Explanation, Idealisation and the Goldilocks Problem

Brian Weatherson

Michael Strevens's book *Depth* is a great achievement.[1] To say anything interesting, useful and true about explanation requires taking on fundamental issues in the metaphysics and epistemology of science. So this book not only tells us a lot about scientific explanation, it has a lot to say about causation, lawhood, probability and the relation between the physical and the special sciences. It should be read by anyone interested in any of those questions, which includes presumably the vast majority of readers of this journal.

One of its many virtues is that it lets us see more clearly what questions about explanation, causation, lawhood and so on need answering, and frames those questions in perspicuous ways. I'm going to focus on one of these questions, what I'll call the Goldilocks problem. As it turns out, I'm not going to agree with all the details of Strevens's answer to this problem, though I suspect that something *like* his answer is right. At least, I hope something like his answer is right; if it isn't, I'm not sure where else we can look.

1   The Goldilocks Problem

Sam has engaged in some unhealthy activity, and is now profusely vomiting in the bathroom. Here are three things that are true of the buildup to this unfortunate turn of events.

(1) Sam either ate a carton of raw eggs, or drank a bottle of vodka.
(2) Sam ate a carton of raw eggs.
(3) Sam ate a carton of raw eggs that were bought at midday.

All three of these claims are interesting things to know about the buildup to the vomiting. But intuitively, or at least according to my intuitions, (2) is the best *explanation* of the lot. That's because intuitively, (1) is too weak, and (3) is too strong, while (2) is just right.

Let's assume for now these intuitions are correct. We then have the puzzle of explaining why explanations of moderate strength, like (2), are strictly better than either weaker explanations, like (1), or stronger explanations, like (3).

---

[1]Strevens (2008). All page references, unless otherwise noted, are to this book.

Put another way, we have to explain what makes (2) 'just right'. Call this the Goldilocks problem.[2]

If the Goldilocks problem was merely a matter of first-pass intuitions, then perhaps the right way to solve it would be to explain why we have quirky intuitons about explanations. But I think we can see that it turns on deeper features than that.

On the one hand, we want explanations, particularly of single events, to locate those events in the causal structure of the world. That's why we're pushed towards saying that (3) is the bext explanation of Sam's current activity. Indeed, in his defence of a causal theory of explanation, David Lewis (1986) says that (3) is really the best explanation, though we might prefer to use, or to offer, (2) for pragmatic reasons.

On the other hand, we want explanations that unify disparate phenomena. If we see that an event is just one instance of the right kind of pattern, it feels more explicable. That pushes us towards explanations that encompass more and more actual and possible outcomes. This pushes us away from (3) as an explanation, and towards (2), but also away from (2) and towards (1). After all, if we accepted (1) as the best explanation for what's going on, we would have an explanation that encompasses even more events.[3]

We can also get pushed towards (1) as being the ideal explanation by considering ways in which (2) is a better explanation than (3). There is a sense in which some of the information in (3) is *redundant*. No matter when Sam bought the eggs, the vomiting would have resulted given that they were eaten. Here is one principle we might draw from that. If E' is logically weaker than E, and the outcome O would have happened even if E' had happened but E had not, then E' is a better explanation than E. This will get the right result that (2) is a better explanation than (3). But it will get the wrong result that (1) is a better explanation than (2).

To some extent, the observations of the last three paragraphs point to a solution to the Goldilocks problem. There are virtues that (2) has over (3), in not being too specific, and over (1), in being specific enough for the task at hand. But as a moment's reflection will show, attempting to turn these ideas into a theory is not exactly trivial. It's much too easy to come up with principles that end up implying that (2) has all the *vices* of (1) and (3), and is really *worse* than each, rather than better. (The attempt to use counterfactuals to give a sufficient condition for

---

[2]Strevens calls the problem of how to explain why (2) is a better explanation than (1) 'the disjunction problem'. Given that the problem arises in the context of a theory that aims to explain why (2) is better than (3), I think the disjunction problem and the Goldilocks problem are not particularly distinct.

