3

Reduction in Real Life

Peter Godfrey-Smith

1. INTRODUCTION

The main message of the paper is that there is a disconnect between what many philosophers of mind think of as the scientific practice of reductive or reductionist explanation, and what the most relevant scientific work is actually like. I will sketch what I see as a better view, drawing on various ideas in recent philosophy of science. I then import these ideas into the philosophy of mind, to see what difference they make.

At the end of the paper I address a possible objection: the familiar package of ideas I reject in the philosophy of science should not be lightly discarded, because other popular views on fundamental issues depend on positions that I want to reject. I reply that those apparently attractive further ideas are not worth holding onto.

So the paper begins with issues in the philosophy of science: reduction, laws, mechanisms, and models. It then turns to philosophy of mind, and returns to broad themes in the philosophy of science at the end.

2. MECHANISMS, MODELS, AND REDUCTION

In this section I contrast two packages of views about reduction and related issues. One is a "traditional" view, the other an "alternative" view. The traditional view is not just the deliverances of older philosophy of science, however. It is a package of ideas that draws on traditional philosophy of science (especially late logical empiricism), but that has been augmented and modified by philosophers

I am indebted to those at the Aarhus conference in 2005 for helpful comments. I am grateful also to Ned Block, Carl Craver, Steven Horst, Kim Sterelny, and the editors of this collection for criticism of an earlier draft.

of mind. The "alternative" view draws on recent philosophy of science, but the position presented will be my own blend of ideas that derive from several different camps.

Here is the package of views about science that I will refer to as standard or traditional, in much philosophy of mind.

(i) Theories are essentially networks of generalizations.
(ii) The best theories feature, as central components, forward-looking causal laws. These laws treat future states and events in their domain as a function of past states.
(iii) We have theories of this kind at different "levels". The lower-level ones either reduce the higher-level ones, or are linked by a weaker explanatory relation (perhaps supervenience of facts or properties at the two levels).
(iv) Physics is at the bottom of this hierarchy of levels. Above it we find chemistry, biology, psychology, and the social sciences.
(v) There is a close link between the notions of law, natural kind, counterfactual dependence, confirmation, and explanation. In particular, not all true universal generalizations specify laws. Those that do specify laws contain predicate terms that pick out natural kinds. Laws support counterfactuals, unlike non-lawlike generalizations. Law-like generalizations, and only them, can be confirmed by their instances. Laws also have a special role in explanation.

There is plenty of debate surrounding these ideas, within mainstream thinking. But some core parts of the picture remain constant across medium-sized differences. An especially important one is the idea that genuine scientific understanding involves knowledge of laws. This package of ideas also has a fairly consistent influence on debates about the relations between "levels" in a total scientific picture. "Reduction" is associated with strong inferential relationships between levels, and the threat of the dispensability of higher-level descriptions. If reduction is possible, the coordination between levels is achieved by something like an additional set of "bridge" laws. Against this we have projects seeking more moderate options; supervenience is seen as a looser relation between levels than reduction, but one potentially preserving physicalism. Much discussion then focuses on the status of higher-level laws, which might capture patterns that cannot be seen from the point of view of a lower-level description.[1]

What is wrong with this package? The answer I offer is not intended as a description of all of science. The aim is to describe sciences that connect most directly to naturalistic philosophy of mind—roughly speaking, biology

[1] Influential versions of the view I am calling "traditional" can be found in Fodor (1974) and Kim (1993). As should be clear, what anti-reductionists sometimes call a "received" reductionist view is included with many forms of anti-reductionism within the larger category I am calling "traditional". For a detailed treatment of supervenience, see McLaughlin and Bennett (2005).

and psychology. If we focus on those areas, then the standard package is almost entirely wrong. It is false that these parts of science are organized around laws. In particular, it is false that the usual form theoretical knowledge takes is a set of forward-looking law-like causal principles that directly describe real systems. Laws appear occasionally, but they are minor players, with none of the organizing role they play in physics. I also reject the usual story about the links between laws, kinds, counterfactuals, confirmation, and explanation, and reject some popular accounts of the relations between levels.

At an earlier time in the history of science it might have been possible to think that these facts reflect badly on the biological sciences themselves. But that would be a difficult case to make now, given what biology has done and become in the last sixty years. And importantly, biology has not achieved its recent progress by moving *closer* to the traditional philosophical ideal.

I now start to present an alternative view, via three moves that draw on different parts of recent philosophy of science.

The first move is drawn (in moderated form) from John Dupré's book *The Disorder of Things* (1993). Dupré argues that when philosophers write about reductionist work in science, they imagine that what we get from such work, when it succeeds, is a low-level theory that tells us *what will happen*, in a system of a certain kind. That is, philosophers imagine science giving us a body of information that tells us how later states in a system are a function of earlier states. The "reductionist" thinks we are learning (or will one day learn) low-level accounts of this kind for the case of complex macroscopic systems like organisms and thinking agents. "Anti-reductionists" deny that this is happening, or deny that it will be possible.

For Dupré, this is a mistaken view of what actual reductionist work in many sciences looks like. He argues that in biology, and other fields in what we might call the "mid-level" part of science, we often have good reductionist theories that tell us a particular kind of thing. They tell us *how* various complex systems do what they do. But they don't tend to tell us, in any detail, *what* the systems will do. That is, we do not find low-level dynamic theories making specific predictions about how the system will change over time, what it will do next. To address such dynamic questions we tend to use a higher-level framework, even when we have a genuine reductionist understanding of the higher-level processes.

This is a useful re-orientation of the discussion. When we look at successful reductionist research programs in areas like biology, we do see an accumulation of information about how various biologically important processes occur. We now have a good understanding of processes like photosynthesis, respiration, protein synthesis, the transmission of signals in the brain, the action of muscles, the immune response, and so on. This sort of work can reasonably be, and often is, described as reductionist. We are taking a high-level process or capacity, and explaining how it works in terms of lower-level mechanisms and entities. In many of these cases, the "lower" level is the level of specific molecules or lower. (In cases like photosynthesis, for example, electrons themselves figure in the story.) But in all these cases, our theory does not take the form of a forward-looking dynamic account. The theory does not say: given this specific configuration of DNA molecules, enzymes, and other cellular mechanisms, the following processes *will* occur. Or: these processes will occur with 0.8 probability. An attempt to give a low-level story of that kind would be overwhelmed by the complexity of the system.[2] But that complexity does not overwhelm our ability to explain how things happen.

