Armchair Science

Abstract

We define the notion of *armchair science* – roughly, a concentration on the development of idealized theory with only a loose relation to possible empirical application, and in particular with no specific real-world target in mind. Work in this style is both very influential and very widespread in contemporary social and biological science. We propose that it be subjected to what we call *efficiency analysis*. To this end, we examine in detail the role of the Prisoner's Dilemma game in explaining the live-and-let-live system in World War One trenches. It yields, together with other considerations, a strong prima facie case that armchair science fails the efficiency analysis, in other words that it absorbs too many scientific resources that would be better spent elsewhere. Philosophy of science should be at the forefront of highlighting this important issue. Yet for several reasons, which we explain, existing philosophical work on modeling has fallen short in this critical role.

> "[In biology] factors in a real-life phenomenon are often misunderstood or never noticed in the first place. The annals of theoretical biology are clogged with mathematical models that either can be safely ignored or, when tested, fail. Possibly no more than 10% have any lasting value. Only those linked solidly to knowledge of real living systems have much chance of being used."
> E.O. Wilson

> "In a biology department, you have people doing all sorts of different things. So some do DNA analysis, others do anatomy, some people go and sit with gorillas in the forests of Burundi, and others do experiments with rats. But they are called biologists because biologists recognise that living organisms are complex things and you cannot understand them only at one level. So why can't economists become like that? Yes, you do need people crunching numbers, but you also need people going to factories and doing surveys, you need people watching political changes to see what's going on."
> Ha-Joon Chang

1. Introduction

There seems to be something especially wrong with the idea of armchair science, more so than with that of armchair philosophy. Armchair science – which, we argue, is what a large part of modeling in the social and biological sciences amounts to – endeavors to find truths about the world without direct observational or experimental input. Of course, spending time in the armchair might be useful sometimes. But how useful? And compared to what?

After introducing the notion of armchair science and proposing a new way of assessing its value, we go through an example to make the concerns vivid. In light of this, we then review the flourishing philosophical literature on modeling, with a view to two things: first, to establish that this literature has been insufficiently critical of armchair science, i.e. of large areas of scientific practice; and second, to diagnose why it has fallen short in this respect. The paper therefore has two targets: first, armchair science itself; and second, the existing philosophical coverage of it.

2. What is armchair science?

‘Armchair science’ is our term for modeling whose goal is neither prediction nor more or less faithful representation. Instead it exists, as it were, at one remove from direct empirical contact, exploring relations of dependence between variables that may bear only a loose relation to real-world entities. Such modeling is very widespread in economics, parts of biology, and elsewhere too.

The contrast is to the many models in science that are targeted at particular real-world systems, and that are judged by successful prediction or explanation or other empirical criteria, rather than exclusively by their theoretical innovation. There is little armchairish about this second kind of model, since correct representation or prediction requires actively drawing a strong connection with the target and importing much detailed empirical information. Examples include many models in engineering, econometrics, climate and weather science, election prediction, and many other fields, which typically have as a goal, at least ostensibly, to represent a particular target system and to explain or predict its behavior.

Much work in applied economic theory is armchair science. It is only called ‘applied’ because it attempts to apply the theoretical tools of game theory, equilibrium analysis, order statistics and other mathematical techniques to social phenomena such as bargaining, auctions, or norms. This application yields bargaining theory, auction theory, theory of norms, etc, where the theories in question are collections of more or less related models. Each model represents an imaginary but possible situation, say, two governments bargaining over some good, or two groups of soldiers deciding whether to co-operate. These models are judged as analytical tools for studying the ‘core features’ of real-world negotiations, auctions, and norm adherence. What they are *not* judged on is their ability accurately to represent these real-world phenomena, let alone on their ability to predict them.

A central characteristic of armchair models is that they are deductive, in other words their conclusions follow deductively from their premises. But the phenomena they study typically exhibit diversity and context-dependence – actual auctions, negotiations and norms turn out very differently depending on who participates, where, and under what conditions. As a number of philosophers of science have noticed, to accommodate these two features together, i.e. both deductivity and a lack of empirical regularities, modelers need to import lots of assumptions to give structure to the modeled situation.[1] These assumptions describe features such as unrealistic decision-making procedures, mathematical properties of utility functions, or statistical properties of information distributions. We typically end up with a model that sounds like it is about a familiar object or process (people negotiating), but in fact is about a much more abstract entity (rational agents embedded in a game).[2] To the extent that such modeling represents at all, it represents generalized abstract systems such as bargaining, collective action problems, predation, or signaling. Sometimes it does not do even this, as when the entities in a model are non-existent or even impossible. Examples of these latter include perfectly rational agents with perfect information, infinite populations, or perfectly competitive markets.

Michael Weisberg (2013) calls armchair science *targetless* modeling. We could just use Weisberg's term, but we wish to be a bit more provocative. Just as philosophers have recently started worrying about armchair philosophy, we think scientists should worry about armchair science.[3] We wish to question how much value we are actually getting from targetless modeling given the many resources that go into it.

By targetless modeling we will mean the development of idealized deductive models either without concern for their application or with just a vague hope that eventually at some point they will become target-directed. It is true that there may not be a sharp boundary between

[1] Cartwright describes the process as the transformation from a Galilean experiment to an over-constrained experiment (2010). Weisberg uses the expression 'construal setting' to describe the necessity of interpreting the targets more abstractly in order to accommodate the presence of idealizing assumptions (2013, 121). Musgrave (1981), Maki (2000) and Hindriks (2006) have all proposed a typology of assumptions that go into this type of modeling.

[2] This reliance on the formalization of familiar situations and common sense concepts distinguishes these models from idealized models in the physical sciences. Although the concept of armchair modeling and our corresponding claims may apply to the latter too, in this paper our target is social and biological science.

[3] Of course, there is a great tradition in natural philosophy and science of criticizing apriorism, idealization, thought experiments and related practices. But we do not think these forebears take us very far here. The proliferation and growth of targetless modeling in 20th and 21st century economics and biological science is unparalleled and should be discussed separately. (For the changing meaning of armchairism in economics, see Maas 2010.)

targetless and target-directed modeling, and hence between armchair and non-armchair science. For example, bargaining models in game theory can be used to devise a specific targeted model in which a democratic government issues a threat, the internal opposition either opposes or supports it, and the target state can then resist or comply (Schultz 2001). This model explicitly represents some real-world geopolitical negotiations and is therefore no longer targetless. Similarly, when soldiers in World War One trenches are claimed to be playing Prisoner's Dilemma – an example we will examine below – this model is very much target-directed. Nevertheless, although there may not be a single point when a modeling research program ceases to be targetless and becomes target-directed, we believe the distinction is still important. Targetless modeling is a sustained attempt at articulating, unifying and refining a family of models with only non-committal gestures at future potential applications; target-directed modeling, by contrast, is the use of these resources to explain existing phenomena. Models clearly on the armchair side of the boundary are very common across the social and biological sciences, and non-laboratory sciences in general. (Indeed, the impossibility of experimentation is armchair science's most popular justification.) Our examples will be drawn mostly from economics, but the points raised should hold more generally.