[3]For more on explanation as unification, see Friedman (1974) and, especially, Kitcher (1989).

superiority of explanation in the last paragraph is illustrative of how we might end up theorising this way.) Having a theory of explanation that avoids these traps is both desirable, and difficult.

2 Idealisations in Explanation

Much more familiar than the Goldilocks problem is the problem of accounting for the role of idealisations in explanation. Explanations seem, after all, factive. The sentence *p because q* just entails both *p* and *q*. And yet explanations involving idealisations seem to be false. Here's an illustrative example.

On a busy suburban corner, there are four gas stations.[4] Although the price for which they offer gas fluctuates a lot from day to day, the four usually have the same price, even to the nearest tenth of a cent. Why might that be? One might suspect collusion, but we'll stipulate that this is a real free market, and the stations are actually competing, not colluding. Another might be that the stations are using 'cost-plus' pricing. But in fact, given the many and varied ways in which the stations (or their corporate parents) have used derivatives to hedge their costs, the four actually face very different input costs. And in any case, a 'cost-plus' theory can't explain the fluctuation of prices.

The real explanation is relatively simple. If any station charges a higher price than its rivals, then no one will come to that station. And that's something the station desparately wants to avoid. So no station charges a higher price than the others. And that means they all charge the same price.

Now why, might we ask, is it that if any station charges a higher price than its rivals, then no one will come to that station? There's a simple explanation here too. First, customers know the prices at each of the four stations, or at least if they don't the cost of getting those prices is zero. Second, the customers are each utility-maximisers who prefer having more money to less. And third, the goods that the stations are offering are perfect substitutes. Those three premises entail that a station with a higher price than the others will have zero customers.

But just wait! Precisely none of those three premises are perfectly true. There is some cost in figuring out the prices at each. If there weren't, we couldn't explain why stations put up such big signs advertising their prices. The point of those signs is to *reduce* the cost of acquiring price information. And, as philosophers of economics never tire of pointing out, customers aren't perfect utility maximisers. And, finally, the goods aren't perfect substitues. The stations might have different

---

[4]'Petrol stations' if that fits your dialect better.

queue lengths, or reputations for quality, or associations with firms that pollute the Gulf of Mexico, and so on.[5]

Strevens has a nice story to tell here about what we should say about the explanations like the one I just offered. When the explainer says that, for instance, the cost of acquiring price information is zero, we should interpret them charitably, and loosely. We should apply the same principles as we apply when interpreting someone's claim that Brazil is triangular. The truth-conditional content of the claim is not that the cost of acquiring price information is *precisely* zero. Rather, it is that the cost is in a not-too-large range that includes zero. How large is 'not-too-large'? That depends on what the person is trying to explain? If they are trying to explain the size of gas station signage, it will be a small range; if they are trying to explain the dynamics of gas station pricing, it will be somewhat larger.

Strevens's theory here is *hermeneutic*, not *revolutionary*. He doesn't say that we should replace the explanations that economists give, which are full of freely available information, perfectly substitutable goods, utility maximising agents and so on, with explanations that involve low cost information, highly substitutable goods, and agents who usually choose high utility outcomes. Rather, he is saying that the explanations those economists give already involve low cost (but not necessarily free) information, highly (but not necessarily perfectly) substitutable goods, and so on. This seems entirely right to me. Well known results showing the limitations of human rationality simply don't undermine the stories like the one I told explaining the correlation between prices at nearby gas stations, even though a cursory glance at those explanations might appear to involve appeal to perfectly rational buyers.

Now what happens when we interpret an explanation as saying not that some value is zero, but that it is near zero? Well, we get an instance of the Goldilocks problem back. We could imagine an explanation of the gas station prices that includes the exact value of the cost of acquiring information about each station's price. That explanation would be more precise than the explanation that merely says the cost of acquiring price information is low. But despite that increase in precision, it would be a *worse* explanation, and it would be worse for just the same reason that (3) is a worse explanation than (2). (Of course, we haven't yet said just what that reason is!)