We should not go too far with Dupré, however. He would use these ideas to take us in the direction of libertarianism, and a very deflationary account of the bearing of low-level sciences on our understanding of human life. We must also be careful not to overstate the size of the separation between knowledge of how things work and knowledge of what will happen. Our knowledge of how things work includes knowledge of capacities and tendencies that can be the basis of predictions and interventions. (If this were not so, there would be acute problems in testing hypotheses.) With this knowledge we can often also formulate new generalizations, about both the characteristic behaviors of the system and how it will respond to abnormal circumstances. But these generalizations do not usually take the form of laws, and are not the central theoretical principles that organize our knowledge. Instead they appear as useful consequences and spin-offs from the growth of our knowledge of how things happen. A further qualification is that it would be a mistake to extend this picture to all of science. (I am not saying that Dupré does this.) These ideas are not intended to give a new account of the relation between thermodynamics and statistical mechanics, or even explanations of chemical reactions.[3]

The important thing is the way that Dupré's criticism re-orients the discussion for philosophy of mind. It is true, as Dupré says, that philosophers routinely picture the advance of knowledge in areas of lower-level science that are relevant to human thought and agency as the accumulation of forward-looking laws. Often, this means that the philosopher must merely *imagine* a future state of knowledge where we have such laws.[4] There is nothing wrong with imagining such a state, and imagining how this kind of knowledge might impact on us. But that state is indeed an imaginary one, and it is not a very natural near-term extrapolation from where we are now. It is not the actual form of well-developed present-day sciences that have a reductionist character. Molecular biology is, by

[2] Here I mean a direct and literal description of what will happen given a certain real-world configuration, not what would happen in an idealized model system that imagines away much of the complexity. See the discussion of models later in this section.

[3] Chemistry may be an interesting in-between case, from the perspective of this paper. For example, Stemwedel (2006) gives an account of the structure of the explanations of individual chemical reactions that includes an interesting mix of forward-looking principles explicitly christened "laws", and information that (at least to me) fits better a models-and-mechanisms framework of the kind discussed below.

[4] See, for example, Sober (1999), and commentary on that paper in Godfrey-Smith (1999).

any measure, an advanced and well-developed branch of science. Perhaps one day in the future it will be organized around a set of forward-looking laws of the kind that philosophers like to imagine. But at present, it is not organized that way at all. It is organized as knowledge of how things work, how things happen, and what structures in living cells do what.

So how might we give a better philosophical account of the content of this kind of scientific knowledge? The second idea I draw on in this section is the theory of "mechanistic explanation" recently developed by a collection of philosophers including Bechtel, Machamer, Craver, Darden, Richardson, and others. I will call these philosophers "new mechanists," and will draw especially on the summary "Thinking About Mechanisms" (2000) given by Machamer, Craver, and Darden.[5]

The aim of the new mechanists is to give a detailed account of what they take to be the predominant mode of explanation in large parts of biology, cognitive science, and some other areas. Neuroscience is often a particular focus. It would probably be appropriate to add a qualification to the analysis the new mechanists offer, and present it as an account of how these sciences work *when* they are in a reductionist mode, which they often are. (Work in different modes will be discussed briefly later in this section.)

The distinctive features of the new mechanists' account are as follows. First, they give an account of the ontology employed by these sciences, an ontology of mechanisms, activities, capacities, and processes. (I would add to their account an emphasis on structures and structural description.) Second, their account is antagonistic towards the traditional philosophical emphasis on laws, and also towards views of causation that are influenced by a focus on laws. Third, they give a very simple treatment of "levels" in these sciences. Levels are understood in terms of ordinary part–whole relations. (In the next section I discuss how a view of levels can diverge from this simple idea.)

The new mechanists take as data such scientific achievements as the explanation of protein synthesis, and the explanation of the transmission of signals across synapses between neurons. This is scientific progress, if anything is. Their argument is that there is little or no apparent role for laws in these sorts of achievements. What *does* figure essentially is a form of explanation in which complex processes are explained in terms of the capacities and organization of lower-level parts.

In mainstream philosophy of mind, the closest cousin to this picture is Cummins' discussion of functional analysis (1975), and some of his follow-up work (2000). But the new mechanists are aiming for more contentious and

[5] See also Bechtel and Richardson (1993), and Bechtel and Abrahamsen (2005). Wimsatt (1972) is an important precursor. The term "new mechanists" is one of several that seems to float around the movement. A more amusing one is Andrew Hamilton's "mechanistas". The generalizations I give here about new mechanism do have exceptions; the movement is new and quite heterogeneous.

general conclusions, as is seen in the negative treatment of laws. Their treatment of causation is also affected by these commitments. A very mild version of the mechanism-oriented view would be one that emphasized mechanisms as the currency of scientific work in these fields, but then employed a traditional regularity or nomological account of causation in the background. That is not the approach of the new mechanists. In Machamer, Craver, and Darden, in particular, the new mechanist view is associated with what we can call a "production-oriented" view of causal relations and their role in explanation. The obvious contrast is with regularity views, but we can also contrast production-oriented views with abstract difference-making accounts, that use counterfactuals and similar constructs to analyze causation (Lewis 1973; Collins, Hall, and Paul 2004). For the new mechanists, all such difference-making facts must be grounded in mechanistic facts. This last set of ideas might suggest that new mechanism is getting too close to *old* mechanism, in which a very restricted range of physical relationships are seen as scientifically legitimate. But new mechanism, properly configured, leaves it open what kinds of relations will be important in such areas as physics and physical chemistry.

The new mechanists have done a good job of giving a positive account of a kind of scientific work that had been badly misdescribed by earlier philosophy of science. They have given a fairly accurate, and philosophically informative, account of mature scientific work within the reductionist family of projects in biology and other "mid-level" sciences.

I should note that, once this picture is in place, the fate of the term "reduction" can become unclear. At a 2005 symposium on the relation between philosophy of science and philosophy of mind at Boston University, Steven Horst and William Bechtel gave talks that, on these points at least, presented fairly similar pictures of how the relevant areas of science operate, and the deficiencies of more traditional views. But Horst saw his message as anti-reductionist; his talk was titled "Beyond Reduction". Bechtel, in contrast, saw himself as describing what real reductionist work, as opposed to the philosophers' image of it, is like. In discussion, Bechtel (and Paul Churchland) argued that the term "reduction" is entirely natural for this kind of scientific work. This is work that engages in the explanation of high-level capacities in terms of lower-level ones, explanation of the big in terms of the small, and it is what most scientists themselves see as reductionist work. It is only if we tie the term "reduction" to the old philosophical picture that this kind of work could be called anti-reductionist. Terminology *per se* is not very important, of course, but I agree with Bechtel and Churchland on this point.