3. The efficiency question

Do we really need another discussion of targetless modeling? The flourishing models literature has offered sophisticated accounts of what models are, the roles they play in scientific practice, the many diverse ways in which they might be idealized, their relations to theories, simulations and experiments, and so on. (See Frigg and Hartmann 2012 for a survey.) Our discussion here rides on the many lessons of this literature. However, we will argue in Section 5, in one way it has been too permissive. In particular, it has failed to subject armchair science to appropriate scrutiny. It has been keen to figure out what models do, but not whether they are doing it too much or too little. Call this latter question the *efficiency question*: are disciplines practicing armchair science to the extent that is appropriate?

We certainly do not want to argue against armchair science always and everywhere. Theory obviously has its place. But we do want to raise the question of whether armchair science justifies the resources thrown into it. It is this focus on efficiency that distinguishes our approach from the Popperian spirit of earlier critiques of theoretical economics (Hutchison 1941, Blaug 2002). Our complaint is not that these models are unfalsifiable, for instance, nor that they do not constitute a progressive research program in Lakatos's sense. One important achievement of the recent modeling literature has already been to show that these criteria do not fit large swathes of contemporary science well (Hausman 1992 among others).[4] How instead, then, do we cash out efficiency? We will conceive of it in a fairly orthodox way: the balance between the resources going into armchair science and the goods coming out of it. Here are three clarifications:

The first is relatively uncontroversial. The resources in question are mainly human. Armchair modeling is, of course, a lot cheaper than building particle accelerators or giant fusion chambers. So what we mean by the costliness of armchair science is the great opportunity cost of having the majority of the best minds in, say, academic economics dedicating their energies to refining armchair models.

The second clarification is more complicated. Science, as a human enterprise, has many stated goals (to educate, to explain, to heal) and sometimes many implicit goals too (to dominate, to exclude, to justify the status quo).[5] A full efficiency analysis should take into account all of these, but in this paper there is room only for a start. So we adopt the narrower criterion: how well does armchair science fare at *establishing causal explanations*? Of course, there is more to science than causal explanations, but as a core ingredient of understanding, policy, and technology, via their intimate connection to interventions, it is hard to imagine a philosophy of science that does not value them. It might be objected that a focus on causal explanations rather than systematic knowledge is more appropriate to history than to social science. In reply, we certainly agree that models of general applicability are highly desirable – but only if they explain or are of heuristic value (see later). Models that do neither are not, especially if they use up huge resources along the way. And when looking at the details, as we will see, it is highly dubious that armchair science makes good on these desiderata. The best way to get a reliable sense of what theoretical input really is useful is by

[4] Our critique here is also distinct from critiques from critical realists (e.g. Lawson 1997), from complexity (e.g. Velupilai 2005), or from moral critiques (e.g. Dupre 2001).

[5] Another, more sociological, way to conceive of efficiency is in terms of influence. The minimalist nature of armchair science, for instance, combined with its relatively low material costs, make it an efficient instrument of rhetoric in the public sphere and a powerful political tool (Erickson et al 2013). But we are interested in the epistemic payoff from armchair science, and want to distinguish this from its rhetorical or political victories.

trying to establish explanations in actual detailed cases. Only then can we know how valuable armchair models truly are or aren’t. So a focus on causal explanation is an excellent epistemic test here (Author-7).[6]

The final clarification: efficiency analysis is not supposed to encompass every epistemic vice that might be connected to modeling. Statistical abuses such as data mining, for example, are problems of a different nature. Rather, armchair science is characterized by a particular bet it makes: idealization is worth the cost. This is the bet that needs to be examined for its rationality.

4. An example: Prisoner’s Dilemma and World War One truces

We cannot give a full defense of the thesis that armchair science has not delivered; no single article could. However, one of our claims is that the limitations of armchair science only become clear in the details. For this reason, it is better to examine one case in depth than many cases superficially. So the particular case study in this section is meant both as an illustration of the efficiency problem and as an example of how to pursue further efficiency analysis. We choose Prisoner’s Dilemma (‘PD’) because it is a vivid example of armchair science with explanatory ambition, and in post-war war social science has become an essential ‘set piece’.[7]

The literature on PD is enormous. According to JSTOR, more than 15,000 articles about it appeared between 1963 and 2013, with no sign of slowing down: 4,400 were just in the last 10 years. Various encyclopedia entries and overviews across economics and philosophy discuss some of this literature’s main developments: asymmetric versions of PD, versions with multiple moves or players, single-person interpretations, versions with asynchronous moves, finitely and infinitely and indefinitely iterated versions, iterated versions with error, evolutionary versions, versions interpreted spatially, and many other variants and nuances besides (Michihiro 2008, Govindan and Wilson 2008, Kuhn 2009). In characteristic armchair fashion, many of these have apparently been motivated by a loose kind of responsiveness to

[6] Our approach is certainly still compatible with according intrinsic value to pure theory development. But armchair science comes with explanatory, not just purely theoretical, ambitions and thus is fair game for an explanation-based conception of efficiency.

[7] This expression is due to Dan Rodgers (2011). We will assume that readers are familiar with the basic PD game.

real-world problems and complications. After all, putative actual players of PDs will often act asynchronously or more than once, or make errors, and so on. Certainly, the subtlety and sophistication of this literature is often impressive. Nevertheless, a striking fact about it is its overwhelmingly theoretical focus. Deeper empirical applications of PD, featuring detailed examination of the evidence of particular real-world cases, are remarkably thin on the ground.[11]

So PD research is very much targetless. Yet its intent isn't. The authors of the New Palgrave Dictionary of Economics entry on co-operation, for instance, are explicit about their explanatory ambitions: "A major goal of economic theory has been to explain how wide-scale co-operation among self-regarding individuals occurs in a decentralized setting." (Bowles and Gintis 2008) But it is hard to find serious attempts at applying PD to explain actual historical or contemporary phenomena.[13] Most applications instead appear in informal sources such as blog posts, teaching tools or off-hand remarks in the media of the sort: "well, that's obviously a Prisoner's Dilemma!"[14] But identifying a casual similarity between PD and an actual situation is clearly not enough, and the closer one looks the more elusive explanatory success becomes. In the limited space here, we will support this claim via an extended analysis of one example. Of course, a single case does not prove much by itself. But, again, it is much more instructive here to look at one case in depth than at many cases superficially.