So puzzles about idealisations in explanation reduce, given Strevens's nice hermeneutic suggestion, to the Goldilocks problem. That raises the interest in solving the Goldilocks problem, so let's turn to Strevens's own solution to it.

---

[5]Given the last point, we'd expect that after the BP disaster in the Gulf of Mexico, stations weren't too worried about being undercut on price by a nearby BP station.

3   The Kairetic Theory of Explanation

I'm going to have to simplify a lot in sketching Strevens's theory of explanation, but I hope the following offers a not-too-inaccurate picture. For Strevens, explanations of individual events are causal models. (Explanations of regularities are basically explanations of the events that make up the regularity.) A causal model is a valid argument, whose premises are all true, and whose conclusion is the event to be explained, such that the conclusion can be derived from the premises using (more-or-less) nothing but modus ponens, with every such step, from C and $C \rightarrow E$ to E, being such that in reality C caused E. When an argument has this property, Strevens says that the premises *causally entail* the conclusion. In practice, these models typically have three (kinds of) premises: a specification of initial conditions, a law (or set of laws) linking those conditions to the eventual result, and a 'no defeaters' condition, since the laws in question will usually not guarantee any outcome.[6]

There will usually be many such explanations. For instance, we could start with either (1), (2) or (3), add an appropriate law and a no defeaters condition, and causally derive that Sam is nauseous. Strevens then puts two extra conditions on causal models, one of which provides a ranking of explanations, the other of which is a necessary condition for an explanation being satisfactory.

The ranking condition is that the weaker the set of initial conditions are, the better the explanation is. If we weaken the initial conditions, but can still causally derive the explanandum, then the stronger set of initial conditions contained redundant information and better explanations excise redundant information. The condition that some information is necessary for the causal entailment to go through is what Strevens calls 'the kairetic condition' on explanatory relevance, and that in turn is why the theory is called a kairetic theory of explanation.

Once we loosen the specification of the initial conditions, a range of different possible causal pathways are compatible with the argument being a causal entailment. The necessary condition Strevens adds is that these possible pathways must be *coherent*. And he defines cohesion as "dynamic contiguity" (105). That is, if we situate all the possible causal chains in a possible space, an argument satisfies the cohesion condition if the set of chains consistent with the argument's premises causally entailing the conclusion form a contiguous set.

Note that contiguity is not that closely related to a similarity condition. The set of all possible causal pathways is perfectly contiguous, although its members

---

[6]For instance, the gravitational law says that there is a downward force on my coffee cup, but it doesn't guarantee that it moves downwards. And, indeed, there are currently sufficiently many forces acting on it that it remains suspended 90 feet above the ground. The 'no defeaters' clause is intended to rule out such mischief.

are severely dissimilar. On the other hand, some small sets of causal pathways are not contiguous. So consider (4) and (5) below. Arguably the set of worlds in which (4) is true is not contiguous – there is a disconnect between the worlds where Suzy throws and the worlds where Billy throws – while the set of worlds in which (5) is true is contiguous.

(4) Either Billy or Suzy threw a brick at the window at exactly 2π mph.
(5) Suzy threw a brick at the window at between 5 and 30 mph.

Although the worlds where Suzy throws hard are very dissimilar from the worlds where Suzy throws softly, there is a chain of worlds connecting the two. And each member of the chain is very similar to the next member. That suffices for contiguity.

It's important to what follows that Strevens takes contiguity here to be *physical* contiguity. That is, two worlds (or causal pathways) are contiguous iff they are contiguous from the perspective of fundamental physics. Contiguity is not meant to be something defined in terms of explanations, and nor is it meant to be contiguity in terms of properties of interest to the special sciences. This will be important for what follows.

