with roles for justification or truth or belief, that would not bring some new advantages to a traditional JTB analysis of knowledge. Employing Lewis's method of philosophical analysis does not *necessarily* bring the features with it that have proved fruitful. But philosophers thinking about philosophical analysis in Lewisian terms are at least nudged in the direction of networks and functional roles; and this nudge has been enough to yield a recognisable family of philosophical theorising. The results of this family, in the work both of Lewis and others, seem fruitful and exciting: at least to an admitted insider like me.

## 3. Counting the Costs

One important part of Lewis's philosophical method is that he saw deciding between philosophical questions as a matter of weighing up costs and benefits, and selecting the theory which, in the judgement of the weigher, did best by that measure. Lewis was not alone in this: many contemporary philosophers will pay at least lip service to the idea that what must be done when deciding a philosophical issue is to weigh up costs and benefits of rival philosophical views. Perhaps because this is such a widespread idiom, it seems worthwhile paying closer attention to the role this played in Lewis's philosophy. What is it to weigh up costs and benefits of a theory? What is a cost, what is a benefit, and how is the weighing to be done?

The first place this way of talking is prominent is in "Holes", a paper co-authored with Stephanie Lewis (Lewis and Lewis 1970). The paper is a dialogue between two characters neither of whom is a spokesperson for the authors, but where the characters agree it is probably safe to assume their views correspond with their authors'. There, the characters talk of "paying a price" of plausibility, when a theory disagrees with common opinion, and a theory "earning credence" through clarity and economy. The characters also suggest that many debates over "ontic parsimony" are a matter of counting the costs of disagreement with common opinion on the nominalist side, with the advantages accruing to a nominalist when his or her theory is economical and/or clear. (Lewis and Lewis 1970 pp 211-212)

Lewis endorses this language of costs and benefits in his own voice in Lewis 1983. Even after we have a stock of counterexamples before us and have heard all the philosophical arguments, "presumably we will still face the question of which prices are worth paying, which theories are on balance credible, which are the unacceptably counterintuitive consequences and which are the acceptably counterintuitive ones". (p x) This might sound like an ultimate appeal to intuitions for assessing theories, but it is not quite that: on Lewis's view, "[o]ur 'intuitions are simply opinions'" (p x). It looks like this assessing of prices, and working out the credibility of theories, must be done by the lights of our opinions taken as a whole.

Lewis also uses the language of costs and benefits in Lewis 1986b (p 135). He admitted that his concrete modal realism involved a "denial of common sense", and said "I think it is entirely right and proper to count that as a serious cost". Nevertheless, he thought, "the price is right" since "the theoretical benefits are worth it". As well as that judgment, though, he thought that he needed to show "that they cannot be had for less": concrete modal realism about possible worlds was better than rejecting possible worlds, but should only be believed if it was also sufficiently better than ersatz modal realist alternatives. In this passage Lewis also contrasts losses from disagreements with common sense with gains from theorising, this time the "earned credence that is gained by making a theory more systematic" (Lewis 1986b p 134).

Enough is said in these passages to get a general sense of what Lewis has in mind. The "costs" and "benefits" seem to be in credibility, or gains and losses of credence theoretical options receive. A theory can gain benefits from agreement with our earlier opinions, but there are other ways to gain benefits as well: clarity, economy, being systematic. I assume Lewis would agree that there are other ways to gain benefits as well: empirical support from scientific inquiry, for example.

One puzzling thing about some of this discussion are the things Lewis thinks can be "benefits", if the goal is to select the most belief-worthy theory. Why would clarity or being more systematic add to the credibility of a theory? Both can sound like mere matters of presentation. They might still play an important role for less than ideal theorisers - perhaps clarity can show some apparent problem to be merely apparent, and a systematic theory can make it obvious how evidence bears on hypotheses in a way that a jumble of evidence statements and theoretical assertions does not. But I suspect Lewis had something more substantial in mind. One way to "clarify" a theory is to stop running two phenomena together; and this can be particularly useful if you initially said inconsistent things. If A is F, and B is not-F, and A and B were initially confused, then instead of one thing described as F in one place and not-F in another, a theory that says there are two things, one F and one not-F, will be much more believable. "Clarification" need not be this extreme, of course, but drawing distinctions can be more valuable than merely providing convenience of presentation.

I suspect fuller-bodied theoretical virtues are relevant in other ways here as well. "Economy" of a theory could just be succinctness; but Lewis's remarks suggest he has in mind a virtue like parsimony of theoretical postulates. "Systematicity" might signal a unified theory, with theoretical postulates each confirmed by, and perhaps explaining, a range of evidence. (Lewis speaks of "trying to improve the unity and economy" of a theory in Lewis 1986b p 134, which also suggests he values unification in theorising.) Lewis never offers us a complete list of the features that confer theoretical "benefits" to a theory: nor, for that matter, do we ever get an exhaustive list of costs. I suspect it would not be easy to discover such lists. It could well be that the task of saying what bears on the costs and benefits of theories, and the weights that should be attached to those costs and benefits, is about as hard, or harder, than working out what is the best philosophical position on each philosophical issue. One might even doubt that there is a general answer available here, if one was enough of a methodological particularist, though there is no evidence that Lewis, at least, was a particularist in these matters. It is worth noting that Lewis thought these costs and benefits were relevant to the question of which philosophical theory to believe (or give credence to): so he seems, at least implicitly, to have been committed to the view that features like clarity, systematicity and economy play a role in belief-worthiness, not merely practical aspects of theory acceptance and manipulation.