The third idea I will use comes from yet another camp in recent philosophy of science, that looks at the role of models and model-building in scientific theorizing.

One strand in recent philosophy of science uses the notion of a model, in roughly the logician's sense, to analyze *all* scientific theorizing. This is the "semantic view" or model-theoretic view, of theories (Suppe 1977, Van Fraassen

1980). That is not the set of ideas I will draw on; I make use of a related line of thought. This view holds that there is a particular kind of science that seeks to represent the world using models. Model-based science is a *strategy*, and often a response to a certain kind of problem.[6] In understanding this work, the logician's sense of "model" is not the right one to use. We need a different concept.

The new mechanists have not generally embraced these ideas.[7] And in the present context, there is a convenient way to approach the relation between the two. Consider the general kind of scientific work that the new mechanists discuss, but focusing on what it tends to look like in its early stages. These are stages where we do not know much about the system and its workings. Our eventual goal is an account of the structure and operation of some set of mechanisms. The goal is a list of real parts and their capacities. So in the early stages, we are dealing with hypothesized parts and their capacities. Machamer, Craver, and Darden (2000) do say a little about this stage. Their term for the products of this early work is "mechanism sketches". These are schematic mechanisms with some black-boxes that need to be filled in. In at least many areas though, the common scientific response to problems of this kind is *model-building*, in a specific sense. Model-based science features an "indirect" strategy for the representation and investigation of unknown systems. A model-builder first describes a hypothetical structure, usually a relatively simple one, and then considers similarity relations between this structure and the real-world "target" system that he is trying to understand.

A good initial sketch of this process was given by Giere (1988). Giere's aim was to describe *all* scientific theorizing, and his starting point was physics as presented in textbooks. The attempt to capture all theorizing in these terms was almost certainly over-reaching. And this paper will not try to defend any claims about physics. But Giere did succeed in giving a compact but informative sketch of one important kind of theoretical work in science, a kind that is relevant to fields impinging on philosophy of mind. This is the style of science in which a paper might begin: "Imagine an infinite population of asexual organisms . . ." "Consider a feed-forward neural network with one layer of hidden units and the following learning rule . . ." In my treatment of model-based science, I take this phenomenon at face value. What the model-builder is doing is specifying and inviting us to consider a hypothetical or fictional system (or class of systems), which he or she can describe exactly. Having done so, we can then consider ways in which the behavior of this hypothetical system might cast light on the behavior of a real system.

[6] See Godfrey-Smith (2006) and Weisberg (2006).

[7] There are exceptions to this. One is the far-seeing Wimsatt (1972). Another is Glennan (2005), but Glennan's paper could be better described as an application of some ideas from the "semantic view of theories" to the case of mechanistic description (as seen in his enthusiasm for state space descriptions of all models, whether they explicitly feature equations or not). I should also note that Horst, whose talk at Boston in 2005 is discussed above, combined a mechanistic view with an emphasis on modeling.

Model-based science gets part of its strength from a certain kind of flexibility, resulting from the indirect strategy employed. In model-based science, a lot of day-to-day discussion is about the model system—the hypothetical or imaginary system—itself. Two scientists can use the same model to help with the same target system, while having different views about the extent and character of the similarity that the model has to the target. One might see the model as a purely predictive device. The other might see it as a causal map, a good representation of a hidden dependency structure inside the target system. And there is no dichotomy between a single realist and single instrumentalist attitude here, but a spectrum or space of possible attitudes on how model and target might be related.

How does this relate to the new mechanists' account? The situation might be summarized like this: in the sciences the new mechanists are interested in, the desired *end-point* is often the sort of conceptual structure that they describe. But a different story should often be told for the early stages—the stages where people do not have a good handle on the components and their capacities. In that situation, model-building is a natural and common approach that is taken. This is not usually permanent. A description of a model can pass into a mechanistic description.

So we now have a sketch of how scientific work proceeds in the case of *early* stages of *reductionist* work on *complex* systems. In that situation, the currency of scientific work is often *models of important processes*; models of possible mechanisms, possible dependency structures, that might in time give us an account of the real mechanisms. Once we say it like this, it becomes apparent that this is what a large proportion of work in the cognitive sciences is concerned with today—models of learning, models of numerical cognition, models of the processing of syntax. And this really is quite different from the picture we would get by applying the standard philosophy of science that philosophers of mind tend to assume. Everyday work is not concerned with the assessment of hypothesized laws governing lower-level entities, with some explanatory relation to higher-level laws. Instead, models of important processes are the currency. The aim of the modeling is to eventually give an account of actual mechanisms and how they work. In the meantime, people model, with the hope that models can evolve into direct descriptions of mechanisms.

Here I have emphasized a transition from modeling to mechanistic description. I see this as specifically important for the kind of science that is relevant to philosophy of mind. But model-based science is not always a way-station. This strategy can be retained when the scientific field is mature. Idealized models may then be developed and retained for their useful generality (Levins 1966), and also for the advantages that come from simplicity. An idealized model system may be described by compact and comprehensible dynamical principles that express the future as a function of the past.

So the third and final main idea of this section is the importance and distinctiveness of modeling. Before moving on, though, I will make some further comments about generalizations and laws.

Antipathy to standard philosophical ideas about laws in science has been a theme of the paper so far. But surely it cannot be denied that scientific work of all kinds constantly deals in generalizations. Is the "alternative" view trying to deny the scientific importance of generalization itself? That would indeed be a mistake. Generalizations of various kinds are ubiquitous, and some generalizations are deeper and more important than others. Even where a science seems overtly focused on mechanisms, there is an obvious role for general statements about the systems being studied; we can often express knowledge of mechanisms in the form of generalizations. (Enzymes are made of protein. Human mitochondria are inherited maternally.) If we admit the importance of generalizations, and make distinctions among them with respect to something like "depth", is the resulting view really so different from the traditional view? It is sociologically interesting that biologists usually do not call even their deeper generalizations "laws", but might this fact be philosophically a superficial one?[8]

There is certainly space for other positions here. Sandra Mitchell (2000) has argued that plenty of generalizations in biology can reasonably be called "laws", provided that we extensively modify the usual philosophical picture of laws. She suggests that we recognize a three-dimensional space in which generalizations can be categorized by their *stability*, *strength*, and *abstractness*. The word "law" might reasonably be used in a context-sensitive way for generalizations that score highly on a relevant mix of the three dimensions, and this is applicable to all scientific fields. Mitchell has no objection to the word "law" being used broadly for "generalizations that ground and inform expectations in a variety of contexts" (p. 262). Her objections are to the usual philosophical account of what these generalizations are like.