We're now in a position to see Strevens's solution to the Goldilocks problem. The detail about when Sam bought the eggs is irrelevant to the conclusion that Sam is nauseous. As long as the eggs were bought, and eaten, Sam's nausea will exist. Indeed, its existence will be guaranteed by a causal law, given the appropriate 'no defeaters' condition. So the kairetic condition says we improve the explanation of Sam's nausea by dropping the time at which the eggs were bought.[7] Now if we start with (1), there will still be a causal law that lets us derive Sam's nausea. But the space of causal pathways consistent with the argument we generate will not be contiguous. It will contain the worlds where the eggs cause nausea, and the worlds where the vodka causes nausea, and nothing in between. So it isn't an eligible explanation. So the kairetic account predicts, correctly, that the best explanation of Sam's nausea starts with (2). QED.

4   Equilibrium Explanations in Economics

But there's a difficulty looming for this nice theory. It isn't at all clear how we're going to generalise it to cover explanations in the social sciences. It's perhaps

---

[7]Of course, if we wanted to explain the *time* of Sam's nausea, and not just its existence, the extra details in (3) might matter.

easiest to see this if we look at an example. This example is originally from Hendricks and Porter (1988), though much of my discussion of it leans heavily on the exposition in Sutton (2000, 47-56).

The fact to be explained concerns the amount that oil exploration firms pay for, and eventually earn from, licences to drill in various tracts of the Gulf of Mexico. At various times, the government opens up the rights to drill on new tracts of sea bed. Firms are allowed to make a single bid for the rights to these tracts, and the highest bid wins. Some firms that bid have, prior to the opening of the new tract, drilling rights to some adjacent tract, and some do not. Having drilling rights to an adjacted tract is useful, because oil deposits tend not to follow the sharp lines on government surveyors' maps. If you have already been working on an area adjacent to the one being auctioned, you have a pretty good idea how much oil that tract contains. If you don't, then you have to make a guess based on more general features of that region of the Gulf. The stylised fact to be explained is that firms that bid on tracts adjacent to their existing tracts made a large profit, on average, while firms that bid on non-adjacent tracts made no net profit. (In fact they averaged a small loss, but the amount is close enough to zero that it's worth treating their net returns as zero.) Why might this be?

The explanation that Hendricks and Porter offer starts with the following game, from Wilson (1967). Assume that two players, A and B, are bidding on a good of some value in $[0, 1]$. A knows exactly how valuable the good is - call this value x. B has no idea how valuable the good is; her credences about its possible value are distributed evenly over $[0, 1]$. Both A and B know these facts about each other. What should each of them do?

Standard game theory has an answer. The game has a single Nash equilibrium. A bids $x/2$, and B plays a mixed strategy, randomly choosing a bid from $[0, 1/2]$. If each of them play these strategies, then A has an expected return of $x^2/2$, and B has an expected return of 0. Moreover, given each of them is playing those strategies, the other party cannot do better by changing their strategy. (That's just what it means for the strategies to form a Nash equilibrium.)

Now Hendricks and Porter go on to suggest that the drilling rights auctions are more or less like these games, with A's role being filled by the firm with an adjacent tract, and B's role by the firm with no adjacent tract.[8] If we apply

---

[8]There are a lot of technical details I'm suppressing here, many of which Hendricks and Porter take into account, and many of which they rightly suppress. Sutton characterises Hendricks and Porter's model as having considerably fewer added complications to the simple game Wilson develops than Hendricks and Porter themselves do. For instance, Sutton suppreses, while Hendricks and Porter explicitly consider, the possible efficiency gains derivable from owning adjacent tracts, but this only makes a small difference to the final result. On the kairetic theory of explanation, these simplifications actually improve the explanation considerably.

that model, we get plausible results for how much profit the two kinds of firms should earn, including a nice story about why the non-adjacent firms earn no profit. Indeed, we even get a satisfactory (at least to economists) story about why firms without adjacent tracts continue to bid even though they earn no net profit by doing so. If they didn't bid, then firms with adjacent tracts could win the bidding by bidding a penny, and then it would be valuable to bid. In other words, the only equilibrium solution requires them to bid, even though they get no gain from it.