The particular case we will examine is the 'live-and-let-live' system that arose in World War One ('WW1') trenches, which Robert Axelrod analyzed in terms of PD in chapter 4 of his (1984).[17] We have chosen it in part because it is perhaps the most famous example of PD's supposed explanatory success. It is also the most detailed application we know of PD to a particular real-world target (and we have looked). Moreover, it is an application widely

[11] A lot of the PD literature is 'empirical' rather than theoretical in the sense that it reports on psychological experiments, e.g. whether the co-operate strategy is played by subjects in a laboratory and the factors that influence this. However, whatever their other merits, these kinds of experiment do not address the shortfall mentioned in the text, namely of detailed PD explanations of field phenomena (Author-7).

[13] Sunstein (2007) comes close, but even here the phenomenon in question (the failure of the Kyoto protocol) is explained in part by the fact that it does *not* have a PD structure. The well-known attempts to use PD to explain the arms race, meanwhile, although compelling to Richard Nixon among others, remained largely informal (Erickson et al 2013).

[14] http://cheaptalk.org/2013/11/13/prisoners-dilemma-everywhere-amazon-source/

[17] We analyse this same example at slightly greater length in (Author-7).

considered a notable success and widely recommended for that reason. We take it that it is therefore not an unduly soft target.

Axelrod draws on the fascinating and detailed account of WW1 trench warfare by the historian John Ashworth (1980), itself based on extensive letters, archives, and interviews with veterans. The 'live-and-let-live' system refers to the many informal truces that arose spontaneously on the Western front despite constant severe pressure against them from senior commanders. How could they have happened? Axelrod's case is that, upon analysis, the implicit pay-offs for each side on the front formed an indefinitely iterated PD, and that co-operation – i.e. a truce – is therefore exactly PD's prediction.[19]

Many historical details seem to support his case, such as the limited retaliations that followed breaches of a truce, or the demonstrations of force capability via harmless means in order to establish a threat credibly but non-disruptively. Perhaps the most striking evidence is how the live-and-let-live system eventually broke down. The (unwitting) cause of this was the beginning of a policy, dictated by senior command, of frequent *raids*. These were carefully prepared attacks on enemy trenches. If successful, prisoners would be taken; if not, casualties would be proof of the attempt. As Axelrod observes:

> "There was no effective way to pretend that a raid had been undertaken when it had not. And there was no effective way to co-operate with the enemy in a raid because neither live soldiers nor dead bodies could be exchanged. The live-and-let-system could not cope with the disruption… since raids could be ordered and monitored from headquarters, the magnitude of the retaliatory raid could also be controlled, preventing a dampening of the process. The battalions were forced to mount real attacks on the enemy, the retaliation was undampened, and the process echoed out of control." (1984, 82)

The conditions that PD predicts as necessary for co-operation were unavoidably disrupted and, Axelrod argues, it is no coincidence that exactly then the truces disappeared.

[19] Indefinitely iterated PD has many other Nash equilibria besides mutual co-operation. The analysis that Axelrod actually applies comes from his well-known PD computer tournaments, which led him to predict a strategy of Tit-for-Tat with initial co-operation, which in turn does indeed predict indefinite mutual co-operation. Throughout this section, we will use 'PD' as shorthand for this richer theoretical analysis of Axelrod's. (The main lesson, namely PD's lack of predictive and explanatory success, would apply still more strongly to PD alone, because then we would be faced with the additional problem of equilibrium selection too.)

We agree that many of the historical details are indeed, in Axelrod's phrase, 'consistent with' the situation being an iterated PD.[23] Nevertheless, upon closer inspection PD does not explain why the WW1 truces occurred, nor why they eventually broke down, contrary both to Axelrod's account and to how that account has been widely reviewed.

Why this negative verdict? To begin, by Axelrod's own admission some elements of the story deviate from his PD predictions. First, the norms of most truces were not Tit-for-Tat but more like Three-Tits-for-Tat. That is, retaliation for the breach of a truce was typically three times stronger than the original breach. Second, in practice two vital elements to sustaining the truces were the development of what Axelrod terms ethics and rituals: local truce norms became ritualized, and their observance quickly acquired a moral tinge in the eyes of soldiers. Both of these developments made truces much more robust and are crucial to explaining truces' persistence, as Axelrod concedes. Yet, as Axelrod also concedes, PD says nothing about either. Indeed, he comments (1984, 85) that this emergence of ethics would most easily be modeled as a change in the players' payoffs, i.e. potentially as a different game altogether, thus telling against the value here of PD.

There are several other important shortfalls in addition to those remarked by Axelrod. First, PD predicts that there should be no truce-breaches, but in fact breaches were common. Second, as a result (and as Axelrod acknowledges), a series of dampening mechanisms therefore had to be developed in order to defuse post-breach cycles of retaliation. Again, the Tit-for-Tat analysis is silent about this vital element for sustaining the truces. Third, it is not just that truces had to be robust against continuous minor breaches; the bigger story is that often no truces arose at all. Indeed, Ashworth examined regimental and other archives in some detail to arrive at the estimate that, overall, truces existed about one-quarter of the time (1980, 171-175). That is, on average, three-quarters of the front was *not* in a condition of live-and-let-live. PD is silent as to why.[24]

[23] As we will see, many other details were *not* so consistent. But even if they all had been, this criterion is far too weak for explanation. After all, presumably the WW1 details are all consistent with the law of gravity too, but that does not render gravity explanatory of them.

[24] Ashworth, by contrast, does develop a detailed explanation, largely in terms of the distinction between elite and non-elite units, and their evolving roles in the war. The escalation in the use of raids, so emphasized by Axelrod, is only one part of this wider story. Most areas of the front were not in a state of truce even before the escalation of raiding.

Moreover, PD does not fully address two other, related issues. The first is how truces originated as opposed to how they persisted, about which it is again silent.[25] The second is how truces ended. This PD does partly address, via Axelrod's discussion of raids. But many truces broke down for other reasons too. Ashworth devotes most of his chapter 7 to a discussion of the intra-army dynamics, especially between frontline and other troops, which were often the underlying cause of these breakdowns.

We have not yet even mentioned some more traditional explanatory worries. An obvious one here is that the explanations are after-the-fact; there are no novel predictions. Thus it is difficult to rule out wishful rationalization, or that other game structures might fit the evidence just as well. A second worry is that Axelrod's arguments for the crucial claim that the pay-off structure fits that of an iterated PD (1984, 75) are rather brief and informal. Do they really convince?[27] And are the other assumptions of PD, such as perfectly rational players and perfect information, satisfied sufficiently well?

In light of these multiple shortcomings, how can it be claimed that PD explains the WW1 truces? It is not empirically adequate, and it misses crucial elements even in those areas where at face value it is empirically adequate. Moreover, it is silent on obvious related explananda: not just why truces persisted but also why they occurred on some occasions not others, how they originated, and (to some degree) when and why they broke down.

