2

New Mechanistic Explanation and the Need for Explanatory Constraints

L.R. Franklin-Hall

Introduction

In the past decade and a half, a new "movement" (Glennan 2005: 443) has arisen in the philosophy of biology, one called a "revolution" (Bechtel 2006: 280) with "broad implications" (ibid: 2) and which has met with "broad consensus" (Campaner 2006: 15). On this "hot topic" (Robert 2004: 159), a vast literature has developed, within it one of the most cited papers in *Philosophy of Science* (viz. Machamer et al. 2000).

What is the subject of such attention? It is the "new mechanistic philosophy" (Skipper and Millstein 2005: 327), articulated by a group of philosophers—including William Bechtel, Carl Craver, Lindley Darden, Peter Machamer, and Stuart Glennan—interested in the nature of mechanisms, complex systems characterized most prominently as "entities or

L.R. Franklin-Hall (✉)
Department of Philosophy, New York University, New York, NY, USA

K. Aizawa, C. Gillett (eds.), *Scientific Composition and Metaphysical Ground*, DOI 10.1057/978-1-137-56216-6_2

activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions" (Machamer et al. 2000: 3). Mechanisms are said to be worthy of attention largely because they are central to a new and superior approach to scientific explanation, one truer to scientific practice than the long defunct deductive-nomological (DN) view. It is also claimed that the mechanistic approach has implications beyond explanation, as it "transforms how one thinks about a host of other issues in the philosophy of science" (Bechtel and Abrahamsen 2005: 426), including causality, laws, kinds, reduction, discovery, and scientific change.

Philosophical movements can be judged by their fruits. We can ask of them: what problems does a movement offer solutions to? Judging by both the language of the new mechanists and the influence of their work, it would appear that the mechanistic approach had served up a bevy of solutions. Yet I argue here that, at least with respect to its core project—that of elucidating the nature of scientific explanation—appearances are deceptive: the mechanisms movement has not yet yielded the advertised results. This is not because mechanisms advocates are committed to claims that are false. My critique is motivated instead by concerns that mechanistic explanatory accounts offered to date—even in their strongest formulations—have failed to move beyond the simple and uncontroversial slogan: "some explanations show how things work." In particular, I argue that proposed constraints on mechanistic explanation are not up to the task required of them: namely, that of distinguishing acceptable explanations from those that, though minimally mechanistic, are uncontrovertibly inadequate.

Sections "The Mechanistic Explanatory Framework" and "Formulating Explanatory Constraints" sketch a version of the new mechanistic explanatory account, one constructed by combining the most promising proposals from across the mechanistic corpus. After articulating three principles at the heart of this picture—concerning causation, parts, and explanatory level—sections "The Causal Standard" through "The Levels Standard" argue that these principles remain promissory notes. The chapter concludes in section "Conclusion" with an evaluation of the mechanistic explanatory program.

The Mechanistic Explanatory Framework

Scientists in many disciplines—but particularly in biology—frequently refer to *mechanisms* when describing the systems they investigate, provoking a natural question: "what is a mechanism?" Answers to this question from advocates of the new mechanistic philosophy—such as from Machamer et al. (2000: 3), Bechtel and Abrahamsen (2005: 423), and Glennan (2002: S344)—differ more in language than content; all agree that a mechanism is a physical system composed of at least somewhat organized parts whose interactions either bring about or constitute some phenomenon.

Though mechanisms may be germane to various philosophical endeavors (Levy 2013; Nicholson 2012), most prominent is their central place in a theory of explanation, one intended to apply to many of the biological sciences. According to that theory, explanations are explanatory in virtue of communicating facts about "how things work" (Craver 2007b: 110) in the system that brings about, or constitutes, the phenomenon to be explained. These facts should be communicated by a largely veridical representation—called a *mechanistic model*—of the system responsible for the explanandum phenomenon (Bechtel and Abrahamsen 2005: 425; Craver 2006; 2007b: vii; Glennan 2005: 446; Machamer et al. 2000: 3).

Mechanistic models need not take some canonical form, nor must they be usable to derive a statement of the explanatory target (Bechtel and Abrahamsen 2005: 430; Bechtel 2011: 537; Craver 2007b: 160; Machamer et al. 2000: 23). What they must do is account for a system's capacity to produce certain outputs in response to certain inputs. To do this in a properly mechanistic style, they should describe the system as having multiple parts that are organized in some respect and that change through time according to dynamic principles, principles that might be understood to reflect activities, laws, or some other species of regularity. When such models bridge inputs and outputs as required, they can directly explain systems-level capacities; they may also explain particular events when supplemented by a statement of initial (i.e., activation) conditions.[1]

[1] Beyond token capacities and events, mechanists also aspire to treat regularities. Though the details are rarely made explicit, a given regularity can be explained via a mechanistic model jointly applicable to all of the particular systems underpinning a regularity's instances; to do this, such a model must be at least somewhat abstract. For a discussion of how this might work, see Strevens' "First Fundamental Theorem of Explanation" (2008: Chap. 7).

The most important variety of model that new mechanists judge as *unexplanatory*—at least when deployed to explain the biological phenomena that interest them most—is the *global model*, one constituted by a single dynamic principle stating that a system experiencing such-and-such inputs will produce such-and-such outputs.[2] Global models treat systems as opaque black boxes; they fall short explanatorily in virtue of failing to look "under the hood" and "beneath the regularities couched at the behavioral level to reveal underlying mechanisms" (Kaplan and Bechtel 2011: 442).

Formulating Explanatory Constraints

The explanatory framework sketched above is plausible yet incomplete. The basic problem is that, for any candidate explanandum phenomenon that the new mechanistic account aims to treat, there exist an enormous range of models that satisfy the above-noted core mechanistic conditions, that is, by representing the system in terms of organized parts that change according to dynamic principles. Yet, only a handful of these models appear to be explanatorily apt. Thus, to fill out the account, we must design constraints capable of distinguishing the good mechanistic models—those that provide adequate explanations—from those that fall explanatorily short.

To illustrate this challenge, and to motivate the new mechanistic contributions that might be used to meet it, I will describe four veridical, mechanistic models for a single phenomenon: a neuron's capacity to release neurotransmitters at its axon terminal when its dendrites are exposed to neurotransmitters, and not otherwise.[3] While the first model, called here the *Standard Model*, is explanatorily acceptable—based as it is on textbook

[2] Mechanists also judge unexplanatory *phenomenological models*—those that don't purport to describe the inner workings of the system at issue—as well as mechanistic models that are false (even allowing for limited idealization) of the systems they purport to describe. As these exclusions will be uncontroversial for any fan of causal explanation, they require no discussion.

[3] Though this phenomenon is often modeled probabilistically, I treat it deterministically for the sake of expository simplicity. This simplification is innocent; over-permissiveness would be found equally on any probabilistic formulation.

accounts—the other three will appear flawed. They each make a distinct variety of explanatory error and can illustrate in the breach the constraints that a mechanistic model must fulfill to be explanatorily acceptable.[4]

According to the Standard Model, a neuron's capacity to release neurotransmitters when exposed to them is explained by describing the neuron as composed of a variety of somewhat organized macromolecular parts, including membranes, channels within them, and ionic concentrations in the internal and external environment—all of which interact according to dynamic principles, such as one stating that neurotransmitter binding is followed by channel opening.[5] Though these details could be communicated in a variety of ways, they are most often presented in narrative form, as follows: neurotransmitter exposure leads neurotransmitter molecules to bind to ligand-gated receptors located in the dendrite membrane. Upon binding, these channels open. Then sodium ions rush into the cell, depolarizing the membrane locally. Next, a population of voltage-gated membrane channels, located in the same region, also open and more sodium enters. This begins a cascade of channel opening, depolarization, and further channel opening, that moves up the neuron until it reaches the neuron's axon terminal where voltage-gated calcium channels open and calcium enters the cell. Finally, vesicles containing neurotransmitters located near the axon terminal bind with the membrane, releasing neurotransmitters to the extracellular environment.