There is obviously a bit of work to do to show that this game provides a good model of Gulf of Mexico auctions. For one thing, we have to show that we can treat the auction as having effectively two players. Hendricks and Porter suggest that the behaviour of firms with adjacent tracts is sufficiently cooperative that this is a legitimate idealisation. There are other idealisations too, all of which I think can be fit nicely into Strevens's kairetic story. We have to treat the firms with adjacent tracts as knowing the value of the tract, when really they'll only know the approximate value. But it is plausible that treating their ignorance as being zero-valued, i.e., treating their knowledge as being perfect, makes no difference to what we need to explain. Similarly, it is not really true that the other firms have no idea how valuable the tracts are. But their knowledge levels are close enough to being represented by a flat probability distribution over the possible values of the tract that it doesn't make a difference to this story to model their knowledge more precisely.

(There is an interesting technical point here. Strevens focusses on cases where the idealisations involve giving some variable a "zero, infinite or some other extreme or default value" (318). In social sciences, one useful 'default' value is that the variable is represented by a flat probability function over some interval. This will rarely be exactly right; whether we interpret the probability function metaphysically or epistemically the 'right' function will presumably have some bumps or kinks in it. But it is an acceptable idealisation.)

So far so good. We started with an interesting set of facts, we found a nice mathematical model that has the facts as a consequence, and we argued (or at least hinted at how one could argue) that the deviation between the model and the facts was irrelevant to the outcome to be explained. So we've arguably fit a widely accepted economic explanation into the kairetic framework.

But once we start looking at the details, some problems start to emerge. Remember that on the kairetic account, explanations must be causal derivations. It doesn't look at first like we've got any causation in the economic explanation. But I think that's wrong. After all, there's a reason why the two types of firms bid they way they do. The structure of the auction, along with other facts, causes

them to make these bids. It isn't something you'll see highlighted in Hendricks and Porter, but it's arguable their story is a causal story.

The problem is the 'other facts' you need to cite to complete this causal explanation. Those don't seem to be sufficiently 'cohesive' for Strevens's story to hold up. What we know is that if the actors follow equilibrium strategies, then we'll get the results that are actually observed. But why should we think that actors will do just that? There are several possible reasons; too many reasons it might seem for the kairetic theory to work.

Possibly the actors are perfectly rational, and perfectly rational beings play Nash equilibrium strategies.[9] Possibly the actors are worried about their strategies leaking out, and are maximising expected utility relative to that assumption.[10] Possibly there are a number of actors playing other strategies, but they don't tend to survive economically, and so the statistics are dominated by firms that do survive, and the survivors generally play equilibrium strategies.[11] Possibly the firms are run by a lot of game theorists, and "game theory is an excellent way of predicting the behaviour of professional game theorists."[12] More likely, some combination of these four reasons, and even some others, is causally relevant to the establishment and maintenance of this equilibrium.

And that is something that's hard to fit into the kairetic framework. We can show how the background facts about the case (i.e., the risks and rewards facing the competing oil firms), and a general causal law (i.e., that firms tend to end up playing equilibrium strategies) entail the conclusions that various firms bid on newly released tracts despite having zero expected profit. The problem is that many distinct causal pathways are compatible with this loosely described causal structure, and these pathways are not 'cohesive'. So the kairetic theory of explanation predicts that the explanation offered in Hendricks and Porter (1988) is not a good explanation of the observed behaviour in the auctions. That should

---

[9]The normative claim here, that perfectly rational beings play Nash equilibrium strategies, seems implausible to me for reasons similar to those set out by Stalnaker (1996; 1998; 1999).