There is no mystery as to what the actual causal explanations of these various phenomena are, for they are given clearly by Ashworth and indeed in many cases are explicit in the letters of the original soldiers. Thus, for instance, elite and non-elite units had different attitudes and incentives, for various well understood reasons. These in turn led to truces occurring overwhelmingly only between non-elite units, again for well understood reasons. The basic logic of reciprocity that PD focuses on, meanwhile, is ubiquitously taken by both Ashworth and the original soldiers to be obvious. Next, why did breaches of truces occur frequently, even before raiding became widespread? Ashworth explains via detailed reference to different incentives for different units (artillery versus frontline infantry, for instance), and to the fallibility of the mechanisms in place for controlling individual hotheads (1980, 153-

[25] Again, Ashworth covers this in detail (as Axelrod does report).

[27] Gelman (2008), for instance, argues that they do not.

171). And so on. Removing our PD lens, we see that we have perfectly adequate explanations already.

Overall, we therefore judge both that PD does not explain the WW1 truces, and that we already have an alternative – namely, historical analysis – that does. Our point here is simply to register this failure of armchair science. We are not making any further claim about any deep reason for this failure (say, because armchair models are tautologies and that tautologies cannot explain).[28]

So if not explanation, what fallback options are available for defenders of PD? It seems to us there are two. The first is that, explanatory failure notwithstanding, PD nevertheless does provide a deeper 'insight' or 'understanding', at least into the specific issue of why the logic of reciprocity sustains truces. We address this response below (section 9). In brief, we will argue that such insight is of no independent value without explanation, except perhaps for heuristic purposes. This leads to the second fallback position – that even if PD does not provide explanations here, still it is of heuristic value. Is it? Alas, the details suggest not, for two reasons.

First, PD did not lead us to any causal explanations that we didn't have already. Ubiquitous quotations in Ashworth show that soldiers were very well aware of the basic strategic logic of reciprocity. They were also well aware of the importance of a credible threat for deterring breaches (Ashworth 1980, 150). And well aware too of why frequent raiding rendered truces impossible to sustain, an outcome indeed that many ruefully anticipated even before the policy was implemented (Ashworth 1980, 191-198). In other words, PD is following here, not leading.

The second reason that PD lacks heuristic value is that it actively diverts attention *away* from aspects that are important. We have in mind the crucial features already mentioned: how truces originated, the causes and management of the many small breaches of them, the importance of ethics and ritualization to their maintenance independent of strategic considerations, why truces occurred in some sections of the front but not in a majority of them, and so on. Understanding exactly these features is crucial if our aim is to encourage co-

[28] The exact manner in which armchair models might explain is a tricky issue, addressed by many philosophers, and by us elsewhere (Author-1, Author-7).

operation in other contexts too – and this wider aim is the headline one of Axelrod's book, and implicitly surely a major motivation for the PD literature as a whole. Yet here, to repeat, armchair analysis is directing our attention away from them!

A final note: a 'basic' iterated PD is all that was required. Even if we accepted Axelrod's claims, that is, the great bulk of the theoretical PD literature would be irrelevant to them.

Overall, in the WW1 case:

> 1) PD is not explanatory.
>
> 2) PD is not even valuable heuristically. Rather, detailed historical research offered much greater heuristic value in this case, as well as much greater explanatory value.
>
> 3) To the extent that PD could have offered something, the basic version would have been quite sufficient.

The conclusion from this case, at least if our currency is causal explanations of real-world phenomena, must be that the huge investment in the armchair science of PD has been inefficient.

5. The modeling literature

Next, we turn from armchair science to philosophers' treatment of it. There is something of an orthodoxy about models in philosophy of science. Paul Teller uses the term 'model view'; Peter Godfrey Smith speaks of 'the strategy of model-based science'. Much of it originates from Giere (1988), with contributions from many others too (e.g. Wimsatt, Cartwright, Suarez, Weisberg, Odenbaugh, Maki, Morgan and Morrison). Although there remain substantive disagreements, we can ignore those here. Very roughly, here are the orthodoxy's core tenets:

> -- Models are objects defined not by some intrinsic feature, but by our use of them for representational purposes. This representational role can be played by models that are abstract or physical.
>
> -- Models do not themselves make claims about the world but instead connect to the world by means of theoretical hypotheses, which typically take the form of similarity claims between the model and the target system.

-- Modeling is an activity more or less independent of theorizing, because modelers self-consciously use fictional objects to learn about the real world.

-- Models have many intellectual roles (not to mention sociological and pedagogical ones too).

For our purposes, the ontological status of models, exactly how they represent, and their relation to theories, are not directly relevant. (We will return to them later.) Rather, the only tenet relevant to the efficiency question is the last one, namely that concerning the various roles that models play. Once we are clear on those, we can start talking about the payoff from armchair science. But even then, in order to get from roles to payoff we need to know how *well* armchair models play these roles. It is here that the literature has been wanting. Philosophers have been creative and productive at articulating the many roles that models *could* play, but not so focused on whether models *actually* play them successfully, let alone efficiently. In effect, they have gone easy on these models. Perhaps naturalistic charity assumes that since they are so widespread they must be useful in some way and our task is to identify how. But the efficiency question requires a much more critical approach than that.

Here are three major roles that armchair science could play, as identified by philosophers:

1) They *explore* possibilities.

2) They *confirm* claims about relationships between variables.

3) They provide *understanding*.

There are other roles for sure, such as helping to formalize inchoate intuitions into structures and equations, making effective teaching and socialization tools, etc. But we take the above three to be the main epistemic players.

Begin with exploration and confirmation. Very roughly, these two roles differ in the strength of the credit they give models. Exploratory tools are heuristics for the discovery of interesting new hypotheses. As such they are weaker than confirmatory tools, which tell us something about the epistemic status of those hypotheses. The distinction is highly relevant to armchair science.

6. Against the confirmatory view – or, why armchair models do not explain

Confirmationism is the view that armchair science can be of explanatory value, in the sense of rationally affecting the probability of an explanatory hypothesis. Such a view is undoubtedly plausible for many targeted models but our concern here is with targetless, i.e. armchair, ones. And there is good reason to doubt that these latter can explain, widespread intuitions to the contrary notwithstanding (Author-6).

Begin with *causal* explanation, which is the most important case because it is by far the most popular candidate for the notion of explanation appropriate to economics, biology and special sciences in general. The essential initial point is that armchair models do not qualify as causally explanatory because they are false and therefore do not identify any actual causes (nor even 'approximate' causes in the manner of, say, Newtonian theory).

Perhaps though, the confirmationist replies, armchair models achieve causal explanations less directly, despite all the idealizing and constraining assumptions contained within them. We learn from PD, for instance, that particular payoff structures in an indefinitely iterated interaction cause agents to co-operate. Never mind the idealizations, says the confirmationist, the hypothesis that PD confirms is not about what actually happens, but rather about what would happen under certain conditions. That is, the model *isolates* an effect – namely co-operation – and shows that it is caused by a particular structure of two agents' interaction. That this effect may not actually obtain or be detectible very often is not fatal.

