Mechanisms in Cognitive Science[1]

Carlos Zednik
carlos.zednik@ovgu.de
Otto-von-Guericke-Universität Magdeburg

forthcoming in S. Glennan & P. Illari (eds.), *The Routledge Handbook of Mechanisms and Mechanical Philosophy.* London: Routledge

1. Introduction

A principal goal of cognitive scientific research is to explain behavioral and cognitive phenomena such as perception, action, categorization, memory, learning, language, and attention. The most influential account of explanation in cognitive science is due to David Marr (1982). On Marr's account, cognitive scientists answer questions at three distinct *levels of analysis*: the *computational* level, which concerns questions about a particular system's computational goals and about the appropriateness thereof; the *algorithmic* level, which is driven by questions about the representations and algorithms that are used by the system to achieve these goals; and the *implementational* level, which addresses questions about the way in which these representations and algorithms are physically realized (Marr 1982: 24ff). According to Marr, questions at all three levels of analysis must be answered in order to "completely understand" a cognitive system, and to thereby explain its behavior.[2]

Although Marr's account remains influential to this day, there are reasons to be unsatisfied about its clarity and scope. First, although it is relatively clear which questions should be asked, it is not quite as clear how these questions might be answered. Among others, it is unclear what it actually takes to show that a computational goal is appropriate, and what it means for an algorithm to be physically realized. Second, because Marr's account is formulated in terms co-opted from computer science, rather than in terms endemic to philosophical discussions of scientific explanation, it is difficult to know how explanations in cognitive science compare to explanations in other disciplines that center on e.g. subsumption under law, the development of models, and/or the discovery of mechanisms. Finally, many of the terms that play a central role in Marr's account are far less prominent in cognitive science today than they were at the time of Marr's writing. In particular, the *computationalist* research program which predominated in the 1970s and 1980s (Pylyshyn 1980) now competes for attention and resources with alternative research programs such as *connectionism* (Rumelhart et al. 1986), *dynamicism* (van Gelder 1998), and the *Bayesian approach* (Zednik & Jäkel 2016), some of which may not rely on notions of 'computation', 'algorithm', and 'implementation' at all.

Despite these reasons to be unsatisfied with Marr's account of explanation in cognitive science, its lasting influence recommends it as a productive starting point for discussion. Indeed, the aim of this chapter is to show that many of the ambiguities in Marr's account can be resolved, and that its scope can be extended, by considering Marr to be an early advocate of *mechanistic explanation* (Chapter 1). Although Marr's account was originally designed to capture explanations in computationalist cognitive science, the questions it identifies at each level of analysis are in fact variations on the types of questions that are asked in any research program that aims to discover and describe (cognitive) mechanisms. Specifically, questions at the computational level can be construed as questions about *what* a mechanism is doing and *why*: questions that concern the mechanism's behavior and containing environment. Questions at the algorithmic level, in contrast, are questions about *how* a mechanism does what it does, and concern its component operations. Finally, questions at the implementational level of analysis can be understood as questions about *where* a particular mechanism's component operations are carried out—that is, questions about the component parts in which such operations might be localized. Although Marr showed how each one of these questions might be answered using the concepts and methods of computationalist cognitive science, they are also answered in the context of other research programs such as connectionism, dynamicism, and the Bayesian approach. In other words, there is reason to believe that all of these research programs seek three-level explanations, and moreover, that all of them aspire to discover and describe mechanisms.

This chapter will outline and defend a mechanistic interpretation of Marr's account of explanation in cognitive science, and thereby attempt to resolve the ambiguities that remain in this account, as well as to extend its scope. Notably, in line with Marr's claim that all three levels of analysis are necessary to "completely understand" a cognitive system, on the current interpretation all three levels must be addressed in order to provide mechanistic explanations of behavioral or cognitive phenomena.[3] In this way, the present interpretation differs from several previous attempts to relate Marr's account to the framework of mechanistic explanation (see e.g. Bechtel & Shagrir 2015; Bickle 2015; Kaplan 2011; Milkowski 2013; Piccinini & Craver 2011). Moreover, the present interpretation helps to address a number of philosophical debates concerning e.g. the role of idealization in cognitive modeling, the role of abstraction in mechanistic explanation, and the nature of the realization-relationship that obtains between functional processes in the mind and physical structures in the brain.