[10]Note that maximising expected utility does not entail playing equilibrium strategies without some extra assumption about strategies leaking, since a mixed strategy can be part of a unique equilibrium, but can never be uniquely utility maximising.

[11]Philosophers tend to overstate how much economists rely on rationality assumptions. One of the attractions of game-theoretic explanations is that they don't require all the agents to be perfectly rational. After all, game-theoretic explanations work well in evolutionary biology, and the players there are certainly not perfectly rational. For more on this point, and especially on how much work economists do to weaken rationality postulates, see Hahn (1996).

[12]The quote is from a blog post by Daniel Davies on October 8, 2010. See http://d-squareddigest.blogspot.com/2010/10/on-not-being-obliged-to-vote-for.html.

worry anyone who either finds it intuitively plausible that it is a good explanation, or thinks that we should defer somewhat to the salient experts on what is a good explanation.

5   Possible Responses

I think these equilibrium explanations are a challenge to Strevens's solution to the Goldilocks problem, and I think that's a problem given the importance of solving the Goldilocks problem to the broader aims of the kairetic theory of explanation. But there are a number of ways Strevens could respond to this challenge. Indeed, we can see three responses already made in *Depth*. So I'll end by noting why I don't think those three responses work.

First, it is true that some equilibrium explanations are cohesive in Strevens's sense. Strevens discusses an example proposed by Elliot Sober (1983). Here is how Strevens describes the case.

> Consider a ball released at the inside lip of a basin. The ball rolls down into, then back and forth inside, the basin, eventually coming to rest at its lowest point. This will happen no matter what the ball's release point. … Sober claims, quite rightly, that [an] equilibrium explanation … is the best explanation of the ball's final resting place. (267-8)

Now there are many ways in which the ball might have reached its equilibrium state. But note that these ways are all fairly similar to one another. The ways are, collectively, *cohesive* in just the sense needed for the kairetic theory.[13] But this is surely an accident of the example. The ways in which agents reach a game-theoretic equilibrium are very different from one another, which makes that case rather unlike the case of a ball descending to the bottom of a basin. In short, while some equilibrium explanations will be suitably cohesive many, perhaps even most, will not.

Second, sometimes we don't want to fully explain why p is true, but merely why p is true rather than q, or why p is true given that r is true. In these cases, Strevens says that we exploit 'explanatory frameworks' (149) which fix certain facts as given for the purposes of explanation. So we might take $p \vee q$, or r, to simply be fixed background facts; part of the framework relative to which explanations are made. When a proposition is part of the framework, its presence

---

[13]Actually, this sentence isn't as obviously true as it seems. Strevens's discussion of the case brings out some unexpected difficulties in accomodating Sober's claim in the kairetic theory. But this doesn't affect my point, which is that the case is relatively *easy* for the kairetic theory to accommodate.

in the derivation of the intended outcome does not contribute to incohesiveness (163). So if we say that, for instance, the fact that games like the tract auction end up at equilibrium is part of the framework, then the orthodox explanation of, say, why firms bid despite a zero expected profit, can work. In short, the story about why firms bid is incohesive, but the story about why firms bid given that firms play equilibrium strategies is cohensive, and it is the latter that economists are trying to explain.[14]

Now perhaps that's true of what some economists are doing some of the time. But it seems too defeatist to me. Part of the appeal of game theoretic explanations is that they are supposed to explain why we get to, and stay at, equilibrium. I don't think a practicing economist would say that they are merely presupposing that players in a game reach equilibrium as opposed to offering a theory where that fact falls out as a nice explanandum. It's true that economists do leave some things in the framework. They generally assume that economic actors are agents, while leaving the story about how agency might be physically realised to other disciplines. But it seems wrong to me to say that all the facts about how equilibrium is established and preserved are simply framework questions.