To formulate a second kind of model that applies to the same explanandum, consider any regular "side-effect" of neurotransmitter binding, such as the mild vibration of the cell membrane surrounding the receptor. Presume that whenever the neuron is exposed to neurotransmitters, this

[4] All four candidate models maintain that a neuron behaves thus because it is constituted in such a way that (1) it does not release neurotransmitters absent neurotransmitter exposure, and (2) exposure initiates a cascade of events, one of which is neurotransmitter release. Yet, the first condition is customarily taken for granted, and explanatory presentations focus on the second by describing the relevant features of the constitution of the neuron, and how exposure—given this constitution—has the specified result.

[5] Just as the overall phenomenon might be treated either probabilistically or deterministically, so it goes with this dynamic principle. Though I will not worry about the details, which sort of treatment is most apt will depend on how the channels are individuated. If *single channels* are separately represented, a probabilistic treatment is most appropriate; if large collections of channels are treated together, deterministic treatment will be preferred.

vibration occurs, but it has no consequences on the remainder of the cell depolarization process. Given this, we can formulate a model identical to the Standard Model, except that it appeals to two alternative dynamic principles, one relating neurotransmitter exposure and membrane vibration, and a second relating vibration and any later event genuinely relevant to neurotransmitter release, for example, the entry of calcium into the axon terminal. With these principles and others, such an alternative model might bridge inputs and outputs, stating first that neurotransmitter exposure is followed by membrane vibration, itself followed by cellular calcium entry, eventuating finally in neurotransmitter release.[6] Like the Standard Model, this model can appeal to organized parts changing according to dynamic principles. Nevertheless, it is flawed in virtue of making what I call a *causation error*.

A third kind of model correctly describes causal connections between a system's parts, but individuates those parts in a non-standard—and explanatorily deficient—way. Consider, for instance, a model that describes just four connected parts of the neuron, large chunks of biomass extending about one-fourth of the way from dendrites to the axon terminal, each capable of taking at least two states. This model might be used to account for the target phenomenon as follows: neurotransmitter exposure changes the state of the first part, which modifies the state of the second part, in turn modifying the third in the same way, and then finally the last hunk of neuronal materials, eventuating in the output of interest—neurotransmitter release. This model, however peculiar, is also properly mechanistic: it describes multiple organized parts, changing according to dynamic principles, and principles that themselves track the causal order. Nevertheless, in virtue of its gerrymandered carving of the system into quarter-neurons, it fails to reflect actual explanatory practice, and is intuitively unexplanatory. It makes what I will call a *carving error*.

[6] Some might suggest that this model isn't mechanistic *at all*, insisting that to be mechanistic a model must satisfy a causal constraint. This would be to cut up the project slightly differently than I have, but with no consequences for the overall argument. The task facing the new mechanist would still be to cash out the causal constraint; it matters not whether that constraint is appealed to in the definition of mechanistic models *simpliciter*, or (as in my exposition) in the characterization of *explanatorily adequate* mechanistic models.

The fourth model characterizes both real causal connections and appeals to "natural," rather than gerrymandered, parts. Its distinctive explanatory shortcoming is that it describes the system at the wrong "level," in terms of organized atomic parts changing according to dynamic principles (in this case, principles aptly called *laws*) describing atomic interactions. Such a model will be so complex that, in contrast with the three rehearsed already, it is not possible here to sketch the course of events it would describe as following from neurotransmitter exposure. Yet such a low-level model will still satisfy the requirements of the mechanistic framework above: it describes organized parts that change over time according to dynamic principles, collectively bridging inputs and outputs. By depicting the neuron in such detail, it makes what I call a *zooming error*, and should, as above, be censured by any explanatory account that takes actual explanatory practice as its touchstone.[7]

The three flawed accounts just sketched were easy to design, and equivalent alternatives are readily available for any explanandum you might choose; they require no real creativity or insight. One starts with the input–output relationship for which the mechanistic model must account. These inputs and outputs, as the mechanists rightly emphasize, will be underpinned, in any particular system, by a complex set of connections between that system's parts. To produce a model that errs *causally*, describe at least some portion of the system underlying the explanandum behavior in terms of correlational—not causal—principles. To produce one that makes a carving error, describe that underlying system veridically, but use a peculiar set of terms, those that individuate the system in a non-traditional way.[8] Finally, to produce a model at the wrong *level*, either zoom in on the parts of the system more than is explanatorily appropriate—by describing, for example, the inner working of entities usually treated as wholes by scientists accounting for the focal phenomenon—or fail to break the system into parts, thereby producing a global model.

[7] A zooming error is a species of carving error, and they are separated largely for rhetorical purposes. The first prototypically concerns using gerrymandered parts, while the second concerns otherwise "natural" parts at too fine (or coarse) a grain, considering the explanandum phenomenon.

[8] Though many peculiar sets will exist, not any will do: they must still be sufficiently expressive that they can be used, in concert with some set of dynamic principles, to bridge inputs and outputs.

At the core of the mechanistic explanatory account, as I reconstruct it, stand three standards that rule out models that suffer from these three types of errors. These should act—either individually or collectively—as a kind of sieve, sifting out the detritus, and revealing the explanatory nuggets.

The Causal Standard The dynamic principles that describe system change should be *causal*. Different workers attempt to spell out this requirement differently, sometimes drawing strategies from theories of causation produced independently of the mechanisms movement. For instance, some mechanists depend on Woodward's (2003) version of the interventionist account of causation (Craver 2007b; Glennan 2002), while others develop their own activities theory (Bogen 2005; Machamer 2004).

The Carving Standard Models should carve mechanisms "at their joints," describing them in terms of the appropriate set of parts (Craver 2006: 367; Bechtel and Abrahamsen 2005). They should not reflect the "arbitrary differentiation of the parts of a whole" (Bechtel 2008: 146). For instance, parts appealed should be good parts, like macromolecules, rather than bad parts, like quarter-neurons.

The Levels Standard Models should represent the system at the right "level," or grain, which in the judgment of many (though not all) new mechanists will not be a fine-grained physical specification but will be in various ways abstract (Levy and Bechtel 2013). In particular, some will hold that an explanatory model should represent systems at the level *just below* that of the explanandum phenomenon. Thus, it may be a mistake to explain neurotransmitter release at the axon terminal in terms of atomic events, or even in terms of an "influx of sodium" into the terminal, rather than in terms of a comparatively coarse-grained event like "depolarization" (Craver 2007: 23).

How successful are these standards? The burden of the next three sections is to argue that they are not yet up to the task assigned to them: that of distinguishing the genuinely explanatorily models from the many that fall short.

The Causal Standard

According to the first standard, the dynamic principles embedded within explanatory models must describe *causal relations*, not mere relations of correlation. As Craver notes, "analyses of explanation must include reference to causal relationships if they are to distinguish good explanations from bad" (2007a: 8).

This basic claim is highly plausible but requires elaboration. After all, though causation is one of the most familiar features of our world, it is also among the most obscure. What is this relation between cause and effect, the basic material out of which a causal explanation is constructed? Are causes related to effects, as Hume thought, just by their "constant conjunction"? Or does causation involve a more metaphysically loaded relation of *dependence* or *necessitation*? In that case, how are we to understand this more substantial connection, for instance, in terms of the truth of certain counterfactuals, or in terms of some relationship between universals?