Roughly, this is the account that Uskali Maki articulated in response to criticisms of economic models (2000). Nancy Cartwright in her early writings (e.g. 1989) developed a similar view, using the notion of a tendency or a capacity to describe what models establish. On her view, the idealization of a model is crucial to its ability to shield the main effect from disturbing factors just as a good controlled experiment would.

If these accounts were correct here, armchair models would be isolating genuine causal tendencies, the manifestation of which only occurs when these tendencies are shielded from disturbing causes. But we do not think that they are correct. The fundamental problem is that we have no empirical evidence for thinking that the models really are successful at isolating capacities. If they were, we would be able to *do* the things with armchair models that we are able to do with good targeted ones, that is, combine their insights with our knowledge of disturbing factors in a given environment to predict results or to intervene. An honest look at experimental and design economics, our only opportunities for genuine tests of microeconomic models, reveals that this is not what happens. Whenever model-based causal claims are made, experimentalists quickly find that they do not hold under disturbances that were not written into the model.[29] This is why they do not deserve the honorific ‘capacity’. We suspect the same is true, for the same reason, in other fields too. When successful causal explanations are achieved, at least in economics, it turns out that it is not armchair models that do the explaining but rather extra-theoretical causal hypotheses that require empirical support.[30] To the extent that the discovery of these hypotheses is inspired or helped by armchair models, these models do get credit for a heuristic role – but that is grist for the exploratory view, not the confirmatory one.

Armchair models equally fall foul of non-causal theories of explanation too. Julian Reiss (2012, 56-59) shows just why such models do not explain according to a unificationist theory such as Philip Kitcher’s (1989). At first sight, economic models, for instance, apparently score well on a unification criterion, potentially capturing many diverse phenomena with the same basic analytical tools. But unfortunately they score very poorly on Kitcher’s stringency criterion – the relevant tools, such as utility maximization, are so thin that too little is ruled out by them. The unification that is achieved is thus too vacuous to count as explanatory. Meanwhile, not stating laws, or at least not any that are empirically vindicated in the necessary way, armchair models also clearly do not explain in the deductive-nomological sense. Finally, neither is the notion of mathematical explanation a likely savior here (Author-7), not least because it would require empirical confirmation of precisely the kind that is typically absent in armchair cases.

Generally, armchair models frequently invoke entities that do not exist, such as perfectly rational agents, infinite populations, perfectly inelastic demand functions, and so on. True enough, other sciences too invoke non-existent entities, such as the frictionless planes of high-school physics. But there is a crucial difference: the false-ontology models of physics

[29] Our own stock example is from auction design. Models say that open auctions are supposed to foster better information exchange, leading to more efficient allocation. Do they do that under any real-world conditions that we actually know about? Maybe. But we also know that introducing the smallest unmodeled detail into the set-up, for instance complementarities between different items for sale, unleashes a cascade of interactive effects. As a result, careful mechanism designers do not trust armchair models in the way they would trust genuine Galilean experiments. Nor should they. (Author-1, Author-5)

[30] This was true, for instance, in the WW1 case above, as well as in the auctions case (Author-1, Author-5), and arguably too in other cases more generally (Author-8).

and other sciences are empirically constrained. If a physics model leads to successful predictions and interventions, its false ontology can be forgiven, at least for instrumental purposes – but such successful prediction and intervention is necessary for that forgiveness (Author-4). The idealizations of armchair models, by contrast, being targetless, have not earned their keep in this way. So the problem is not the idealizations in themselves, so much as the lack of empirical success they buy us in exchange. As long as this problem remains, claims of explanatory credit will be unwarranted.

In summary, according to all relevant philosophical theory armchair models are not explanatory. What might have led to erroneous intuitions to the contrary (see also Author-6, and section 9 below)? One reason, at least in economics, is casual empiricism again: 'explanations' often amount to little more than a vague and intuitively appealing analogy between model and phenomenon. Consider, for example, the press release of the Nobel prize committee, which stated that Thomas Schelling's "analysis of strategic commitments has *explained* a wide range of phenomena, from the competitive strategies of firms to the delegation of political decision power" (Nobel 2005, italics added). Yet Schelling's models, for all their importance, have not scored any major predictive or experimental successes. Arguably, even Axelrod's detailed claims on behalf of PD are another example.

7. An aside on robustness

Recently, debate between explorationists and confirmationists has shifted in a different direction. This has been in response to another reason for explanatory skepticism, separate from those of the previous section, namely that many of the assumptions made by armchair models have nothing to do with shielding from disturbances. All sides agree that armchair models have (at least) two kinds of assumptions. The first kind defines the situation under consideration by specifying its putative causal features, such as that there are two parties bargaining, with particular information sets, particular goals, and so on. Following convention, label these substantive assumptions (Kuorikoski et al 2010). The second kind of assumption enables derivations from the first kind. An example is the mathematical properties of the functions that characterize the entities described by the substantive assumptions, such as that a utility function is continuous, twice differentiable, etc. Label this second kind, tractability assumptions. They are a problem for confirmationists because they plausibly (or even necessarily, if we know them to be false) do not describe any actual conditions that shield the operation of the mechanism described by the substantive assumptions.[32] Yet tractability assumptions could never be relaxed without losing the derivation at the heart of the model, so if they are false there is therefore no prospect of the derivation holding in any actual situation. Consequently, even a capacities-based defense of armchair models' causal-explanatory powers must founder here.

In response, confirmationists have appealed to a technique called robustness analysis, to which we now turn. It studies which assumptions are crucial to a given model derivation and which are not. If several different versions of a tractability assumption all lead to the same basic result, then the model's conclusion is said to be robust. The claim is that if many different tractability assumptions all imply the same result, then this result is probably out there in the world, in analogy with the epistemic force gained when multiple independent estimates of some physical constant all agree.

This technique was brought to the attention of philosophers by Michael Weisberg (2006) among others, but he did not explicitly endorse the view that robustness analysis is, by itself, a method of confirmation. Kuorikoski, Lehtinen and Marchionni ('KLM'), on the other hand, do argue that the mechanism specified by the model's assumptions is confirmed (in the sense of 'made more likely') to the extent that the precise form of the tractability assumptions does not matter. There is an ongoing debate about this (e.g. Author-2, Kuorikoski et al 2012). The biggest skeptical counter-argument is, roughly speaking, that because many (even most) of the tractability assumptions common to all the versions of the model cannot be de-idealized, so there is no warrant to interpret relations in the model causally – as it were, even the robust intersection of many toy models remains just a toy model. Therefore even robust models remain empirically, and thus explanatorily, questionable. KLM respond that that just shows that robustness analysis does not *fully* confirm, not that it doesn't confirm at all.