There is a risk of the discussion becoming terminological here. But even that fact is of some interest. Mitchell, unlike me, is motivated by the fact that *some* biologists *do* want to call their claims "laws". Her examples are mostly far from the reductionist style of work that is my focus here, but I do not deny that some biologists talk this way.[9] Ecologists, in particular, worry

[8] An example of a very important generalization might be a suitably hedged version of the "Central Dogma" of molecular biology. A reasonable (though unconventional) formulation might be as follows: the linear structure of protein molecules is specified in a template process by the linear structure of nucleic acids, and this process does not occur in reverse. Note also that in this discussion of biology, I do not treat important theorems generated purely analytically from idealized mathematical models (like Fisher's fundamental theorem) as "laws".

[9] I do not agree with all her cases. One, for example, is "Mendel's law of segregation". I am always puzzled when this is called a law (except in a purely historical discussion). There are many exceptions, and these do not involve unusual breakdowns in the system. They just involve the appearance of segregation distorter alleles, which can appear easily and whose action falls squarely within the domain of ordinary biological activity (see Burt and Trivers 2006 for an extensive

more about laws more than other biologists do (Turchin 2001, Ginzburg and Colyvan 2004). So let me first emphasize my common ground with Mitchell. For Mitchell, the standard idea of a binary distinction between laws and "accidental" generalizations is mistaken. She also accepts that "laws" in biology (and elsewhere) are dependent on historical contingencies. And I think that Mitchell would probably accept the following striking difference between physics and biology. In physics, laws *matter* to the organization of knowledge. Textbooks explicitly name and discuss laws. In biology, laws rarely appear in textbooks and research articles. If no biologist ever said the word "law" again, it would make almost no difference to day-to-day work. If no physicist was allowed to say "law", the result would be wholesale reorganization of the field. The laws in physics textbooks may eventually receive unobvious and perhaps deflationary analyses by philosophers, but there is no denying their overt role in day-to-day work. The contrast with biology here is sharp. It is not the case in biology, as it is in physics, that a select group of compact, formal generalizations is installed in a central position in the theoretical structure, and used to derive and organize other information.

Having made this contrast between physics and biology, it is interesting to note the special status of some parts of psychology. If no psychologist was allowed to say "law" ever again, most of psychology would be unaffected, but a few specific sub-disciplines would be. As I understand it, psychophysics still takes laws seriously, and learning theory used to take laws seriously but does so less and less as time passes. Here it is important that the laws in question have been inherited from much earlier work. Psychophysics inherited principles known as laws from work done in the late nineteenth century, and has had reason to hang onto them. Learning theory inherited candidate laws from behaviorist work in the early to mid-twentieth century, and is showing rather less attachment to them.

In any case, when I make no attempt to defend a softened and unorthodox conception of "law" in this paper, that is because: (i) the discussion is being guided by a contrast between fields where laws matter and fields where they do not, and (ii) I think that the traditional strong connotations of "law" will seep back in to undermine revised usages like Mitchell's.

This completes my sketch of an alternative package of ideas in the philosophy of science that might be applied to philosophy of mind. The overall picture is something like this. Suppose we imagine a future science of the mind that has an organization similar to that of the reductionist parts of present-day biology. What would it look like? We would have little overt role for things called "laws". Our knowledge would be organized largely in the form of descriptions of

review). Note also that the term "law" for this and the other two main Mendelian principles was introduced by a *critic* of Mendelism, W. R. F. Weldon (1902). Counterexamples have more bite against attempts to lay down laws. However, though Mendel did not christen the three "laws" attributed to him, he did describe other principles (in particular, the 3:1 ratios in the offspring of hybrids) as laws in his 1865 paper.

mechanisms—how they are structured and how they work. High-level capacities would be explained in terms of the capacities of lower-level parts. "Levels" would be understood in terms of part–whole relations. In early stages of mechanistic investigation, in contexts where high degrees of generality are sought, and in the study of dynamics, we would see an important role for model-building, the investigation of idealized imaginary structures with complicated resemblance relations to real-world systems.

3. QUESTIONS ABOUT MODERN FUNCTIONALISM

What effect would accepting the package of ideas outlined in the previous section have on the philosophy of mind?

This is a difficult question. Late at night in the bar at the Philosophy of Science Association meetings, one might hear grumbling: "People in metaphysics and philosophy of mind have such an antiquated view of philosophy of science!" But the people in metaphysics and philosophy of mind are well within their rights to march into the bar and reply: "What *difference* does it make, to the truly foundational issues? If I fussily re-express everything in the language of the philosophy of science *du jour*, will the issues be much altered, or will they reappear more or less as before?"

In that spirit, my aim in this section is to use the preceding discussion to reexamine some issues in the philosophy of mind surrounding mainstream functionalism. I argue that there are hidden tensions within the usual picture of functionalism and functional description.[10]

My target is a position I will call "modern functionalism". Typical definitions of the view look like this: "Functionalism says that mental states are constituted by their causal relations to one another and to sensory inputs and behavioral outputs."[11] Such a view depends on the more general idea of the functional *profile*, or a total set of functional properties, of a system. A description of a system's functional profile is achieved through a certain kind of abstraction. My focus will be on the nature of these functional profiles. The argument will proceed by comparing what I see as mainstream functionalism with two slightly different views. One is "machine functionalism", an early position that has now been abandoned.[12] The other is David

[10] There is a link between the worries expressed here and some of those discussed by Ned Block (1990).

[11] This formulation from a summary given in the unpublished paper "Functionalism" on Ned Block's website, http://www.nyu.edu/gsas/dept/philo/faculty/block/papers/functionalism.pdf. Other advocates of what I call here "modern functionalism" include Fodor (1981), Stich (1983), Braddon-Mitchell and Jackson (1996), and Crane (1995).