Before discussing the new mechanistic approach to the causal relation, consider an alternative strategy that connects mechanisms and causation, pursued by an earlier generation whom we might call the "old mechanists." Peter Railton, Wesley Salmon, and J.L. Mackie aimed to use mechanisms to contribute to our understanding of the causal relation, specifically to what distinguished causal connections from mere correlations. Mackie, for example, hoped that what he called a "mechanism" might constitute "the long-searched for link between individual cause and effect" (Mackie 1974: 228–229). And both Salmon (1984) and Railton (1978) attempted to give an account of causation in terms of "mechanism." Many believe that these accounts failed on their own terms (Hitchcock 1995), though it was clear what these philosophers were up to: they were using mechanisms to do battle with "Hume's Ghost," and attempting to "glimpse the secret connexion" between cause and effect.

The relationship between this work and that of the new mechanists has not always been transparent. Machamer et al. (2000) explicitly compared the new mechanists' project to Salmon's and Mackie's but lamented that "it is unclear how to apply [Salmon's and Mackie's] concepts to our biological cases" (2000: 7). Glennan (2002) also suggests that the new mechanists' approach was a successor project, writing that while "philosophers of science typically associate the causal-mechanical view of scientific explanation with the work of Railton and Salmon, [….I] shall argue that the defects of this view arise from an inadequate analysis of the concept of mechanism" (S342).

Yet a clear contrast exists between the old mechanists and the new, and it may be misleading to see their projects as continuous. The key difference concerns the relationship between *cause* and *mechanism*. The old mechanists were trying to reduce causation to mechanism; however, most new mechanists use accounts of causation to understand the relations *between parts* (or, properties of parts) of mechanisms. Speaking metaphorically, old mechanisms were the causal glue, while new mechanisms are glued together by causes. Along these lines, recent commentary calls for abandoning "the idea that causation can be reduced to mechanism. On closer inspection, it appears that the concept of mechanism presupposes that of causation, far from being reducible to it" (Kistler 2009: 599).

Given that mechanisms don't reduce causation but instead require an account of it, what account should that be? Clearly, it must differentiate dynamic principles that reflect relations of correlation from those of causation. To this end, two paths have been taken. The first is to tie the mechanistic approach to an independent account of causation, one that may lack any interestingly mechanistic character, for instance, to Woodward's interventionism or Lewis' counterfactual account. Craver (2007), Glennan (2005), and Leuridan (2010) have pursued this strategy, adopting Woodward's (2003) account of causation, according to which causal relations are those "potentially exploitable for the purposes of manipulation and control" (Woodward 2003: 17). The second is to develop an account of causation with mechanistic contexts in mind. For example, Bogen (2008) and Machamer (2004) have pursued this option, developing an "activities view" of causation.

Though the first approach—that of adopting an independent, non-mechanist account of causation—is perfectly reasonable, I will not explore it. Given the uncontroversial nature of the basic mechanistic conditions—at least for fans of causal explanation—those who fill out the mechanistic picture by adopting a self-standing account of causation are not much advancing the explanatory project. Needless to say, outsourcing causation may well be the right move for mechanists to make, and those who do so may still contribute to our understanding of scientific explanation; however, their contributions must come from elsewhere, presumably from their elucidations of the other two constraints on mechanistic explanations—on parts and level—which will be explored in due course.

Some mechanists have attempted to make sense of the causal relation via the notion of activities (see Bogen 2005, 2008; Machamer 2004; Waskan 2011). Here is an early statement of the view:

> An entity acts as a cause when it engages in a productive activity. [...] A mechanism is the series of activities of entities that bring about the finish or termination conditions in a regular way. These regularities are non-accidental and support counterfactuals to the extent that they describe activities (Machamer et al. 2000: 6–8).

The basic idea is that *X causes Y when related by an activity*. Focusing in this way on activities appears to provide a simple, scientifically informed analysis of causation that avoids many of the thorny matters—such as the nature of laws, regularities, or counterfactuals—that consume those more metaphysically minded. As Bogen puts it, "[i]f the production of an effect by activities which constitute the operation of a mechanism is what makes the difference between a causal and a non-causal sequence of events, mechanists need not include regularities and invariant generalizations in their account" (Bogen 2005: 399).

This activities account, also called the "actualist-mechanist theory" (Waskan 2011), is offered as one of many *process* or *production* theories of causation (Hall 2004). In this case, what makes for a causal connection is an *actual process* of a certain type. Early advocates of the process approach had empiricist sympathies: they were suspicious of the counterfactuals that seemed necessary to make sense of a dependence relation, and wanted to do without them. Their task was to distinguish, in general terms, causal processes from what are sometimes called "pseudo-processes" that may reflect merely correlated events, and all without a counterfactual crutch.

There appear to be two ways of making "activities" part of a philosophically informative theory of causation. Most obviously, activities might be the "special sauce" that distinguishes the genuinely causal processes. The philosophical task would be to describe these activities, characterizing precisely how they are special. The activity approach would, in this case, be structurally similar to the old mechanists' accounts, noted above, which offered not "activities" but "mechanisms," understood in terms of the capacity "of transmitting a local modification in structure (a 'mark')" (Salmon 1984: 147) or "the exchange [or persistence] of a conserved quantities" (Dowe 1995: 323) as tools with which to separate the causal wheat from the correlational chaff.

Second, the activities approach might, though refraining from the above task, identify what the activities in fact are. This could be likened to Descartes' attempt to characterize the causally efficacious properties—such as extension and velocity—as part of a quest to banish the "substantial forms" and "final causes" which Descartes' contemporaries appealed to, in his view, willy-nilly. Jon Elster's work on functions in the social sciences also has this character. He emphasizes the importance of uncovering the "nuts and bolts" of social mechanisms because he believes that—absent a selection process—the functional properties that are appealed to in social–scientific explanations are actually explanatorily empty (Elster 1989). This sort of project would be particularly well motivated if the new mechanists suspected that biologists were likewise appealing to non-explanatory, non-causal features.

Yet those developing an activities account of causation have refrained from both of these tasks. Advocates dodge the first project by claiming that "activities" have merely verbal unity. Scientists do somehow distinguish "causally productive" activities from those that are not, but the distinction cannot be "captured informatively by any single account" (Bogen 2008: 116). This is because "there is no informative general characterization which discriminates causally productive activities from goings-on which are not causally productive of the effect of interest" (ibid: 113; Machamer 2004). Mechanists also refrain from the second undertaking. Unlike Descartes and Elster, they evince no general skepticism regarding the activities appealed to by the competent scientists whose work they study, noting instead that "acceptable causal relations are those that our scientific investigations reveal to us as how the world works" (Machamer 2009: 4). And they claim, wisely enough, that there is no definitive list that philosophers might produce of the activities, and that it is the job of scientists, in any case, to compile it.

The central feature of this account of causation—the activity—is, from a philosophical perspective, brute. Scientists identify activities, but they have nothing generally in common; short of listing those taken seriously by scientists at a given time, we can't say anything about what they are. It remains possible that the quest to find "a general account of causality like Hume's, Hempel's, or Woodward's" is misguided, and that we'd be better off talking only of particular activities (Bogen 2008: 214). Yet, if we take these claims seriously, the content of the first restriction on explanatory models—that they call on causal dynamic principles—is completely opaque. Were I to offer a model containing a dynamic principle which (intuitively) reflected relations of correlation—such as the model above that referred to membrane vibration—all that could be said is that such a model is bad because it doesn't reflect activities, and that activities themselves were just the things that competent scientists talk about.