2. The computational level: "What?" and "Why?"

In Marr's original formulation, the computational level of analysis is defined by questions about a cognitive system's computational goals, as well as questions about the appropriateness thereof (Marr 1982: 24ff). More generally, these can be understood as questions about *what* a cognitive system is doing, and questions about *why* it does what it does (see also McClamrock 1991; Shagrir 2010).

What-questions can be answered by describing the relevant system's behavior. Within the computationalist research program, this involves specifying an information-processing function that maps the system's (sensory) inputs onto its (behavioral) outputs. That said, recent contributions by e.g. dynamicist researchers suggest that many behavioral and cognitive capacities—especially those that depend on continuous feedback from the environment—cannot be easily described as mappings between input and output (van Gelder 1995). Rather, these capacities are better described as continuous trajectories through state space, the dimensions of which might correspond to neural activity, a system's bodily position or motion, and/or features of the environment (Kelso 1995). Although it may be terminologically confusing to associate dynamical state-space trajectories with the "computational" level of analysis, they are analogous to information-processing functions in that they too can be used to describe a cognitive system's behavior, and thus, to answer questions about *what* the system is doing.

Answers to what-questions play an important role not only in Marr's account of explanation in cognitive science, but also in the framework of mechanistic explanation (see also Chapter 16). Descriptions of a cognitive system's behavior are descriptions of an *explanandum phenomenon* (Cummins 2000). In many scientific disciplines, such descriptions are the starting point of mechanistic explanation: The explanandum phenomenon is identified with the overall behavior of a mechanism, and an attempt is made to describe that mechanism's component parts, operations, and overall organization (Bechtel & Richardson 1993; Darden & Craver 2012; Chapter 19). Notably, many mechanisms are known whose overall behavior has been described as a form of information-processing (Craver 2013; Kaplan 2011), but also mechanisms whose behavior has been more effectively described as a trajectory through state space (Bechtel & Abrahamsen 2010). Therefore, these ways of answering what-questions in cognitive science are by no means inconsistent with the principles of mechanistic explanation (compare Chemero & Silberstein 2008). That said, merely describing a mechanism's overall behavior is insufficient for the purposes of mechanistic explanation (Kaplan & Craver 2011; Chapter 20). Indeed, insofar as descriptions of a mechanism's overall behavior often resemble law-like regularities (Chapter 12), they may also feature in other kinds of sci-

entific explanation (Chemero & Silberstein 2008; Zednik 2011). For this reason, determining whether cognitive scientists are in fact in the business of mechanistic explanation requires taking a closer look at the way in which they answer questions beyond "what?".

One such question is the question of "why?". In the context of his celebrated cash register example, Marr deems it important to ask "why the cash register performs addition and not, for instance, multiplication when combining the prices of the purchased items to arrive at a final bill" (Marr 1982: 22). Marr claims that such questions are answered by considering the "appropriateness" of the system's behavior with respect to the "task at hand" (Marr 1982: 24). Whereas many previous discussions of Marr overlook this aspect of the computational level, Shagrir (2010) has provided a compelling analysis according to which a cognitive system's behavior should be deemed appropriate when a mathematical description of that behavior can be mapped onto a relevant (potentially abstract or even counterfactual) property of the system's environment. One analytic technique that is well-suited for answering why-questions in this way is *rational analysis* (Anderson 1991; Oaksford & Chater 2007), which lies at the heart of the recently-influential Bayesian approach in cognitive science (Zednik & Jäkel 2016). This technique involves formally characterizing a cognitive system's task environment as a form of probabilistic inference, and deriving an optimal solution in the sense prescribed by probability theory. A widely-reported—and initially surprising—finding of this approach is that many different kinds of cognition and behavior closely approximate optimal solutions (Pouget et al. 2013). Whenever this is the case, why-questions can be answered because the optimal solutions describe a cognitive system's behavior while simultaneously reflecting a mathematical property of the system's environment—namely, the Bayes-optimal solution to a particular task. Intuitively, the method of rational analysis allows researchers to show that cognitive systems behave as they do *because* that way of behaving is optimal in the sense prescribed by probability theory (but compare Danks 2008).