[12] The term machine functionalism is a more recent one, coined (so far as I know) after the demise of the view.

Lewis's view about causal roles and the identification of mental states (1972, 1994).[13]

Each of these views gives a special role to a particular form of functional description. In the case of machine functionalism, this is *machine table* description. In the case of the other two views, it is something a bit different. So here is an obvious-looking question. Suppose we have a candidate functional description of a complex system. Perhaps it is a car engine, or a human agent. Which facts, described in other terms, is the functional description *answerable* to? How might it be disconfirmed? The question is asked purely in principle; we ignore all epistemic problems.

I discuss Lewis's view first. The key idea here is a distinction between *roles* and *occupants*. For Lewis, we often begin by describing a system in terms of an interlocking set of causal roles, and then we look for physical states (or maybe non-physical ones) that occupy those roles. As I understand Lewis, this process is guided by the principle that for each *bona fide* role, there should be at least an approximate occupant. And crucially, occupants have to be ordinary parts of the system, or states instantiated by ordinary parts of the system. We can employ a liberal concept here, but not a trivial one. If we find there is no *bona fide* occupant for some role that we have become accustomed to positing, then we should stop describing the system in terms of that role.[14]

So within the Lewisian style of functional description, if we have a candidate set of functional roles that might be used to describe some system, there is a straightforward way (in principle) to see if the description is OK. We "pop the hood" on the system. (For those unfamiliar with American slang, this means to lift the bonnet of a car, in order to look at the engine.) We look at its physical composition and see whether the roles we have been talking about have occupants or not. So Lewisian functional description is constrained by facts about the physical layout and organization of the system, facts we could discover by popping the hood.

I now turn to machine functionalism. In some ways this view is at the opposite end of a spectrum from the Lewisian view. Machine functionalism makes use of a special kind of analysis, in which a system is described in terms of its inputs, outputs, and a very abstract notion of inner state, or "machine state". A hypothesized functional profile for a system can be expressed in a machine table, which describes transitions between these three kinds of thing. Table 3.1 gives a

[13] Lewis's view is sometimes seen as akin to functionalism, but strictly speaking a form of the identity theory. For discussion of the subtelties here, see Braddon-Mitchell and Jackson (1996).

[14] At the Aarhus conference, Philip Pettit suggested that I am misdescribing the aims and emphasis of Lewis's work here. The aim, Pettit said, is to find ways to fit our common-sense concepts around a scientific picture of the world. The aim is not to outline a research program or a way of further developing our scientific picture. I am unsure whether this contrast captures Lewis's work well or not. If it does, then it would be more accurate to say that the "Lewisian" form of analysis discussed in this section is one that adopts Lewis-style role and occupant description, and puts it to slightly different work from that envisaged by Lewis himself.

Table 3.1. Machine table for a coke machine

Input	Current State	Next state	Output
5	1	2	
5	2	3	
5	3	1	Coke
10	1	3	
10	2	1	Coke
10	3	1	Coke + 5c

standard type of example, a simple coke machine that accepts only 5c and 10c coins, and charges 15c for a coke.[15]

We now ask the question that was asked about Lewis's view. Which facts is the machine table answerable to? In particular, when might we need to pop the hood?

This is a question about how exactly we are supposed to read machine tables, and not everyone reads them the same way. Sometimes it is said that a machine table answers only to the system's input–output profile. Two systems with same total input–output profile must have the same machine table. Then machine functionalism becomes hard to distinguish from logical behaviorism. Machine tables become a means for compact behavioral description. Indeed, without something like a machine table, describing a set of behavioral dispositions that has significant temporal structure (so that some actions occur after a specific sequence of inputs) becomes difficult.

In other discussions, however, machine table analyses are seen as making weak commitments to hypotheses about internal workings. They say something about *how* a behavioral profile is generated. This is certainly how machine tables look *prima facie*; they look as if they introduce "hidden variable" hypotheses of some kind.

What is crucial to this question is the identity conditions for machine states themselves. This is illustrated by a feature of the coke machine in Table 3.1. According to this machine table, there are two different routes by which the system can get to State 3. The coke machine can get to State 3 via receiving a 10c coin, or by receiving two 5c coins. Is there supposed to be an independent sense in which State 3 is the *same* state when reached via these two routes?[16] Could the machine table be disconfirmed if we look inside and see that there is no common physical state that these two causal paths converge on? If a machine table that is behaviorally adequate cannot ever be disconfirmed by popping the hood, then the machine table is a compact description of behavioral facts. If it

[15] Turing machines are sometimes used, instead of simple finite state automata like the coke machine, to illustrate machine functionalism, but for my purposes the coke-machine cases are much better illustrations of the key features of the view.

[16] An analogous question could be asked about the entities quantified over in Ramsey sentence formulations of functionalism.

is, we find no occupants for our roles in an inventory of the system given in independent terms. Do we then discard our initial functional description, or decide to regard it merely as a predictive device? No, we are told by the modern functionalist. The way in which we peered in when we popped the hood was too crude! The entities posited in the higher-level description are abstract, functionally defined entities. They *need not be visible* from the point of view of lower-level description. (Do not look for a "belief box". Do not look for a language of thought as if it involved inscriptions on a little blackboard.)

This seems to mean that the functionally characterized components are not just higher-level, but level-*bound*. They need not be visible at all from other points of view. But they are supposed to be real causal players in the system. We are supposed to be able to give true explanations of the system's behavior in terms of their activities and interactions.

In a discussion of this issue, Mark Johnston suggested that only careless formulations of modern functionalism give rise to these peculiar apparent consequences. If the modern functionalism was telling us to believe in higher-level *particulars* that are invisible from any other point of view, that would be odd. But modern functionalism is properly formulated as doctrine about *properties* and (hence) *states*. We should not use modern functionalism to try to treat beliefs and pains (for example) as level-bound particulars that somehow compose a thinking agent. Instead, the view gives us an account of what it is for a whole agent to have the property of believing that it is raining (or the property of being in pain). And if states are the instantiations of properties at times, then beliefs and pains are states of the whole system.

This distinction does clarify things, but I do not think it greatly ameliorates the situation for modern functionalism. A first indication that things are still awry comes from reflecting on what becomes of causal explanation within such a view.