The Carving Standard

The second mechanistic explanatory standard insists that explanatory models truck in the *good parts* of a mechanism. These are sometimes called "working parts" (Bechtel 2008) or "working entities" (Darden 2008), though they are most commonly labeled "components" (Bechtel and Abrahamsen 2005: 425; Craver 2006: 369; 2007: 188), terms I use interchangeably. In contrast to a gerrymandered part or "piece," which can result from any conceivable decomposition, including those that "slice," "dice" or "spiral cut" a mechanism, component "cut mechanisms at their joints" (Craver 2007b: 187–188; see also 2007a: 5). As such, components are not mere results of "arbitrary differentiation" (Bechtel 2008: 146).

Requirements on *components* aim to solve the *carving problem*. Though all mechanistic explanations bridge the inputs and outputs of a system with a veridical mechanistic model, there are multiple ways of decomposing a system into organized parts. Furthermore, multiple mechanistic models—that is, those reflecting different decompositions—can bridge inputs and outputs as required. Such alternative models describe the internal working of the same system(s) using different vocabularies. In these alternative terms, the models package some of the same information—most notably, information about how output states depend on input states. Yet, none of these models can be censored for being non-mechanistic or false.

In the face of these false riches, the carving problem is that of providing a principle that distinguishes the good explanatory models from the bad. On the one hand, it is very clear that, in explaining various goings-on, scientists routinely carve mechanisms into *good parts*, rather than gerrymandered entities. But, on the other hand, it isn't transparent what—if anything—this practice is tracking. Fundamentalists may try to sidestep the issue by asserting that—appearances aside—the only appropriate explanations are those that "carve" systems into their fundamental physical constituents governed by physical laws. In contrast, however, many new mechanists do embrace explanations appealing to non-fundamental parts and properties. This gets them much closer to actual scientific practice, at the cost of then needing to specify which "high-level" mechanistic models are appropriate.

Good Parts as Components

In the context of addressing a variety of different topics, including but not limited to the carving problem, Carl Craver has articulated a number of features that "good" or "real" parts, also called "components," must possess (2007: 128–133, 187–195).[9] These features are a mix of epistemological and more metaphysical requirements. All are rather undisputed as necessary conditions on the parts described by mechanistic models, and are frequently mentioned by proponents of the mechanistic approach to explanation.[10]

[9] In particular, in addition to potentially addressing the carving problem, these conditions are offered as standards for distinguishing models that appeal to "real parts" from those that describe "fictional posits"(Craver 2007: 128–133).

[10] I focus on Craver's presentation because it is the most systematic available, but it is characteristic of the new mechanist literature. For instance, compare Darden's (2008: 961–962) discussion of "working entities" and Machamer et al.'s (2000: 5–6) comments on individuation of entities and activities.

1. *Robustness*: components should "be detectable with a variety of causally and theoretically independent devices" (2007: 132).
2. *Manipulability*: it should be possible "to manipulate the entity in such a way as to change other entities" (2007: 132).
3. *Plausibility*: components should be "physiologically plausible" (2007: 132).
4. *Stability*: components should have a "stable cluster of properties" (2007: 131) and should be "loci of stable generalizations" (2007: 190).

The first standard is that components be *robust*. Though some discussions of robustness have a more metaphysical cast, the variety of robustness at issue here is epistemic. To say that a component is robust is simply to say that it is detectable by different kinds of devices, optimally those operating on different principles. This standard is inspired by the usefulness of multi-device detection in helping scientists to distinguish genuine features of a system from artifacts (Culp 1994).

Yet, robust detectability will not address the carving problem. First, no device detects individuated parts *as such*, and consequently no part—component or otherwise—can be detected by more or fewer devices than another. To illustrate, consider an electron micrograph of a cell. Such a micrograph is (roughly) a representation of the electron density of material in different regions. Patterns in the density revealed by electron microscopy can provide evidence about the features of particular components, such as the shape of a membrane channel. The micrograph itself, however, does not detect which *of the pieces are components*; a carving into components is something that the scientist brings to the micrograph to interpret it.

An alternative to insisting that components be detectable by different devices is to suggest that the *properties* of components, as opposed to parts, be so detectable. The problem with this alternative is that components and gerrymandered pieces will pass the test equally: we can detect the properties of protein channels as well as quarter-neurons using a variety of normal neurophysiological devices. Thus, it does not appear that robustness will contribute to solving the carving problem.

The second standard is *manipulability*. This standard requires that a good part be *itself* manipulable in the service of affecting *something else*, a constraint inspired by Ian Hacking's (1983) famous call for "entity realism," according to which we deem theoretical entities "real" when it is possible to *do things* with them. As he put it, "if you can spray [them] then they are real" (1983: 24). Craver explains his particular application of this idea as follows: "[i]t should be possible… to manipulate the entity in such a way as to change other entities, properties, or activities" (132). Understood in this way, the quarter-neuron model—one of many that we must censor—will pass the test, as it is perfectly possible to manipulate a hunk of a neuron to affect something else. In consequence, manipulability appears no better off than robustness in distinguishing components from gerrymandered parts.

The third standard on good parts is called *plausibility*. Here, Craver requires that components actually exist in the systems under consideration, rather than "only under highly contrived laboratory conditions or in otherwise pathological states" (Craver 2007: 132). This suggestion is designed to rule out models that describe parts not present in the systems whose behavior is being explained. Here again, we have a principle that does not help address the carving problem. Just as do components, gerrymandered parts can "exist" in non-pathological conditions, and are thus "plausible" to treat as entities with respect to a behavior that a mechanistic model aims to explain.

This brings us to the final standard on components: that they have "a stable cluster of properties" (Craver 2007: 131). In a related discussion, Craver suggests that components—which themselves can be understood as submechanisms composing larger mechanisms—be "loci of stable generalizations" (Craver 2007: 190). In contrast to the three conditions just reviewed, there are prospects for developing this constraint in a way that allows mechanists to address the carving problem.

The stability condition asserts that a part's status as a component depends on its possessing a stable cluster of properties. A component's properties are stable, I will presume, if they would be maintained across some range of background conditions. Any component with such a property cluster will be one about which we can frame generalizations that are, to some degree, counterfactually stable. The virtues of such generalizations are legion, and a preference for them in explanatory contexts is uncontroversial (Mitchell 2000; Woodward 2001). It thus appears to make sense to "carve" systems, in explanatory contexts, in ways that allow those systems to be described by stable generalizations.

However, any appeal to stability to solve the carving problem must provide more analysis than this. First, the two most straightforward interpretations of the requirement, which are positioned on opposite extremes, either will fail to distinguish good carvings from bad or will be at odds with other commitments of the new mechanistic program. On the one hand, one cannot simply insist that components possess a cluster of properties that is *in some respects* stable, since gerrymandered parts will meet this minimal standard. Yet, on the other hand, mechanists also cannot say that components are pieces with *the most stable* property clusters. This position is unavailable because it is in direct tension with one of the animating motivations of the mechanisms movement: the rejection of proper laws as explanatorily central. The problem is straightforward: to insist on carving mechanisms into components with the maximally stable cluster of properties—that which can be described in terms of maximally stable generalizations—would require modeling mechanisms in terms of basic physical components, governed by "causal dynamic principles" which are physical laws. But to explain system functioning in these terms is clearly not to the mechanists' taste—and for good reason. Scientists, particularly life scientists, explain systems functioning without appealing to proper laws, and do so in terms of parts with property clusters that are often wildly unstable from a physical point of view—for example, proteins which denature in all but a narrow range of pHs, or cell membranes which fragment in all but specially tuned barometric circumstances—yet these models at least appear to provide superior explanations to those provided by lower-level physical accounts. In light of these complexities, mechanists require a version of the stability condition that is substantially more nuanced.