Shagrir's analysis of 'appropriateness' implies that answers to why-questions cannot be found by looking solely at the cognitive system whose behavior is being explained; it also involves looking at the environment in which that behavior unfolds. That said, it is far from obvious how looking at properties external to a cognitive system might be conducive to revealing the mechanisms internal to it. Moreover, it is tempting to think of why-questions as pertaining to a teleological approach to scientific explanation which is traditionally contrasted with the mechanistic approach.[4] Nevertheless, philosophical proponents of mechanistic explanation have recently argued that environmental and teleological considerations play a significant role in mechanistic explanation. In particular, Carl Craver argues that understanding the role a mechanism is supposed to play

in a containing environment can greatly facilitate the task of describing that mechanism's actual behavior, and that many investigators for this reason "search for a higher-level mechanism within which it has a role" (Craver 2013: 153). Similarly, William Bechtel (2009) argues that mechanistic explanation involves "looking up" at the environment in which a mechanism is naturally embedded, because it greatly facilitates the task of characterizing the mechanism's actual behavior: It might reveal complexities that the mechanism must accommodate, as well as regularities that it might exploit.

These considerations suggest that answers to why-questions do in fact play a role in mechanistic explanation (see also Chapter 8), and indeed, that they may be instrumental for coming up with answers to what-questions. Notably, the Bayesian approach illustrates this kind of dependence of the *what* on the *why.* Recall that many different kinds of behavior and cognition have been found to approximate optimal solutions within a particular task environment. Although initially surprising, this finding is not accidental. Investigators regularly find discrepancies between an optimal solution and the observed behavioral data, but then often go on to tweak the specification of the task environment until the optimal solution is closely approximated by the data (Anderson 1991). Although some commentators denounce this kind of tweaking as a post-hoc model-fitting exercise of limited explanatory value (see e.g. Bowers & Davis 2012), others take it to be an efficient means of deriving mathematical formalisms that simultaneously answer what- and why-questions at the computational level of analysis, and that even facilitate the search for answers at lower levels (Zednik & Jäkel 2016).

In short, Marr's computational level of analysis concerns two types of questions: questions about *what* a cognitive system is doing, and questions about *why*. It is a mistake to consider either kind of question to be antithetical to the principles of mechanistic explanation. On the contrary, both kinds of questions play an important role in the discovery and description of mechanisms in several different research programs. But although the computational level of analysis can therefore be seen to play a critical role in mechanistic explanation, it is by no means sufficient. Just as Marr deems it necessary to answer questions below the computational level for the purposes of "completely understanding" a cognitive system and its behavior, mechanistic explanation also involves describing the internal features of a mechanism—its component parts, operations, and organization.

3. The algorithmic level: "How?"

Marr's algorithmic level of analysis is defined by questions about "the representation for the input and output", and about the "algorithm by which the transformation [between them] may actually be accomplished" (Marr 1982: 23). Questions of this kind are traditionally answered through *cognitive*

modeling, which involves *functionally analyzing* the complex behavioral or cognitive capacity being explained into an organized collection of simpler capacities (Cummins 1983), and subsequently describing this collection in formal mathematical or computational terms (Busemeyer & Diederich 2010; Luce 1995). In the computationalist research program, cognitive models often consist of lists of production rules with which to manipulate symbolic expressions (Pylyshyn 1980). These models are quite naturally viewed as descriptions of algorithms for transforming representations so as to achieve a particular computational goal. That said, cognitive models are a commonplace even in non-computationalist research programs. Indeed, some of the most influential cognitive models today consist of mathematical equations that determine the numerical values of output variables (Rumelhart et al. 1986; Nosofsky 1986), the evolution of state variables over time (Busemeyer & Townsend 1993), or the probability distribution over a hypothesis space (Pouget et al. 2013). It may not always be useful or even possible to think of these models as describing representation-transforming algorithms (Ramsey 2007; van Gelder 1995). Nevertheless, insofar as they can be used to reproduce a particular input-output transformation or a series of state-changes, they can still be said to compute a particular information-processing function or state-space trajectory. Insofar as the function or trajectory being computed accurately describes a particular cognitive system's behavior, the relevant cognitive model is a possible answer to a question about *how* that cognitive system does what it does (Marr 1982: 23, see also Cummins 2000; McClamrock 1991).