Our own view is that non-empirical robustness analysis cannot offer confirmation and is at best of exploratory value. But suppose we granted the contrary to the confirmationist. Even then, that still leaves hanging a crucial question: how *much* confirmation? Is robustness analysis an efficient method? That is, even if robustness analysis did have some confirmatory power, still it might be the case that economists and ecologists are doing way more of it than

[32] This is why Cartwright in her later work talks about these models being "overconstrained" (2010).

is reasonable. After all, many non-substantive assumptions in armchair models cannot be subjected to robustness analysis anyway, e.g. because analytic solutions are absent or because mathematical tools break down (Author-2). There is no way to answer the question seriously without comparing robustness analysis to *other* routes to explanation and thus broadening the discussion considerably.

Confirmationists often appeal to the fact that economists and ecologists cannot conduct experiments and hence need modeling and robustness because there is nothing else they can do. In effect, this takes disciplinary inertia as given by nature and unchangeable. But it is neither as, for instance, the rise of experimental economics, field methods, and the use of randomized controlled trials in development economics attests (see also section 11). It is perfectly conceivable that these latter methods along with traditional anthropological observation might be just as, or more, effective for increasing our confidence in given causal claims. The WW1 truces are a perfect illustration: much more explanatory value was achieved, and much more efficiently, by Ashworth's historical research than by the application of the huge PD literature. The debate surrounding robustness analysis has ignored the efficiency question.

8. The exploratory view

What exactly is the exploratory view of models? There are early clues to it in Sidney Morgenbesser's 1956 PhD thesis at the University of Pennsylvania, which spoke of "schematic expressions" as a way of characterizing theories in social science (Morgenbesser 1956, chapter 2). Such an expression is merely a framework, neither a statement of a mechanism nor an explanation. It can become the latter only with additional content. In Wimsatt and Humphreys we find the notion of a 'template'. Wimsatt spoke of false models as "means to truer theories", as the title of his much cited (1987) paper puts it. More recently philosophers have talked of models delineating the space of possibilities (Forber 2010), being tools of conceptual exploration (Hausman 1992), providing potential or possible explanations (Gruene-Yanoff 2009, Aydinonat 2008, Forber 2010; see also Brandon 1990), or serving as raw material out of which explanations can be constructed (Odenbaugh 2005, Author-1, Author-5).

We will not rehearse here any specific exploratory account. It is sufficient to mention two of their core features. The negative one is that armchair models do not by themselves provide causal explanations. The positive one is that they can help us achieve them indirectly, hence the talk of ‘templates’, ‘frameworks’, and ‘open formulas’. A key feature of a template is that it is not itself up for confirmation, or at least not in the way that fully fledged causal claims are. The flip side is that when one is used during the development of a successful explanation, the warrant from that eventual success goes to the explanation – but not to the template.

A separate claim that can also be part of an exploratory view is that armchair models play an *agenda-setting* role. The influence of PD in economics may be a good example of this. Recent work in philosophy of biology supports it too. In particular, it has been claimed that armchair models serve to structure and inspire subsequent research by providing concepts and ideas, but that they do not themselves tell us what to be realist about nor do they themselves explain (Pincock 2012, see also work by Jay Odenbaugh and Patrick Forber). It is only the subsequent research, often featuring close empirical study, which leads to explanations. This picture, of course, is more or less exactly a statement of an exploratory rather than confirmatory role.

Overall then, where do things stand with regard to the efficiency question? If the exploratory view is correct, armchair modeling is a tool for discovery, no more than that, and we don’t know how good a one. If the confirmationist view is correct, armchair modeling is also a tool for confirmation too, but still we don’t know how good a one. Either way, the efficiency problem looms larger than ever. There are lots of good tools of discovery and confirmation and yet armchair science is sticking to just one! Is this defensible? The point is that merely deciding between the exploratory and confirmatory views does not yet address the efficiency question. Why this persistent lacuna in the modeling literature? Before finally answering that, we need to examine one further issue.

9. Understanding: no third way

Return now to the third epistemic role mentioned earlier. Might the notions of ‘understanding’ or ‘insight’ represent a route to epistemic value distinct from either explanation or heuristic value? Armchair models are often claimed to provide just such benefits. Unfortunately, the mere subjective feeling of understanding is a very unreliable indicator of explanation (Trout 2007, Author-6). In order for claims of understanding to support armchair science, therefore, it is necessary to demonstrate some value for them quite independent of explanation.

A large segment of the literature denies that this can be done. This was the position originally of Hempel and more recently is endorsed, in one form or another, by Trout, de Regt (2009), Strevens (forthcoming), Khalifa (2012), and others – including us. On this view, regardless of the understanding or insight that exploratory models might subjectively seem to give us, if they do not also give us explanations then in fact they are of no value.[33]

The opposing view is that understanding can offer something independent of explanation. But what is this additional thing? The leading answer is *how-possibly* explanations, which are taken to be distinct from actual ones. The original notion of a how-possibly explanation was advanced by William Dray (1957). He had in mind an explanation that did not show why an outcome should be expected or had to happen, but instead merely showed why it was possible that it had happened. In a sense its point is negative, explaining why an impossibility or low-probability claim is incorrect. But contemporary writers usually have in mind something slightly stronger. Gruene-Yanoff (2009), for instance, holds that models such as Schelling's checker-board representation of race and location serve to suggest causal hypotheses about the actual world. They do this by establishing that such hypotheses could *possibly* hold – in an idealized, and hence non-actual, world. These hypotheses, as in the Schelling case, are often surprising and therefore potentially useful. Nevertheless, the mere fact that a result holds in a non-actual world is not itself sufficient to warrant the claim that it holds in the actual one, nor therefore that it explains an actual-world outcome, as Gruene-Yanoff concedes. But then it is not clear what other value it can provide beyond the heuristic one of suggesting hypotheses about the actual world. There is no extra third path here besides explanation and heuristic.

[33] In our view, two separate errors can occur here: first, thinking that the feeling of understanding implies explanation when in fact it doesn't. And second, when faced with good arguments that it doesn't, insisting then that in that case there must be some *other* epistemic value, distinct from explanation, indicated by the feeling of understanding. Both errors reflect a refusal to acknowledge the frequent epistemic worthlessness of feelings of understanding, and we do not think that this refusal stands up to philosophical scrutiny.

There is a parallel discussion in philosophy of biology. In particular, a tradition in evolutionary biology interprets the idealized models of population genetics to be how-possibly explanations, and a literature has arisen about exactly what these explanations in turn amount to. Brandon (1990) sees them as incomplete attempts at actual-world explanations. Subsequent empirical work then attempts to fill in the details, to decide whether the potential explanation given by the model can indeed be turned into an actual one. Notice that again this leaves no third way: a model can have a heuristic role, with the promise of eventually being converted into an explanation. But if it is not converted into an explanation, there is no separate epistemic value left over. Others too advocate variations on a similar approach, viewing the models not as confirmatory in themselves but rather as offering a library of possible explanations, further empirical work then being required to establish any particular explanation as correct (Resnik 1991, Reydon 2012).