Such a nuanced requirement could be constructed in a number of ways and which would be impossible to exhaustively survey here. Instead, consider one intermediate approach to the stability standard that seems in line with the basic commitments of the new mechanistic program: provide principled guidelines on the range of background conditions over which part properties must be stable, with that range being somewhere between the minimal and maximal standards just considered. Parts with properties stable over that range are "components"; those which are not are "mere pieces." This stability range can be extracted from the stability properties of the explanandum. In particular, consider this constraint: a part is a component of a mechanism for a behavior if the relevant properties of the part—in particular, those of its properties that underpin the mechanisms' behavior—are stable, at a minimum, throughout the range of conditions over which the mechanism's overall behavior is stable.

Why might one want a part's property stability range to be determined by that of the stability of the overall system's behavior? Arguably, because it is only a part with this characteristic that could actually underpin the behavior to be explained. After all, mechanism-level behaviors—such as the input–output relationships that are the target of most mechanistic explanations—themselves have modal scope, holding in at least some range of background conditions. If a mechanistic model is to fully account for such a modally robust explanandum, the parts appealed to in the model must themselves survive—maintaining their property clusters—over that same range.[11]

In illustration, recall the explanandum behavior discussed above that neurons release neurotransmitters when exposed to neurotransmitters. This behavior holds of neurons over a range of conditions—in different temperatures, different ionic environments, and so on. Among the neuronal components critical for the behavior are the ligand-gated ion channels located in the dendrite membrane. The channel properties relevant to the overall mechanism's behavior—most notably, their disposition to open in response to neurotransmitter binding—must be stable over a range of background conditions in order for the mechanistic model to account for the stable systems behavior. Imagine, for instance, that in some condition in which the system behavior was maintained, the ion channel was denatured, and thus no longer possessed the property relevant to the behavior under analysis. Were this to be the case, one could not model the behavior in terms of these parts.

How might this standard reject gerrymandered parts? The contrast between the quarter-neuron model and the Standard Model can illustrate. Consider the range of background conditions over which the parts

[11] There are situations more complicated than this. If a mechanism contains a variety of redundant subsystems—each of which has a different range of stable functioning—the overall mechanism behavior could have a range of stability greater than that of any particular component, or component pathway. Yet, this possibility doesn't undermine the more generic suggestion that some identifiable relationship exists between the stability of a mechanism's parts' properties and the mechanism's systems-level behavior, and that this connection might be used to determine the relevant stability range required of mechanism parts.

represented in the quarter-neuron model would maintain their properties, as well as the range over which the macromolecules in the Standard Model would do so. At first glance, in contrast to the macromolecules, it may appear that the quarter-neuron will fail to possess properties as stable as required. Its properties will change in a broad range of circumstances, as the quarter-neuron will be modified in some way just in case any of its proper parts is so modified. Thus, this revised standard maybe effective, and the carving problem solved.

Unfortunately, the proposal just described is not strong enough to distinguish good parts from mere pieces, and can only be used to rule out non-veridical models, not those reflecting inferior carvings. The problem is that many gerrymandered pieces, correctly characterized, will in fact possess properties that are just as stable as required by the constraint—that is, as stable as the behavior of the overall mechanism. This is because only the properties *that underpin the mechanism's* behavior need to be so stable, according to the standard under consideration here. Although it is true that a relatively large part—gerrymandered or otherwise—like the quarter-neuron, will change *in some ways* in the face of a wide array of background circumstances, it will not change as often with respect to the properties that underpin mechanism behavior—those determining its capacity to bridge the relevant inputs and outputs. In fact, with the caveat noted above, it will maintain these properties at least over the range for which the system-level behavior is stable. One might be tempted to reject such properties as peculiar or gerrymandered—and thus not those whose stability is relevant for determining component-hood. However, this would be to make one's account of "good parts" dependent upon a substantive account of "good properties," which mechanists don't provide. Thus, the tactic shows little promise. A stability constraint—at least in the version I've proposed—cannot solve the carving problem.

Good Parts as Mutually Manipulable

A more sophisticated tool that might better address the carving problem is the mutual manipulability (MM) standard, proposed in Craver (2007a, b). It aims to provide conditions for when "a part is a component in a mechanism" (Craver 2007b: 141). Given that the term "component" is used in explicit contrast with mere "pieces" or "parts" (Craver 2007b: 188), the MM standard appears to be framed to solve the carving problem. It offers conditions for what are called "relevant" components via two basic requirements on the relationship between a component and a whole mechanism. These conditions require that something about the whole mechanism depends on the features of the component, and conversely, that something about the component depends on the features of the whole mechanism. More particularly, a part is a component of a mechanism for a behavior if the following conditions are satisfied:

(A) Intervening to change the component can change the behavior of the mechanism as a whole; *and*
(B) Intervening to change the behavior as a whole can change the behavior of the component. (Craver 2007b: 141)[12]

These conditions are loosely inspired by the interventionist account of causation, and both (A) and (B) are counterfactual conditions.[13] They are either true or false depending on whether some ideal causal manipulation—here called an "intervention"—which need not be possible to actually carry out, *would have* the specified result. Depending on the particular intervention–result pairing, this result might be a causal consequence of the change brought about by the intervention, or it could follow constitutively from that change, just as an intervention to increase the mass of my foot would change the mass of my whole body.

The first part of the MM condition, labeled (A) above, has two elements in need of refinement, one involving the *intervention to change the component*, and the other the *change in the behavior of the mechanism as a whole*. With respect to the first element, what would it mean to *intervene*

[12] Craver sometimes presents the standard, quite reasonably, using his own symbolism. For instance, another version of (A) requires that "there is some change to X's φ-ing that changes S's ψ-ing" (2007b: 153). Though these alternative statements are compatible with the interpretation I give of the MM standard, and have informed my presentation, I do not use Craver's notation because it would require too much space to adequately explain.

[13] Though this statement is from Craver's (2007a), in explicating the view I am very influenced by Craver's presentation in his (2007b). In correspondence, he reports that his presentation of the standard there is particularly careful.

to change the component? Note that use of the term "intervention" here, though clearly inspired by its use by causal interventionists, should not be understood in the precise technical sense defined by them (e.g., Woodward 2003) but instead as another sort of in-principle causal manipulation, sometimes glossed simply as "wiggling" (Craver 2007b: 153; 2007a: 15). With this in mind, there are two genres of change that might be intended. First, the manipulation might change the *input to the component*. For instance, in the case of a part like a ligand-gated ion channel, a change might involve exposing the channel to neurotransmitters, something that would have a variety of downstream effects, the most direct of which is the opening of the channel. Second, such a change might be made to the features underpinning the *input–output regularity* realized by the component itself. Again, focusing on the ligand-gated ion channel, a "wiggling" of the input–output relationship could involve a modification of the channel's disposition to open upon neurotransmitter binding. A parallel ambiguity faces the second half of the (A) condition—that involving the resulting change to the *behavior of the mechanism as a whole*. This could involve a change (from some default) of the output produced in a particular circumstance, or a change to the overall input–output relationship that the mechanism underpins.

In light of these alternative versions of the condition—both with respect to the feature intervened upon and the consequent change—I distinguish between two versions of Craver's condition (A).

(A_i) intervening to change the input to a component (from a default input)
changes the output of the mechanism as a whole (from a default output).