Notably, there are often many different ways to compute a particular information-processing function or state-space trajectory. For this reason, many different cognitive models may be developed to answer a particular how-question, and investigators need a way of distinguishing good answers from bad ones. Interestingly, Robert Cummins is sometimes interpreted as being unwilling or unable to make such a distinction. Consider the following oft-quoted passage:

> "Any way of interpreting the transactions causally mediating the input-output connection as steps in a program for doing ϕ will, provided it is systematic and not *ad hoc*, make the capacity to ϕ intelligible. Alternative interpretations, provided they are possible, are not competitors; the availability of one in no way undermines the explanatory force of another." (Cummins 1983: 43)

Because Cummins attributes the same degree of explanatory force to a (potentially) wide variety of models, he has been accused of being unable to distinguish between answers to how-questions that capture the way a cognitive system *actually* does what it does, from answers that merely capture the way it might *possibly* do so (Kaplan 2011; Piccinini & Craver 2011). This accusation seems

unwarranted, however; it fails to acknowledge Cummins' demand that cognitive models accurately reflect the "transactions causally mediating the input-output connection". Moreover, Cummins goes on to acknowledge that the elements of some cognitive models are more likely than others to be instantiated in a particular cognitive system—and that these models are for this reason to be preferred (Cummins 1983: 44, see also Feest 2003). Thus, Cummins does in fact outline a criterion for distinguishing good answers to how-questions from bad ones: Good answers are provided by cognitive models whose elements are in fact instantiated by the cognitive system being investigated, and that actually reflect the causally relevant factors that contribute to that system's behavior.

This criterion can be fleshed out by aligning it with the idea that cognitive models should accurately describe the internal features of the mechanism responsible for the explanandum phenomenon: its component parts, operations, and overall organization. More precisely, cognitive models should satisfy what has come to be known as the *model-to-mechanism mapping constraint* (3M):

> "In successful explanatory models...(*a*) the variables in the model correspond to components, activities, properties, and organizational features of the target mechanism that produces, maintains, or underlies the phenomenon, and (*b*) the (perhaps mathematical) dependencies posited among these variables in the model correspond to the (perhaps quantifiable) causal relations among the components of the target mechanism." (Kaplan & Craver 2011: 611, see also Chapter 17)

3M allows for the possibility that many different models have equal explanatory force—as long as the rules, symbols, equations and variables posited by each one of these models correspond to the features of the target mechanism. Moreover, although 3M requires that the features of a mechanism be described correctly, it does not require that they be described completely. In this spirit, Piccinini & Craver (2011) argue that cognitive models are designed to provide *mechanism sketches,* elliptical descriptions that correctly, albeit incompletely, describe a mechanism's component parts, operations, and/or organization. Although mechanism sketches satisfy 3M by correctly describing some of a mechanism's features, additional "filling in" is required in order to transform these sketches into full-fledged mechanistic explanations (Craver 2007; Machamer et al. 2000, see also Section 4 below).

Several commentators have argued that cognitive models can in fact be viewed as mechanism sketches. To this end, they have considered examples from a variety of research programs (for discussion see e.g. Abrahamsen & Bechtel 2006; Milkowski 2013; Zednik 2011).

That said, other commentators have identified counterexamples in the form of models that contain *idealizations*: constructs that do not correspond to any one of a particular mechanism's features (see also Chapter 17). For example, Weiskopf (2011) introduces Hummel & Biederman's (1992) model of object-recognition, which contains *Fast Enabling Links* (FELs) that "possess physically impossible characteristics such as infinite speed" (Weiskopf 2011: 331). Similarly, Buckner (2015) considers the role of backpropagation learning in connectionist networks, the execution of which depends on the biologically implausible "backwards transmission of information across neural synapses, the need for prior knowledge of correct output, and the distinct, individualized error signals used to adjust the thresholds and weights of each node and link in the network" (Buckner 2015: 3925). Because FELs and backpropagation learning are psychologically, biologically, and/or physically implausible, cognitive models that incorporate these constructs can be known with relative certainty to *not* satisfy 3M.