A different justification for armchair models is that although not explanatory in themselves, they do form essential *parts* of actual-world explanations (e.g. Weisberg 2013, see also Forber 2010). A well-known quotation from R.A. Fisher makes the point: "No practical biologist interested in sexual reproduction would be led to work out the detailed consequences experienced by organisms having three or more sexes; yet what else should he do if he wishes to understand why sexes are, in fact, always two?" (quoted in Weisberg 2013, 121).

This appeals to contrastive theories of confirmation and explanation. We are sympathetic to these[34], and so agree that part of any actual explanation is an account of a contrast case, which in turn may well require a model of a counterfactual scenario. So, for example, a model of three sexes may indeed form part of the explanation for why in the actual world there are only two. Nevertheless, like all such counterfactuals, these scenarios must be in the closest possible world, i.e. as 'realistic' as possible. In the sex example, we would want to investigate whether three sexes could have evolved in an ecological world *as similar to the actual one as possible* – not in some simplistic or abstract world far removed from our own. To fulfill their contrastive role in actual explanations, therefore, investigations of counterfactual scenarios cannot be idealized any more than can those of actual ones. Why else would experiments in general be controlled so carefully, if not to ensure that the actual

[34] And indeed, in one of our cases, to a contrastive theory of causation too (Author-3).

control observation was as good a proxy as possible for the counterfactual observation relevant to explanation? Accordingly, appeal to contrastive explanation offers no salvation to armchair science.

Overall, it has proven elusive just what understanding or insight could amount to that is not either heuristic value or else the valueless mere subjective feeling of understanding.

10. Why has efficiency been neglected?

Why, despite its importance, has the efficiency question been largely neglected by the modeling literature? Several explanations of this lacuna are possible: perhaps it is too difficult to assess what would have happened with alternative research strategies; or perhaps naturalistically inclined philosophers will tend to rationalize existing epistemic practice if possible rather than critique it wholesale.[35] Two further explanations now arise from the analysis of this paper.

First, there has been a widespread implicit commitment to the confirmatory view of models. Although often appropriate for targeted models, this commitment is much less appropriate for armchair ones. We think that this in turn has had two important consequences:

1) Armchair science's actual role has not been appreciated accurately – it is exploratory not confirmatory. As a result, there has been insufficient critical analysis of whether it has fulfilled this exploratory role efficiently. Has it been a worthy agenda-setter? Does the often meager empirical bounty justify such resources and prestige?[36]
2) There has been a great focus in the literature on issues of *representation*. In confirmatory cases such a focus is defensible, as it bears directly on explanation: successfully representing a cause immediately yields an explanation. But in non-confirmatory cases no model is explaining anyway; what matters to explanation is whether an eventual causal hypothesis represents, not whether the initial heuristic model does. A large portion of the modelling literature is therefore irrelevant to the armchair case.

[35] We thank an anonymous referee for emphasizing these possibilities.

[36] True enough, the efficiency question arises for confirmatory models too. But given their frequent success at generating explanations and predictions, failure at these tasks tends to cut short a line of enquiry relatively quickly. The same is not true for armchair models, so in practice the efficiency question becomes much more pressing there. But this can only be recognized once we have an accurate diagnosis of armchair models' true role, which is exactly what a concentration on the confirmatory view misses.

The second explanation is that, as we have seen, some of the literature gives credit to armchair models for providing understanding, independent of explanation. But in our view:

1) Such credit is mistaken.
2) It serves merely to insulate armchair models from rightful criticism for their poor performance regarding explanation and agenda-setting.

11. Conclusion: a way to proceed

What would answering the efficiency question require? It would mean evaluating a counterfactual claim, namely how a scientific project would look if instead of armchair science it had adopted a different methodology. In practice, the main contrasts to armchair science are field or experimental methods. In the case of the WW1 truces, for instance, the main alternative to PD analysis was the detailed historical research by Ashworth. The critical question is: would the goals of the field have been achieved better via these alternative means? In the WW1 case, they clearly would have. Recently political science and sociology have seen a renaissance of new qualitative and quantitative methods of causal inference such as case study methods, process tracing, and small-N causal inference (Mahoney 2012, Bennett and Elman 2006, Ragin 2008). Field, natural and quasi-experiments are gaining ground across economics, geography and biology (Meyer 1995, Diamond and Robinson 2011). The causal modeling literature in philosophy promises causal inference from statistics free from the idealizations of current armchair models (Pearl 2000, Spirtes et al 2000). Of course, each of these methods will have its own costs: cases might be too idiosyncratic, field experiments improperly controlled, process tracing too dependent on one's theoretical perspective, causal inference hostage to its own idealizations, and so on. But armchair science comes with costs too, as we have seen.

Is the efficiency question simply too complex to answer? We grant that it is indeed complex, and that it might be unrealistic to expect much agreement. But we disagree that the question is flawed in itself or not worth asking. Besides, in practice it is unavoidable. After all, whenever a researcher selects their methodology, they are implicitly answering it already. The proliferation of armchair science is in effect a bet that has *already been made* about the answer to the efficiency question. It is surely best to consider such an important matter explicitly. Given the prima facie case that it may represent a bad bet, consideration of armchair science is all the more urgent. Does the proliferation of armchair science represent a massive misuse of intellectual resources? Philosophy of science should be attending to the efficiency question; it is too important to ignore.[38]

[38] Here we see ourselves as following Nancy Cartwright's call: "My aim then is to urge us to direct our efforts away from the more abstract questions that usually entertain us—from highly general questions of warrant (like do we have reason to believe our theories are true rather than merely empirically adequate, is simplicity a symptom of truth, the 'principal principle', and the like) to much more specific questions about particular methods and their problems of implementation, their range of validity, their strengths and weaknesses, and their costs and benefits." (2006, 982)

References

Ashworth, J. (1980). *Trench Warfare 1914-1918*. London: MacMillan.
Axelrod, R. (1984). *The Evolution of Co-operation*. London: Penguin.

Aydinonat, N. (2008). *The Invisible Hand In Economics: How Economists Explain Unintended Social Consequences*. Routledge.

Bennett, A., and C. Elman (2006). ‘Qualitative research: Recent developments in case study methods’, *Annu. Rev. Polit. Sci.* 9, 455-476.

Blaug, M. (2002). ‘Ugly currents in modern economics’, in U. Maki (ed) *Fact and fiction in economics*. Cambridge, pp35-56.

Bowles, S. and H. Gintis (2008). ‘Co-operation’, in *The New Palgrave Dictionary of Economics* (eds) Steven Durlauf and Lawrence Blume. Palgrave.

Brandon, R. (1990). *Adaptation and Environment*. Princeton: Princeton University Press.

Cartwright, N. (1989). *Nature’s Capacities and their Measurement*. Oxford: Clarendon.