According to this version of modern functionalism, we treat the system as a whole as having a total set of physical properties at time t_1, that give rise (non-causally) to a range of distinct higher-level properties at that time. The system may then go into a new total physical state at t_2, which gives rise to a range of new higher-level properties. It is not supposed to be the case that the various higher-level properties at t_1 are each instantiated by different physical components of the system. What then seems questionable is the idea that the higher-level states present at t_1 causally interact *with each other* such that there is a legitimate causal description of the system at the higher level, according to which its higher-level states at t_2 are consequences of interactions among its higher-level states at t_1. In the most familiar ways of thinking about causes that interact to produce an effect, the various causes are treated as distinct from each other. Here, by explicitly treating the whole system as the only relevant particular, instantiating all the various mental properties, we have "entangled" the physical bases of each of the mental states whose interactions we might have wanted to describe in causal terms. The problem can be put by saying that there seems to be no difference between this version of modern functionalism, and a version of machine functionalism that expresses its machine states as long conjunctions without positing interactions between the "components" of the total machine state.

This problem is distinct from the more standard problems about mental causation within a physicalist world view (Kim 1993, Bennett 2003). This is because the problem does not arise if the distinct mental states present at a time involve properties instantiated by different physical parts of the system. The problem only arises from the entanglement of the supposedly distinct causal players with each other at the physical level.

This argument is not intended to be decisive. It depends on difficult questions about causation, and the modern functionalist could in any case adopt a mild revisionism about causal description and explanation. But this argument has a more rigorous relative, developed by David Chalmers (1996) for different purposes.

Chalmers' argument forms part of an account of the "implementation" of computational structures by physical systems. It depends on a distinction between two kinds of computational formalisms, which are called FSA (Finite State Automaton) and CSA (Combinatorial State Automaton) descriptions. The key points in Chalmers' treatment bear generally on functionalism, however, and do not depend on linking functionalism to computationalism about the mind.[18]

In formal terms, Chalmers shows that an obvious and straightforward way of understanding what is required for a physical system to implement a CSA is far too weak. This criterion on CSA implementation turns out to require little more of a physical system than that it matches the input–output profile of the CSA. This is important because the CSA formalism is, essentially, the kind of functional specification envisaged in modern functionalism. An extra constraint on the implementation of a CSA is needed to avoid this collapse into near-triviality, and the obvious way to add such a constraint involves a move back towards (what I am calling) a Lewis-style view.

I will sketch some details briefly (though this paragraph and the next can be skipped). A pair of arguments is given. One concerns the implementation of an FSA, which is basically the sort of structure represented by a machine table. In particular, inner states of the system are treated in an atomic way, without internal structure. Surprisingly, any physical system that has the right input–output profile, has some way of recording its input history, and has a "dial" that can be set to various persisting states, implements an FSA, on a natural understanding of implementation.[19] Chalmers accepts this consequence.

[18] This is discussed in more detail in Godfrey-Smith (forthcoming).

[19] Here I only treat the case where FSAs have inputs and outputs in their specification. There are also "inputless FSAs" which are even easier to implement.

A CSA, however, is richer than an FSA. Each overall machine state is broken down into a vector (or list) of substates, and the CSA transition rule takes the system from one vector of substates (plus an input), to a new vector of substates (plus an ouput). So treating a system as a CSA seems to involve positing a number of interacting internal states present at any given time, each with its own role in the system. But suppose we say that any physical system implements a CSA if there is a mapping between the states of the physical system and those of the CSA such that causal processes in the physical system correspond to all the possible transitions in the CSA's formal specification. This simple criterion for implementation can be shown to be too weak. Any CSA can then be mapped to an FSA with a suitably large number of atomic inner states, in such a way that it inherits the weak implementation requirements of that FSA. So any system with the right input–output profile (plus an input memory and "dial") will implement the CSA. The appearance of further constraints on implementation deriving from the interactions among the substates of the CSA is illusory.

If implementing a CSA is to require more than this, some extra requirement is needed. In his discussion, Chalmers considers a simple and clearly sufficient candidate, and some weaker options that may or may not suffice. In my terms, the simple option is one that involves a move back towards the Lewisian view discussed above. This is the requirement that each CSA substate be mapped onto a *distinct spatial region* of the implementing system. Chalmers discusses the possibility that a weaker condition than this will suffice, but an extra requirement of something *like* this kind is needed. In particular, a theory of implementation must exclude a mapping in which each CSA substate is mapped holistically to a partial specification of the physical state of the entire system.

So to know whether a CSA is non-trivially implemented by some physical system, we have to work out whether the CSA substates can be mapped to something like distinct parts of the physical system. We have to pop the hood, and the aim when we do so is to see whether the roles in the CSA specification have occupants that are *bona fide* parts, or states of *bona fide* parts.

Two conclusions can be drawn. One is that the overt form of description standardly seen in modern functionalism, on its own, exerts far less constraint on the physical system being described than one might think. The other is that the obvious way (probably not the only way) to restore the lost content to functional description is to move back towards the requirement that occupants of roles have independent standing as real parts of the system.

Another moral I take from Chalmers' argument is that modern functionalism is a less worked-out and coherent doctrine than it looks. Chalmers himself does not draw this conclusion, perhaps because he sees the extra constraint that is needed on CSA implementation as being more in the spirit of standard functionalism than I do. In any case, in the remainder of this section I will put a different option on the table. This option may be a better way of making sense of the phenomena that functionalists want to capture, and a better way of describing the scientific work that is taken to support a functionalist attitude.

This alternative view distinguishes two kinds of thing that can look like "functional" description in the philosophers' sense, and that can shade into each other in some cases. Both were introduced in the previous section; they are mechanistic description, in roughly the sense of the new mechanists, and modeling. These are two real kinds of scientific work, a bit different from each other, with particular relations between them.

Scientific analysis in the style of the new mechanists is quite close to Lewisian functional description. The mechanists and Lewis use different terminologies and have different agendas, of course. Their treatments of causation are also very different. But in other ways, the two pictures are quite similar. The aim in both cases is to describe how the abstract causal analysis of a complex system works. The kind of description that results is answerable to what you see when you pop the hood. Both use a simple notion of levels of analysis, based on ordinary part–whole relations. There are no mysterious level-bound objects. In the previous section I said that the new mechanists had given a fairly good account of the explanatory style of fields like cell biology. In a considerably more qualified way, the same could be said for Lewis's framework.