(A_{ii}) intervening to change the input–output relationship realized by the component changes the input–output relationship realized by the mechanism as a whole.

Some examples used to illustrate the MM standard fall under (A_i), while others align more with (A_{ii}). For instance, indicating the relevance of the first version, Craver (2007a, b) suggests that what he calls "activation experiments" can (sometimes) test the fulfillment of the condition, experiments in which one activates a component, apparently by setting its inputs in a certain way, and evaluates the consequences of this intervention on the system-wide output. On the other hand, indicating the relevance of the second version, Craver describes "interference experiments." In this case, the intervention can involve completely destroying, or more subtly modifying, the characteristics of the candidate component, and investigates change to the capacity of the whole mechanism. Fortunately, it will not be necessary to determine which refinement of the (A) condition is most defensible. Instead, I will probe the efficacy of both versions.

The second half of the MM standard, (B), requires *that intervening to change the behavior as a whole can change the behavior of the component*. The most obvious uncertainty here concerns what it means to intervene "on the behavior of the whole." A prima facie worry is that one can only intervene on the behavior of a whole by intervening on the behavior of its parts (individually or in combination); if so, triviality threatens, since there will always be some change to the behavior of the whole that changes the behavior of the component, namely, an intervention that changes the behavior of the whole just by changing the behavior of the component.

Fortunately, Craver suggests a more substantive reading of (B). An "intervention on the behavior of the whole" is just one that *sets the input conditions on the mechanism* in a certain way, that is, one that sets the inputs to those required to bring about the particular system-wide output that is of interest (Craver 2007b: 146). The resulting "change in the behavior of the component" is a change to its output (rather than to the features underlying its capacity to produce certain outputs given certain inputs). Thus, reconsider (B) as follows:

(B*) Intervening to change the input to the whole mechanism, such that it will bring about a particular output of interest, can change the output of the component.

Can these standards—A_i, A_{ii}, and B*—distinguish parts and components? According to (A_i), intervening to change the input to a component (from a default input) can change the output of the mechanism as a whole (from a default output). This will not help rule out gerrymandered pieces, since some changes to the inputs to such pieces—such as quarter-neurons—can change the outputs to whole mechanisms. In particular, changing the input to any of the quarter-neurons can lead a neuron to release neurotransmitters. According to (A_{ii}), intervening to change the input–output relationship realized by the component should be able to change the input–output relationship realized by the mechanism as a whole. Again, bad parts, such as quarter-neurons, pass this test without event. After all, changes to the disposition of a quarter-neuron can change the relevant disposition of the neuron as a whole. Finally, consider (B*), which requires that intervening to change the input to the whole mechanism, such that it will bring about a particular output of interest, can change the output of the component. Again, this cuts no ice against the bad parts. If we were to "intervene on the whole" by setting the inputs to the whole system in the right way, perhaps by exposing the system to neurotransmitters, the output of any of the quarter-neurons would change. Consequently, even bad parts—those we'd loathe to consider components—will pass the MM test, and that test proves not to be the constraint on components that was needed to fill out the explanatory account.

Good Parts as Scientifically Approved

Given the above difficulties, consider a very different kind of reaction to the carving problem. This down-to-earth reply is inspired by the explanatory practice of scientists themselves. Scientists don't break up the world any-which-way but rather have cultivated schemes of division which are somewhat (though not entirely) uniform within subdisciplines. These schemes award certain parts a scientific seal of approval. Such a practice might appear to provide a solution to the carving problem, one that simply insists that it is to *these only* that mechanistic models must refer. Machamer et al. (2000) gesture at such a proposal when they write that "the components [are those] that are accepted as relatively fundamental or taken to be unproblematic for the purposes of a given scientist, research group, or field" (13).

While this is a reasonable starting point for an inquiry into partitioning practices, as an answer to the carving problem it should be rejected as philosophically deflationary. Leaving the solution here is to make one's philosophical account into a science-reporting task. The philosopher offering it has made little progress in explaining scientific explanatory activity but has simply insisted that—with respect to the parts described—good explanations are just what competent scientists offer as such. This no more illuminates the nature of explanation than the cynic's account of species—according to which, species are groups of organisms recognized as species by taxonomists—illuminates the nature of kinds. The day may come when philosophers, having failed to solve the carving problem, should proclaim a cynic's slogan. Yet this will be a retreat, and a major concession with respect to the intelligibility of the scientific enterprise.

The Levels Standard

The final guideline on explanatory mechanistic models favors models that describe systems at the right "level," usually the one *just below* (in a sense to be explored) the phenomenon to be explained. There are both reductive and (arguably) non-reductive dimensions to this suggestion. First, in insisting that phenomena be explained in lower-level terms—by describing organized components of mechanisms and their interactions—the mechanistic approach to explanation is, undoubtedly, somewhat reductive. However, the approach is also in some measure non-reductive, in view of advocates' resistance to what we might call "fundamentalism," according to which every phenomenon is best explained at the physical level, by a model referring exclusively to physical parts, properties, and laws. Bechtel, for instance, explicitly contrasts his semi-reductive mechanistic view with a fundamentalist account, suggesting that "knowing how the components [of a mechanism] behave and understanding how they are organized is sufficient for the purposes of explaining how the mechanism as a whole behaves" (Bechtel 2008: 151) and that, in most cases, "there is no incentive for performing further decomposition" (ibid). Similarly, Craver, while acknowledging reductive dimensions to the mechanistic approach, still promises to provide mechanists "with the tools to challenge reduction as a normative model" (Craver 2007b: 111).

I will call this alternative to fundamentalism the "cascade view." According to it, whole-mechanism behaviors should be explained in terms of the mechanism's immediate component parts and relations. While such components can *themselves* be seen as even smaller mechanisms, and their behaviors explained using mechanistic models describing each of their own parts, relations, and dynamic principles, the cascade view denies that explanations for the functioning of submechanisms (e.g., components) can be plugged into the explanation for the functioning of the mechanism as a whole. Instead, "successively lower-level mechanisms account for *different* phenomena. Scientists construct a cascade of explanations, each appropriate to its level and not replaced by those below" (Bechtel and Abrahamsen 2005: 426). If the cascade view is correct, an endeavor to explain some phenomenon via a mechanistic model that describes parts and relations located at a non-adjacent level—say, one explaining the regular cardiac rhythm by appeal to a mechanistic model that trucked in atomic constituents—would blunder; its explanatory power would be weaker than that of a comparatively high-level model. In this way, the cascade view rules out the *zooming errors* from section "Formulating Explanatory Constraints."

This basic take on proper explanatory levels is enormously attractive, as it appears to mesh perfectly with scientific explanatory practice, particularly in the life sciences. It seems that, with respect to level, scientists offer just the kind of explanations that the cascade view would recommend—reductive but almost invariably "just below" the phenomenon to be explained, and far more abstract than fully fundamental ones. Yet, the move from this feature of explanatory practice to the more ambitious normative claim about "explanatory power" stated in the previous paragraph—though natural—is not irresistible. And fundamentalists will resist it, partly by trying to make sense of this aspect of scientific practice in pragmatic terms, explaining the fact that the explanations offered by scientific papers and textbooks are high level while not conceding that these explanations are objectively superior to fundamentalist ones. For instance, perhaps full explanations are not offered simply because human minds are too weak to grasp them at once. More important than the particulars of any "error theory" is the fact that reductionists will deny that the lack of fully spelled-out fundamentalist explanations in the scientific literature should be explained by the fact that such explanations are not, in principle, explanatorily optimal.