Another puzzling thing about Lewis's discussion is what he means by "credence" here. Philosophical conjectures are supposed to gain or lose "credence" as we discover clashes with intuition or pre-theoretic belief, or come to appreciate their economy or other systematic virtues. Elsewhere (e.g. Lewis 1983 pp 83, 110, Lewis 1999 p 404), Lewis uses "credence" elsewhere to pick out a rational degree of belief, and sometimes this rationality is quite idealised. A suitably idealised agent gives probability 1 to every proposition true at all possible worlds, and so we might expect that our "credence", our idealised degree of belief, in any necessary truth is also 1. Insofar as a philosopher is considering a hypothesis that is necessarily true or necessarily false, that philosopher's "credence" in that proposition, if idealised enough, is already either 1 or 0, and would not change by noticing things like its agreement with commonsense, its economical capturing of phenomena of interest, and so on.

Whatever Lewis means elsewhere, I think we should not take him to intend that "credence" is such an idealised matter when we discuss philosophical method. Lewis seems to be offering advice that is not merely descriptive – it is not that whatever we happen to do with philosophical arguments is automatically right – but is not idealised to the point where we already endorse every necessarily true philosophical hypothesis. Or at the very least, we do not endorse every necessarily true philosophical hypothesis *in the guise that it is presented*. Lewis does hold that there is only one necessary proposition and we already believe it, in at least one good sense of the term "proposition". But when he says this, he allows that we might be mistaken about exactly what our sentences express (see Lewis 1986b p 36, where he cites Stalnaker 1984 approvingly.) So when we are trying to work out which philosophical doctrine deserves our approval, it may be that sometimes Lewis would characterise what are doing as trying to work out which sentences express the necessary proposition. (Indeed, this might also happen with contingent philosophical matters: we might in some sense know all the relevant facts, but be unsure which philosophical sentences correctly capture the phenomenon we are already familiar with.)

To sum up, once we have a range of options before us, preferably clarified and systematised, we have to weigh them up before coming to a verdict. In doing this weighing, a crucial role must be given to their agreement with starting opinion: it is a cost insofar as they depart from what we believed before (and, according to Lewis, a decisive cost if they vary from any "Moorean facts"). Consistency with scientific findings, particularly well-established scientific results, is also a plus. Theories must also be assessed for unity, economy of postulation, and other such virtues, and these can count for a lot: for example, Lewis concedes that his modal realist metaphysics amounts to a "severe" denial of common sense, but nevertheless thinks that the theoretical benefits it brings make it worth the cost (Lewis 1986b 135). Incidentally, this is a case where we are plausibly choosing between hypotheses that are all necessary if true and impossible if false: it is not as if the question of whether there are a plurality of concrete possible worlds is one which has an answer that varies from possible world to possible world.

My own view, for what it is worth, is that assessment of philosophical theories for theoretical virtues such as simplicity, unificatory power, track record of apparently successful problem solving, and so on does, and should, play an important role in philosophical theory choice. (And not just philosophical theory choice - in choices among more theoretical hypotheses in general.) Lewis's use of evidence from common sense and the findings of science also seems to me right: attempts to find a completely disconnected method for specifically philosophical questions seem misguided, especially as our philosophical opinions need to mesh with our other opinions about the world. Despite Lewis's protestations that common sense has no "absolute" authority in philosophy, insisting that we can never move too far from our starting position seems an overly strong application of theoretical conservative principles: once we have used our starting points to make progress from, it should be an option that as we discover more, we eventually move ever further and further from our beginnings, whether those are the beliefs of individuals before systematic inquiry or the beliefs of our ancestors (Quine's *homo javenensis*). I also suspect that Lewis's scattered appeals to "clarity", "unity and economy" and so on risk downplaying the methodological heavy lifting that principles of simplicity and parsimony, unity, and perhaps other forms of theoretical support must play if we are to justify philosophical systems. And not just philosophical systems - even an explanation of how Newton's physics was justified given the evidence he had available will require heavy reliance on the epistemic role of theoretical considerations, it seems to me, let alone a justification for a contemporary overall scientifically informed view of the natural world.

Lewis also suggests that there is a certain lack of decisiveness in philosophy as a result. "Philosophical theories are never refuted conclusively", Lewis says, though he adds “Or hardly ever”. (Lewis 1983 p x). Lewis thinks there are no “knock down” arguments in philosophy, or hardly any, because in practice there is almost always some option open to someone who holds a view and who wishes to maintain it. Lewis might just be making a psychological point about philosophers here—that they are too bloody-minded to all be convinced by an argument – but I suspect he is suggesting in addition that there is almost always some way of holding on to an antecedently-held view in the face of problems, if someone has an unusual enough calculus of costs and benefits, that is at least somewhat rational.