As discussed in the previous section, though, when faced with complex systems that are poorly understood it can be wise to temporarily eschew the aim of direct mechanistic description. We may not have the right kind of inventory of parts; we may not know what kinds of structures to be looking for as the bearers of key causal roles. In that situation, we model. We describe possible networks of dependence relations, idealized possible machineries. We hope for similarity relations between these hypothetical structures and the real workings of the system. Modeling in this sense is different from the analysis envisaged in modern functionalism in at least two ways. First, this sort of modeling does not traffic in level-bound objects, and secondly, a crucial role is played in modeling by the complex nature of the similarity relations that may hold between model and target.

So we might consider replacing the special form of functional analysis seen in much recent philosophy of mind with two slightly different tools: mechanistic description (which is fairly close to Lewis), and modeling. This combination provides a better framework for thinking about psychological phenomena than modern functionalism does. (Indeed, it is what psychology and cognitive science have mostly been employing all along.) The important functionalist notion of multiple realizability survives intact in this view, because a given role can have physically different occupants in different cases. From this point of view, however, modern functionalism seems to be an attempt to devise a hybrid form of analysis that has some characteristics of each of two legitimate kinds of description. Sometimes it looks like abstract description of real mechanisms, and sometimes it looks like modeling, but it is supposed to be a single thing distinct from each of these. I suggest that this might be an illusion.

Here is one other way to look at the situation. Earlier I said that modern functionalism is designed to enable us to say two things at once. First, people want to treat the various components of a total psychological profile as picking out distinct things that can interact causally. Second, they do not want the useability and legitimacy of folk psychological concepts (like belief and pain) to depend on there being localized physical occupants of these roles in the brain. The suggestion I am making here enables people to say both these things, but not about the same states at the same time. Folk psychology might be something like a model, rather than a theory, of the mind.[20] As a model, it can be useable without there being a simple mapping between its structure and the machinery of the brain. But when the aim is to come up with a literally correct causal description of how mental processes work, using either folk psychological concepts or scientific ones, then we should expect and aspire to engage in the description of mechanisms.

4. LAWS, CONFIRMATION, AND KINDS

The previous section sought to export a package of ideas from philosophy of science into philosophy of mind. In this final section I return to philosophy of science. I will briefly confront a possible objection that might make people reluctant to embrace the package of ideas presented earlier. Here we leave the general topic of reduction, though, which is why this section is at the end.

The objection runs as follows. The familiar body of ideas in philosophy of science that was discarded in Section 2 is essential to the treatment of various other issues. There is a larger network of views whose viability is being questioned here.

The network of ideas I have in mind here posits a set of connections between laws, kinds, counterfactuals, and confirmation. Here I will focus on confirmation. Especially since the work of Goodman, it has been common to hold that the concept of law and the concept of confirmation are closely linked. Only law-like generalizations are confirmed by their instances; "accidental" generalizations are not. If our analysis of some part of science does not take seriously the notion of law, then, it may seem that we will not be able to understand how the confirmation of hypotheses works in that part of science. And for many philosophers, the link between law and confirmation is just one element in a rich network of ideas which it would be very costly to abandon.

My response is that the familiar network of ideas about laws, kinds, and confirmation is much overrated. We would probably be better off without it. I

[20] This idea is developed in more detail in Maibom (2003) and Godfrey-Smith (2005).

will not give a general defence of this claim in this section, but will indicate what one part of a better package of views might look like.[21]

The alleged link between laws and confirmation arises in the attempt to make sense of "instance confirmation", the support that some generalizations receive from observations of particular cases that satisfy them. Goodman's "grue" problem teaches us that not all generalizations receive support from observations of their instances (1955). Perhaps, however, instance confirmation is real when the generalization in question is law-like? Goodman linked both law-likeness and confirmation to a conception of "projectibility" based on the historical role of a predicate or category in a linguistic community, but other philosophers have generally rejected that idea while hanging onto the link between laws and confirmation.

The whole idea of "instance confirmation" is in much worse shape than even Goodman supposed. It is the creature of a particular kind of philosophical system-building, and not a genuine scientific phenomenon that needs philosophical explanation. The philosophical concept of instance confirmation is, I suggest, an unholy amalgam of two genuine inference patterns in science. One is statistical inference from samples. The other is what is usually called "inference to the best explanation" (IBE).[22] These are both real and legitimate, and each has *some* of the features that philosophers associate with confirmation by instances.

In statistical inference from samples, the *size* of sample is usually very important. Many observations are better than a few. *Randomness* of sampling is usually very important. But there is no "naturalness" constraint, of the type familiar from philosophical discussions of Goodman's problem. Roughly speaking, any predicate can be used in statistical inference from a random sample. There are problems of sample bias and confounding that have connections to Goodman's problem (Godfrey-Smith 2003*a*). But the overall status of kinds—their naturalness or lack of it—is not an issue.

In inference to the best explanation, there is no essential role for number of observations, for size of sample. Size may have some practical importance, but it is not evidentially central as it is in statistics. What is important in IBE is the specific causal and nomological structure that is relevant to the case. This is related to the "naturalness" of kinds, though it is not the same thing.

What we see in much post-Goodman thinking about confirmation, however, is a mixture of the features of these two kinds of inference. It is common to think that both the number of observations and the naturalness of kinds are important, while randomness of sampling is rarely discussed. This category is a philosophical fiction. And the idea that positive instances confirm law-like

[21] A more detailed discussion is found in Godfrey-Smith (2003*a*), especially the final section.

[22] In Godfrey-Smith (2003*b*) I preferred the modified term "explanatory inference" because I think IBE suggests the wrong kind of link to an independent notion of goodness of explanation (in the sense discussed in the Hempel, Salmon, Van Fraassen (etc.) literature on explanation). Here I will use the more common term.

generalizations, and only them, is not a feature of either statistical inference or IBE.[23]

This last section has traveled some distance from the topic of reduction. But these points do play a role in the earlier discussion. It seemed for some time that philosophy of science had generated a tightly-knit and plausible package of ideas about laws, confirmation, and kinds. When someone argues, as I did earlier, that there is no important role for laws in some part of science, the appeal of the larger package of ideas linking laws and confirmation (etc.) is one motivation for attempts to find a *hidden* role for laws, lurking in work that is ostensibly quite different in organization. But at least in the scientific fields that border on philosophy of mind, the lawless nature of reduction in real life is something we can, and should, take at face value.