Under pressure from such an alternative, the cascade view requires articulation and defense. In particular, there are two (related) dimensions—one descriptive and one normative—along which buttressing is mandatory. The first and most pressing concern, the *levels problem*, involves simply filling out the cascade proposal by making sense of what "levels" are, including an adjacency relation between them. Though some philosophers can afford to remain silent on this topic—and may even deny any genuinely "leveled" aspect of nature that explanatory levels could track (as in Heil 2005; Strevens 2008)—the advocate of the cascade view cannot skate over it: it lies at the heart of her scheme.

The second topic, the *stop problem*, concerns the respect(s) in which locally reductive explanations are better than those that describe systems in terms of even more basic parts, relations, and principles. At first blush, fans of the cascade view may try to reject this question and to shift the burden of proof back to the fundamentalist. Why not instead insist that *she* defend her diabolical drive to explanatorily descend to the basic physical level, rather than resting satisfied with what most scientists actually dole out—locally reductive explanations? While dialectically tempting, this move is suspect. The cascader and fundamentalist are not equivalently positioned, as the cascade view is distinctively threatened with internal inconsistency. This is because the cascader has taken one step toward reduction, believing as she does that global models—those that treat systems as opaque black boxes—are *not* explanatory, and that the behavior of a complex system *should be* explained by breaking it into organized lower-level parts and their interactions—but then denies the value of further deepening. Yet, whatever explanatory oomph the mechanist gets from analyzing systems in terms of their immediate components, it seems she would get even more from analyzing them into their ultimate components. So, by her own lights, analysis all the way down to the physical should be preferred. In consequence, mechanists must say what is gained (or, at minimum, what is not lost) by stopping mechanistic explanations just one level down.

Though I would prefer to explore both of these issues, for reasons of space I restrict attention to the problem of levels.[14] After all, to even evaluate the mechanists' solution to the stop problem, we would need to know *where* we are advised to stop our mechanistic decomposition.

Though it is customary to see the world as "leveled," just what this involves is notoriously murky. When levels are judged to be "features of the world rather than… features of the units or products of science" (Craver 2007b: 177), they may still be understood in a number of ways. Lacking the space to consider all options, my focus here will be on the view of ontological levels clearly ascendant in the new mechanist literature: *levels of mechanisms*.[15] These levels are species of levels of composition, where the composites in question are whole mechanisms. Since mechanisms are (at least often) embedded within one another, levels of mechanism lend themselves to an adjacency relation: X is one level below Y just in case X is *an immediate component of the mechanism that is* Y. Founding figures in the mechanistic program have expressed sympathy for this view, with Glennan "construing the layers that make up the world in terms of nested mechanisms" (2010a: 363), Craver seeing levels as "levels of mechanisms," in which "lower levels… are the components in mechanisms for the phenomena at higher levels" (2007b: 170), and Bechtel sketching a largely comparable view of "levels within a mechanism" (2008: 147).

This view has three principal features. First, because each of a mechanism's immediate components—themselves understood as smaller mechanisms—may have its own immediate components, which possess components likewise, mechanistic levels can be multiply embedded. Second, all facts about the relative "level" of two things will be a joint function of mind and world; thus, to call these levels "ontological" or

[14] For a critique of the mechanists' most promising response to the stop problem, that offered by difference-making accounts of causal explanation as articulated by Woodward (2003, 2010), and adopted explicitly by Craver (2007), see Franklin-Hall (2016). For my own positive proposal on the stop problem, see Franklin-Hall (forthcoming). A recent paper on this problem that came out too late for me to consider is Harbecke (2015).

[15] For a detailed account of the different things philosophers have meant by "level," see Craver (2007, Chap. 5).

"features of the world" would be, by my lights, to overreach. This follows from the fact that mechanisms themselves—and their componential specifications—are only well defined (if at all) relative to some behavior. And no behavior is delivered to us by the world, but must be picked out by us. Third, even relative to a chosen behavior, questions about the relative level of any two things can be ill-posed. Such questions are only kosher when both entities in question are components (either immediate or otherwise) of the mechanism in question.

The suggestion just sketched offers a more scientifically plausible, and nuanced, understanding of our folk conception of levels than do the global, flat stratifications advanced by Oppenheim and Putnam (1958). And compared to levels defined in terms of the philosophically esoteric—laws, properties, and causes—levels of composition can appear innocent and straightforward. Furthermore, and of central importance here, levels of mechanistic composition can be naturally recruited to provide constraints on proper mechanistic explanation, as follows: for any phenomenon that one might want to explain, there is a mechanism responsible for it, positioned at level *n*. To explain the phenomenon, an explanatory mechanistic model should describe entities—that is, the immediate component parts of the mechanism—at one level down, at *n*–1.

Yet does this proposal address the levels problem, characterizing what it means for one thing to be *one level below* another? If so, it is only by way of a substantial promissory note. The problem follows immediately from the difficulties already encountered in distinguishing "components" or "good parts," thus this discussion can be brief. Levels of mechanistic composition are only well defined if linked to an account of what is required for a part to be an *immediate component* of a mechanism for a behavior. Immediate components must themselves meet two conditions. First, they must be genuine *components*, not gerrymandered parts or "pieces." Second, these good parts must be, in some sense, *just below* the mechanism as a whole (level *n*–1), and not components *of* components (level *n*–2). If supplied with a standard that, for any mechanism for a behavior, specified all of its nested components, one could make sense of which components were *immediate*; however, lacking a distinction between parts and components, the immediacy requirement is impotent, having no material on which to work. In light of this lacuna, even those willing to grant a response to the stop problem, and who see the cascade view as normatively superior to explanatory fundamentalism, should not yet consider it to be a genuine alternative; the levels standard cannot, from the surplus of minimally adequate mechanistic models, tell the good explanations from the bad.

Conclusion

Though attractive at first glance, none of the new mechanists' explanatory guidelines have survived scrutiny, successfully discharging the work assigned to them. This work, it is worth emphasizing, is extremely difficult. So, even granting that I am right that mechanists have yet to complete it, this hardly shows that their general framework, and particularly their commitment to causal explanation, is mistaken. Rather, it suggests that the mechanistic account is but a story half-told. Thus far, proponents have labeled some important distinctions—such as between causal and correlational relationships, between components and mere pieces, and between appropriate and inappropriate explanatory levels. But the task of filling them out remains.

As I see it, the present shortcomings of the mechanistic explanatory account are the flip side of an admirable feature of the mechanism movement, one which has had a salutatory influence on contemporary philosophy of biology (and science): that of taking science (and particularly biology) seriously. I conclude by recalling the origins of the mechanists' explanatory project, in doing so noting both its merits and its limits.

From early writings to the present day, the new mechanists have been struck by what appears to be an evident mismatch between the DN analysis of explanation and explanatory practice in the life sciences. On the DN view, explanations are deductively valid arguments, in which a statement of the explanatory target is derived from true sentences, including one stating a law of nature. Reasonably enough, mechanists have found it difficult to make this jive with what scientists actually did. Where were these supposed arguments in scientific articles and textbooks? What were these strict laws in a science like biology, where exceptions are more than just distracting litter on a landscape of regularity? Something seemed to have gone wrong.

In the face of this apparently naïve philosophical precision, the new mechanists returned to the basics. Rather than imposing a highly regimented account of explanation on the science—one, quite typically, reflecting the philosopher's penchant for argument and logic—we were encouraged to look with fresh eyes *at the science*.[16] What kind of explanations did scientists really offer? Immediately clear was that explanations often showed *how things work*. Yet in moving beyond that, the situation became complicated. Scientists obviously provided explanations using a large number of different representational schemes, with deductive logic nowhere in view. They described causes but usually talked only of particular activities. And they talked frequently of these things they called *mechanisms* but provided no account of what they (in general) were.