Finally, Strevens notes that we can often refer to causal processes in explanations without being able to fully describe them. If someone asks why the temperature in this room stays so even while the temperature in other rooms fluctuates, I can explain the stability by saying that a thermostat regulates the temperature. Now at first this might look like a very incohesive explanation. There are many ways that a thermostat might work, and they don't form anything like a coherent set. But perhaps that's the wrong way to take my explanation. We could take the explanation as *referring* to the particular thermostat that is present, and the particular way in which it regulates the temperature. That explanation will be *very* cohesive; indeed, the real worry is that it is too precise. Of course, I might not be able to *describe* the process by which the thermostat regulates temperature. But this is no barrier to my being able to refer to it, any more than ignorance of chemistry is a barrier to my being able to refer to H_2O.

Could this help with the tract auction we are discussing? At first glance it seems like it might. Perhaps the explanation can simply refer to the means by which a particular firm ends up playing an equilibrium strategy, even if it cannot describe that means. But the second glance is more troubling. Remember that what we're trying to explain here is an average, not a particular firm's behaviour.

_____

[14]This might be reading too much into Strevens's discussion. What he says about a related example is that the existence of communicative channels within firms is part of the 'framework' when making economic explanations. I don't know whether he would extend this story to cover all means by which firms get to equilibrium˙

And it is meant to be consistent with the explanations that different firms get to equilibrium in very different ways. So we can't really just refer to those different methods; we can only describe what they have in common. And that leaves us back with an incohesive explanation. Indeed, Strevens notes this point in a similar context when he says that "in aggregative and regularity explanation ... there is a real risk" that we won't pick out a cohesive causal mechanism. (154)

So I'm left thinking that we need somehow to supplement the story Strevens offers to make it plausible as an account of explanation in the special sciences. The kind of equilibrium explanations game theorists offer of economic outcomes are at least sometimes good explanations. But what makes them good is not the cohesiveness of their underlying physical mechanisms. It is, at least intuitively, the cohesiveness of the explanations from the perspective of the special science in question. If that intuition is right, we theorists still have work to do in characterising this notion of cohesiveness.

References

Friedman, Michael. 1974. "Explanation and Scientific Understanding." *Journal of Philosophy* 71:5–19.

Hahn, Frank. 1996. "Rerum Cognoscere Causas." *Economics and Philosophy* 12:183–195, doi:10.1017/S0266267100004156.

Hendricks, Kenneth and Porter, Robert H. 1988. "An Empirical Study of an Auction with Asymmetric Information." *The American Economic Review* 78:865–883.

Kitcher, Philip. 1989. "Explanatory Unification and the Causal Structure of the World." In Philip Kitcher and Wesley Salmon (eds.), *Scientific Explanation*, volume 13 of *Minnesota Studies in Philosophy of Science*, 410–505. Minneapolis: University of Minnesota Press.

Lewis, David. 1986. "Causal Explanation." In *Philosophical Papers*, volume II, 214–240. Oxford: Oxford University Press.

Sober, Elliot. 1983. "Equilibrium Explanation." *Philosophical Studies* 43:201–210, doi:10.1007/BF00372383.

Stalnaker, Robert. 1996. "Knowledge, Belief and Counterfactual Reasoning in Games." *Economics and Philosophy* 12:133–163, doi:10.1017/S0266267100004132.

—. 1998. "Belief revision in games: forward and backward induction." *Mathematical Social Sciences* 36:31 – 56. ISSN 0165-4896, doi:10.1016/S0165-4896(98)00007-9.

—. 1999. "Extensive and strategic forms: Games and models for games." *Research in Economics* 53:293 – 319. ISSN 1090-9443, doi:10.1006/reec.1999.0200.

Strevens, Michael. 2008. *Depth: An Account of Scientific Explanations.* Cambridge, MA: Harvard University Press.

Sutton, John. 2000. *Marshall's Tendencies: What Can Economists Know?* Cambridge, MA: MIT Press.

Wilson, Robert B. 1967. "Competitive Bidding with Asymmetric Information." *Management Science* 13:816–820.