REFERENCES

Bechtel, W. and A. Abrahamsen (2005). "Explanation: A Mechanistic Alternative." *Studies in History of Philosophy of the Biological and Biomedical Sciences* 36: 421–41.

—— and R. Richardson (1993). *Discovering Complexity.* Princeton: Princeton University Press.

Bennett, K. (2003). "Why the Exclusion Problem Seems so Intractable, and How, Just Maybe, to Tract it." *Noûs* 37: 471–97.

Block, N. (1990). "Can the Mind Change the World?" In G. Boolos (ed.), *Meaning and Method: Essays in Honor of Hilary Putnam*: Cambridge: Cambridge University Press, 137–70.

—— and J. A. Fodor (1972). "What Psychological States are Not." *Philosophical Review* 83: 159–81.

Braddon-Mitchell, D. and F. Jackson (1996). *The Philosophy of Mind and Cognition*. Oxford: Blackwell.

Burt, A. and R. Trivers (2006). *Genes in Conflict: The Biology of Selfish Genetic Elements*. Cambridge MA: Harvard University Press.

Chalmers, D. J. (1996). "Does a Rock Implement Every Finite-State Automaton?" *Synthese* 108: 309–33.

Collins, J., N. Hall, and L. Paul (eds.) (2004). *Causation and Counterfactuals*. Cambridge MA: MIT Press.

Crane, T. (1995). *The Mechanical Mind: A Philosophical Introduction to Minds, Machines and Mental Representation*. London: Routledge.

Cummins, R. (1975). "Functional Analysis." *Journal of Philosophy* 72: 741–65.

—— (2000). "'How Does it Work?' vs. 'What are the Laws?' Two Conceptions of Psychological Explanation." In F. Keil and R. Wilson (eds.), *Explanation and Cognition*. Cambridge MA: MIT Press, 117–45.

[23] The argument in this paragraph is structurally similar to an argument from the preceding section. In each case (confirmation, functional analysis) the recent philosophical tradition has drawn on two real phenomena and combined their elements in the wrong way.

Dupré, J. (1993). *The Disorder of Things.* Cambridge MA: Harvard University Pres.
Fodor, J. A. (1974). "Special Sciences." *Synthese* 28: 97–115.
—— (1981). *Representations*. Cambridge MA: MIT Press.
Giere, R. (1988). *Explaining Science: A Cognitive Approach*. Chicago: Chicago University Press.
—— (1999). "Using Models to Represent Reality." In L. Magnani, N. J. Nersessian, and P. Thagard (eds.), *Model-Based Reasoning in Scientific Discovery*. New York: Kluwer/Plenum, 1999, 41–57.
Ginzburg, L. and M. Colyvan (2004). *Ecological Orbits: How Planets Move and Populations Grow*. Oxford: Oxford University Press.
Glennan, S. (2005). "Modeling Mechanisms." *Studies in the History and Philosophy of the Biomedical Sciences* 36: 443–64.
Godfrey-Smith, P. (1999). "Procrustes Probably: Comments on Sober's Physicalism from a Probabilistic Point of View." *Philosophical Studies* 95: 175–81.
—— (2003*a*). "Goodman's Problem and Scientific Methodology." *Journal of Philosophy* 100: 573–90.
—— (2003*b*). *Theory and Reality: An Introduction to the Philosophy of Science*. Chicago: Chicago University Press.
—— (2005). "Folk Psychology as Model." *Philosopher's Imprint* 5/6: 1–16.
—— (2006). The Strategy of Model-Based Science. *Biology and Philosophy* 21: 725–40.
—— (forthcoming). "Triviality Arguments Against Functionalism." *Philosophical Studies.*
Goodman, N. (1955). *Fact, Fiction, and Forecast*. Cambridge MA: Harvard University Press.
Harman, G. (1965). "The Inference to the Best Explanation." *Philosophical Review* 74: 88–95.
Horst, S. (unpublished). "Beyond Reduction: What Can Philosophy of Mind Learn from Post-Reductionist Philosophy of Science?" Presented at Boston Colloquium in the Philosophy of Science, 2005.
Kim, J. (1993). *Supervenience and Mind: Selected Philosophical Essays*, Cambridge: Cambridge University Press.
Levins, R. (1966). "The Strategy of Model-Building in Population Biology." *American Scientist* 54: 421–31.
Lewis, D. (1972). "Psychophysical and Theoretical Identifications." *Australasian Journal of Philosophy* 50: 249–58.
—— (1973). "Causation." *Journal of Philosophy* 70: 556–67.
—— (1994). "Reduction of Mind." In S. Guttenplan (ed.), *A Companion to the Philosophy of Mind*, Oxford: Blackwell, 413–31.
Machamer, P., C. Craver, and L. Darden (2000). "Thinking About Mechanisms." *Philosophy of Science* 67: 1–25.
McLaughlin, B. and K. Bennett (2005). "Supervenience." *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/supervenience/
Maibom, H. (2003). "The Mindreader and the Scientist." *Mind and Language* 18: 296–315.
Mitchell, Sandra D. (2000). "Dimensions of Scientific Law." *Philosophy of Science* 67: 242–65.
Sober, E. (1999). "Physicalism from a Probabilistic Point of View." Philosophical Studies 95: 135–74.

Stemwedel, J. (2006). "Getting More with Less: Experimental Constraints and Stringent Tests of Model Mechanisms of Chemical Oscillators." *Philosophy of Science* 73: 743–54.

Stich, S. (1983). *From Folk Psychology to Cognitive Science: The Case Against Belief.* Cambridge MA: MIT Press.

Suppe, F. (ed.) (1977). *The Structure of Scientific Theories*. 2nd edition. Urbana: University of Illinois Press.

Turchin, P. (2001). "Does Population Ecology have General Laws?" *Oikos* 94: 17–26.

Van Fraassen, Bas C. 1980. *The Scientific Image.* Oxford: Oxford University Press.

Weisberg, M. (2006). "Who is a Modeler?" *British Journal for the Philosophy of Science* 58: 207–33.

Weldon, W. (1902). "Mendel's Laws of Alternative Inheritance in Peas." *Biometrika* 1: 228–54.

Wimsatt, W. (1972). "Teleology and the Logical Structure of Function Statements." *Studies in the History and Philosophy of Science*, 3: 1–80.