All of these are important observations. A rich, scientifically responsible philosophy of science must be accountable to what scientists do—they are our subjects, and their practices, our data. Thus, a mechanist contribution has been in bringing interesting details of these practices to philosophical attention, from cell biology to studies of metabolism, neuroscience, and most recently to systems biology. But what tasks await, once these phenomena are in view? To say, I will apply to *philosophical* practice language that mechanists often use to describe *scientific* practice.

When studying explanation, philosophers aim not to explain "how things work" in the physical world but instead "how things work" when scientists show "how things work." To do so, philosophers must, after characterizing the surface features of explanatory practice, pry open its "black boxes," displaying the underlying "mechanisms" that account for scientists' very explanatory judgments. Is this just what new mechanists have done? Have they looked "under the hood" of explanatory practice, and detailed its workings? The results of this inquiry suggest not. Or, more sympathetically, it suggests that they have peaked under the hood, but have not yet gotten their hands dirty taking the engine apart.

[16] As Lindley Darden explains in her overview of the movement, "[t]his work on mechanisms in biology originated (primarily) not as a response to past work in philosophy of science but from consideration of the work of biologists themselves, especially in molecular biology and neurobiology and biochemistry and cell biology" (2008: 958–959). Similarly, Bechtel writes that "these accounts of mechanistic explanation attempt to capture what biologists themselves provide when they offer explanations of such phenomena as digestion, cell division and protein synthesis" (2007: 270).

In particular, rather than opening the black boxes of the scientific enterprise—with respect to causation, part individuation, and explanatory level—philosophers have (largely) taken those practices for granted.[17] Perhaps this results from a too-successful enculturation of philosophers into the scientific mindset, making it difficult to achieve the critical distance needed to philosophize *about* science. If so, while mechanists may be right that advocates of the DN account were *too far* from science to say anything true about it, perhaps the new mechanists have remained *too close* to science to say anything surprising about it.

References

Bechtel, W. (2006). *Discovering cell mechanisms: The creation of modern biology*. Cambridge: Cambridge University Press.

Bechtel, W. (2007). Biological mechanisms: Organized to maintain autonomy. In F. Boogerd (Ed.), *Systems biology: Philosophical foundations*. Amsterdam: Elsevier.

Bechtel, W. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. New York: Routledge.

Bechtel, W. (2011). Mechanism and biological explanation. *Philosophy of Science, 78*(4), 533–57.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanistic alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences, 36*, 421–41.

Bogen, J. (2005). Regularities and causality; generalizations and causal explanation. *Studies in History and Philosophy of Biological and Biomedical Sciences, 36*, 397–420.

Bogen, J. (2008). Causally productive activities. *Studies in the History and Philosophy of Science, 39*, 112–23.

Campaner, R. (2006). Mechanisms and counterfactuals: A different glimpse of the (secret?) connexion. *Philosophica, 77*, 15–44.

[17] There is one mildly ironic exception to my general diagnosis. The only putative black box that mechanists have opened is the scientists' concept of "mechanism." On reflection, this focus was imprudent. Not every concept used by scientists is meaty, and not every term reflects a genuine black box; "mechanism" is not a theoretical term within the science, but is a mere pointer, or placeholder—similar perhaps to the philosopher's term "conception."

Craver, C. F. (2006). When mechanistic models explain. *Synthese, 153*, 355–76.

Craver, C. F. (2007a). Constitutive explanatory relevance. *Journal of Philosophical Research, 32*, 3–20.

Craver, C. F. (2007b). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford: Clarendon.

Culp, S. (1994). Defending robustness: The bacterial mesosome as a test case. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, 1*, 46–57.

Darden, L. (2008). Thinking again about biological mechanisms. *Philosophy of Science, 75*, 958–69.

Dowe, P. (1995). Causality and conserved quantities: A reply to Salmon. *Philosophy of Science, 62*(2), 321–33.

Elster, J. (1989). *Nuts and bolts for the social sciences*. Cambridge: Cambridge University Press.

Franklin-Hall, L. R. (2016) High-level explanation and the interventionist's 'variables problem'. *British Journal for the Philosophy of Science*, 67(2), 553–577.

Franklin-Hall, L. R. (forthcoming). The causal economy account of scientific explanation. In C. K. Waters & J. Woodward (Ed.), *Minnesota studies in the philosophy of science*. Minneapolis, MN: University of Minnesota Press.

Glennan, S. (2002). Rethinking mechanistic explanation. *Philosophy of Science, 69*, S342–S53.

Glennan, S. (2005). Modeling mechanisms. *Studies in History and Philosophy of Biological and Biomedical Sciences, 36*, 443–64.

Glennan, S. (2010a). Mechanisms, causes, and the layered model of the world. *Philosophy and Phenomenological Research, 81*(2), 362–81.

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge: Cambridge University Press.

Hall, N. (2004). Two concepts of causation. In J. Collins, N. Hall, & L. A. Paul (Eds.), *Causation and counterfactuals* (pp. 225–276). Cambridge, MA: MIT Press.

Harbecke, J. (2015). Regularity constitution and the location of mechanistic levels. *Foundations of Science, 20*(3), 323–338.

Heil, J. (2005). *From an ontological point of view*. Oxford: Clarendon.

Hitchcock, C. (1995). Salmon on explanatory relevance. *Philosophy of Science, 62*, 304–20.

Kaplan, D. M., & Bechtel, W. (2011). Dynamical models: An alternative or complement to mechanistic explanations? *Topics in Cognitive Science, 3*(2), 438–44.

Kistler, M. (2009). Mechanisms and downward causation. *Philosophical Psychology, 22*(5), 595–609.

Leuridan, B. (2010). Can mechanisms really replace laws of nature? *Philosophy of Science, 77*, 317–40.

Levy, A. (2013). Three kinds of new mechanism. *Biology & Philosophy, 28*(1), 99–114.

Levy, A., & Bechtel, W. (2013). Abstraction and the organization of mechanisms. *Philosophy of Science, 80*(2), 241–61.

Machamer, P. (2004). Activities and causation: The metaphysics and epistemology of mechanisms. *International Studies in the Philosophy of Science, 18*, 27–39.

Machamer, P. (2009). Explaining mechanisms. http://philsci-archive.pitt.edu/5197/.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*, 1–25.

Mackie, J. L. (1974). *The cement of the universe*. Oxford: Clarendon.

Mitchell, S. D. (2000). Dimensions of scientific law. *Philosophy of Science, 67*, 242–65.

Nicholson, D. J. (2012). The concept of mechanism in biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 43*(1), 152–63.

Oppenheim, P., & Putnam, H. (1958). Unity of science as a working hypothesis. *The philosophy of science*, 405–427.

Railton, P. (1978). A deductive-nomological model of probabilistic explanation. *Philosophy of Science, 45*, 206–26.

Robert, J. S. (2004). *Embryology, epigensis, and evolution: Taking development seriously*. Cambridge: Cambridge University Press.

Salmon, W. (1984). *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.

Skipper, R. A., & Millstein, R. L. (2005). Thinking about evolutionary mechanisms: Natural selection. *Studies in History and Philosophy of Biological and Biomedical Sciences, 36*, 327–47.

Strevens, M. (2008). *Depth: An account of scientific explanation*. Harvard: Harvard University Press.

Waskan, J. (2011). Mechanistic explanation at the limit. *Synthese, 183*(3), 389–408.

Woodward, J. (2001). Law and explanation in biology: Invariance is the kind of stablity that matters. *Philosophy of Science, 68*(1), 1–20.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.