“Once the menu of well-worked out theories is before us, philosophy is a matter of opinion” (Lewis 1983 xi). It is unclear from Lewis’s remarks whether he thinks this is a matter of what philosophical disagreement is like *in principle*, or whether, for example, eventually ideally rational philosophers in possession of all relevant evidence would converge. One reason to think that they may not converge is the Bayesian one mentioned on p <10>: if they rationally start from different places, and rationally allow the evidence to impact on them differently, there is no guarantee they will end up in the same place.[4] It probably does not matter very much which, methodologically speaking: if it is not feasible to secure rational agreement between careful, intelligent philosophers of goodwill possessed of the same considerations, we are likely to be stuck with these divergences, and with the absence of “knock down” considerations in many cases, for the forseeable future.

## Conclusion

The methods of philosophers are often not very clearly defined compared to the methods of those pursuing archival work, or studying gene expression, or doing many of the other research tasks in contemporary disciplines. While this can make it hard for students to see how to do philosophy well, or to get consensus even among professional philosophers about the quality of different pieces of philosophical work, it is clear that not anything goes, and there are ways of investigating philosophical questions that are better than others. The challenge for philosophical methodologists is to say something useful about how philosophy is done well, and how to do it better,

[4] There are a number of Bayesian “convergence theorems” around, but even they do not prove that *any* set of admissible priors will *definitely* converge with *all* others, even in the limit.

navigating between the one danger of being too trite and the other danger of proposing a theory of philosophical method that is innovative but wrong-headed.

The recommendation to following the example of philosophers who seem to be doing philosophy well seems like one piece of advice that is potentially useful and unlikely to lead completely astray. Even among those who have extreme disagreements about Lewis's conclusions often allow that he produced high-quality philosophy. So Lewis is one place to start for those looking for an exemplar of contemporary philosophy. My view is that we can do even better, methodologically speaking, than Lewis: but even if we could do as well, that would be doing well indeed.[5]

*Daniel Nolan*
*School of Philosophy, College of Arts and Social Sciences*
*Australian National University*
*ACT 200*
*Australia*
*Daniel.Nolan@anu.edu.au*

**References**

Braddon-Mitchell, D. and Nola, R. (eds) 2009. *Conceptual Analysis and Philosophical Naturalism*. MIT Press, Cambridge MA.

Carnap, R. 1963. "Replies and Expositions" in Schlipp, P.A. (ed) *The Philosophy of Rudolf Carnap*. Open Court, LaSalle, Illinois, 859-1013

French, S. and McKenzie, K. forthcoming. "Thinking Outside the (Tool)Box: Towards a More Productive Engagement Between Metaphysics and Philosophy of Physics". *European Journal of Analytic Philosophy*.

Goodman, N. 1954. *Fact, Fiction and Forecast*. Harvard University Press, Cambridge MA

Jeffrey, R. 1983. *The Logic of Decision* 2nd edition. Chicago University Press, Chicago.

Lewis, D. 1983. *Philosophical Papers Volume I*. Oxford University Press, Oxford.

Lewis, D. 1986a. *Philosophical Papers Volume II*. Oxford University Press, Oxford.

[5] Thanks to Chris Daly and Wolfgang Schwarz for discussion of Lewis's method.

Lewis, D. 1986b. *On the Plurality of Worlds*. Blackwell, Oxford.

Lewis, D. 1989. "Finite Counterforce" in Shue, H. (ed) *Nuclear Deterrence and Moral Restraint*. Cambridge University Press, Cambridge, pp 51-114

Lewis, D. 1991. *Parts of Classes*. Blackwell, Oxford.

Lewis, D. 1995. "Ern Malley's Namesake". *Quadrant*, March 1995: 14-15

Lewis, D. 1998. *Papers in Philosophical Logic*. Cambridge University Press, Cambridge.

Lewis, D. 1999. *Papers in Metaphysics and Epistemology*. Cambridge University Press, Cambridge.

Lewis, D. 2004. "How Many Lives Has Schrödinger's Cat?". *Australasian Journal of Philosophy* 82.1: 3-22

Lewis, D. and Lewis, S. 1970. "Holes". *Australasian Journal of Philosophy* 48: 206-12. Reprinted in Lewis 1983.

Moore, G.E. 1939. "Proof of an External World" in Moore, G. E. 1959. *Philosophical Papers*. George Allen & Unwin, London, pp 127-50

Nolan, D. 2005. *David Lewis*. Acumen, Chesham.

Nolan, D. 2009. "Platitudes and Metaphysics" in Braddon-Mitchell and Nola 2009, pp 267-300

Rawls, J. 1971. *A Theory of Justice*. Harvard University Press, Cambridge MA

Smith, M. 1994. *The Moral Problem*. Blackwell, Oxford.

Stalnaker, R. 1984. *Inquiry*. MIT Press, Cambridge MA

van Fraassen, B.C. 1989. *Laws and Symmetries*. Clarendon Press, Oxford.