Cartwright, N. (2006). ‘Well-Ordered Science: Evidence for Use’, *Philosophy of Science* 73, 981-990.

Cartwright, N. (2010). ‘Models: Parables v Fables’, in Frigg, R. and M. Hunter (eds.) *Beyond Mimesis and Convention: Representation in Art and Science,* New York: Springer, pp19-31.

Diamond, J. and J. Robinson (eds) (2011). *Natural Experiments in History*. Harvard.
Dray, W. (1957). *Laws and Explanation in History*. Oxford: Oxford University Press.

Dupré, J. (2001). *Human Nature and the Limits of Science*. Oxford

Erickson, P., J. Klein, L. Daston, R. Lemov, T. Sturm, and M. Gordin (2013). *How reason almost lost its mind: The strange career of Cold War rationality*. Chicago.

Forber, P. (2010). ‘Confirmation and explaining how possible’, *Studies in History and Philosophy of Science Part C* 41, 32-40.

Frigg, R., and S. Hartmann (2012). ‘Models in Science’, *The Stanford Encyclopedia of Philosophy* (Fall 2012 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2012/entries/models-science/>.

Gelman, A. (2008). ‘Methodology as Ideology: Some comments on Robert Axelrod’s *The Evolution of Cooperation*’, *QA-Rivista dell'Associazione Rossi-Doria*, 167-176.

Giere, R (1988). *Explaining Science*, Chicago: University of Chicago Press.

Govindan, S. and R. Wilson (2008). ‘Nash equilibrium, refinements of’, in *The New Palgrave Dictionary of Economics* (eds) Steven Durlauf and Lawrence Blume. Palgrave.

Gruene-Yanoff, T. (2009). ‘Learning from Minimal Economic Models’, *Erkenntnis* 70, 81-99.

Hausman, D. (1992). *The Inexact and Separate Science of Economics*. Cambridge: Cambridge University Press.

Hindriks, F. (2006). ‘Tractability Assumptions and the Musgrave-Mäki Typology’, *Journal of Economic Methodology* 13, 401–23.

Hutchison, T. W. (1941). ‘The significance and basic postulates of economic theory: a reply to Professor Knight’, *The Journal of Political Economy*, 732-750.

Khalifa, K. (2012). ‘Inaugurating Understanding or Repackaging Explanation?’ *Philosophy of Science* 79, 15-37.

Kitcher, P. (1989). ‘Explanatory Unification and the Causal Structure of the World’, in P. Kitcher and W. Salmon (eds) *Scientific Explanation*, Minneapolis: University of Minnesota Press, pp410–505.

Kuhn, S. (2009). ‘Prisoner’s Dilemma’, *The Stanford Encyclopedia of Philosophy* (Spring 2009 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2009/entries/prisoner-dilemma/>.

Kuorikoski, J., A. Lehtinen and C. Marchionni (2010) ‘Economic Modelling as Robustness Analysis’, *British Journal for Philosophy of Science* 61, 541-567.

Kuorikoski, J., A. Lehtinen and C. Marchionni (2012). ‘Robustness analysis disclaimer: please read the manual before use!’ *Biology and Philosophy* 27, 891-902.

Lawson, T. (1997). *Economics and Reality*. Routledge.

Maas, H. (2010). ‘Sorting Things Out: The Economist as an Armchair Observer’, in L. Daston & E. Lunbeck (eds.), *Histories of Scientific Observation*. Chicago, pp206-229.

Mahoney, J. (2012). ‘The logic of process tracing tests in the social sciences’, *Sociological Methods & Research* 41.4, 570-597.

Mäki, U. (2000). ‘Kinds of assumptions and their truth: shaking an untwisted F-twist’, *Kyklos* 53, 317–35.

Meyer, B. (1995). ‘Natural and quasi-experiments in economics’, *Journal of business & economic statistics* 13.2, 151-161.

Michihiro, K. (2008). ‘Repeated games’, in *The New Palgrave Dictionary of Economics* (eds) Steven Durlauf and Lawrence Blume. Palgrave

Morgenbesser, S. (1956). *Theories and Schematas in Social Sciences*. PhD Dissertation, University of Pennsylvania.

Musgrave, A. (1981). ‘Unreal Assumptions in Economic Theory: The F-Twist Untwisted’, *Kyklos* 34, 377-387.

Nobel (2005). Press Release: The Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel 2005, 10 October 2005.

Odenbaugh, J. (2005). ‘Idealized, Inaccurate, and Successful: A Pragmatic Approach to Evaluating Models in Theoretical Ecology’, *Biology and Philosophy* 20, 231-255.

Pearl, J. (2000). *Causality*. New York: Cambridge University Press.

Pincock, C. (2012). ‘Mathematical models of biological patterns: Lessons from Hamilton's selfish herd’, *Biology and Philosophy* 27, 481-496.

Ragin, C. (2008). *Redesigning social inquiry: Fuzzy sets and beyond*. Chicago: University of Chicago Press.

de Regt, H. (2009). ‘The epistemic value of understanding’, *Philosophy of Science* 76.

Reiss, J. (2012). ‘The explanation paradox’, *Journal of Economic Methodology* 19, 43-62.

Resnik, D. (1991). ‘How-possibly explanations in biology’, *Acta Biotheoretica* 39.

Reydon, T. (2012). ‘How-possibly explanations as genuine explanations and helpful heuristics: A comment on Forber’, *Studies in History and Philosophy of Science Part C* 43, 302-310.

Rodgers, D. (2011). *Age of Fracture*. Harvard.

Schultz, K. (2001). *Democracy and Coercive Diplomacy*. Cambridge: Cambridge University Press.

Spirtes, P., C. Glymour, and R. Scheines (2000). *Causation, Prediction, and Search* (2nd edition). Cambridge: MIT Press.

Strevens, M. (forthcoming). ‘No understanding without explanation’, *Studies in History and Philosophy of Science Part A.*

Sunstein, C. (2007). ‘Of Montreal and Kyoto: a tale of two protocols’, *Harvard Environmental Law Review* 31.1.

Trout, J. (2007). ‘The psychology of scientific explanation’, *Philosophy Compass* 2, 564–591.

Velupillai, K. (2005). ‘The unreasonable ineffectiveness of mathematics in economics’, *Cambridge Journal of Economics* 29.6, 849-872.

Weisberg, M. (2006). ‘Robustness analysis’, *Philosophy of Science* 73, 730–742.

Weisberg, M. (2013). *Simulation and Similarity: Using Models to Understand the World*. Oxford University Press.

Wimsatt W. (1987). ‘False Models as a Means to Truer Theories’, in M. Nitecki and A. Hoffman (eds.) *Neutral Models in Biology*. London: Oxford University Press, pp23–55.

Anonymised author references

-- Author-1
-- Author-2
-- Author-3
-- Author-4
-- Author-5
-- Author-6
-- Author-7
-- Author-8