Although there are counterexamples to the claim that cognitive models generally *satisfy* 3M, there is nevertheless reason to believe that these models are *constrained* by 3M. That is, there is reason to believe that FELs and backpropagation learning are subject to replacement by less-idealized constructs as the explanatory demands on the relevant models increase. Notably, Weiskopf denies this claim, arguing that FELs are "not clearly intended to be eliminated by any better construct in later iterations" (Weiskopf 2011: 331). The key premise in Weiskopf's argument is that FELs allow Hummel & Biederman to understand the unique contribution of synchronous firing to object-recognition, independent of other factors such as mutual excitation and/or inhibition. However, it is important not to conflate intelligibility with explanation: Although FELs may allow investigators to understand the unique contribution of one causally relevant factor, the use of such idealizations often comes at the cost of obscuring or altogether neglecting the contribution of other factors. Thus, although models that contain idealizations may reproduce gross behavioral trends, they are unlikely to capture precise quantitative detail such as reaction-times and learning curves that may be of significant explanatory interest. Indeed, as Buckner goes on to argue, it is in order to capture just this kind of detail that connectionist researchers often seek to replace backpropagation learning with "more biologically plausible training rules" (Buckner 2015: 3925). In general, as the explanatory demands on a particular model increase, it seems likely that idealizations will eventually be replaced by constructs that more accurately reflect the causal structure of a mechanism. Even if the cognitive models being used today do not satisfy 3M, future iterations of these models will presumably strive to do so. For this reason, the requirement that cognitive models correctly describe mechanisms in the sense of 3M is useful for determining whether these models provide good or bad answers to questions about *how* cognitive systems do what they do.

4. The implementational level: "Where?"

How-questions at the algorithmic level of analysis are often the primary concern of research programs in cognitive science. Nevertheless, it would be a mistake to think that an explanation has been provided just as soon as these questions have been answered. As has already been stressed repeatedly, on Marr's account all three levels of analysis are needed to "completely understand" a cognitive system and explain its behavior. Accordingly, investigations at the computational and algorithmic levels must be supplemented by investigations at the implementational level, which are driven by questions about the way in which the constructs of a cognitive model are "realized physically" (Marr 1982: 25; see also Polger 2004). Unfortunately, it remains unclear what it actually takes to show that the production rules, symbols, equations and/or variables specified by a cognitive model are realized by complex and potentially unruly physical systems such as the brain. The aim of this section is to show that this lack of clarity can be remedied by considering what it takes to "fill in" a sketchy cognitive model so as to deliver a full-fledged mechanistic explanation.[5]

In general, "filling in" is a matter of describing a mechanism in greater detail than before (Craver 2007; Machamer et al. 2000). Although the present discussion embraces the view that cognitive models are mechanism sketches which leave out certain details, it has not yet been discussed exactly what kinds of detail are typically left out. Because most cognitive models specify mathematical constructs such as lists of production rules or systems of equations, it is tempting to think of them as specifing details about a mechanism's abstract mathematical or computational properties, rather than about its concrete physical properties. Indeed, this view is quite widespread in the literature, despite often being left implicit (see e.g. Bechtel 2008; Bechtel & Shagrir 2015; Chemero & Silberstein 2008; Craver 2007; Danks 2008; McClamrock 1991; Piccinini & Craver 2011; Shagrir 2010; Stinson 2016). Notably, on this view the details that are included may pertain to any one or more of a mechanism's internal features. For example, on this view cognitive models may describe a mechanism's component operations in terms of e.g. their informational properties, rather than in terms of neuronal spike trains (Pylyshyn 1980; Weiskopf 2011). They may also specify a mechanism's component parts as e.g. filters, without identifying these with any particular neural structures (Stinson 2016). Finally, they might characterize a mechanism's overall organization as a graph, abstracting over anatomical detail (Bechtel & Shagrir 2015; Levy & Bechtel 2013).

Thus understood, the realization-relationship that obtains between the algorithmic and implementational levels is one of *instantiation*; "filling in" involves specifying the concrete physical properties that instantiate a particular set of abstract mathematical properties. Thus,

whereas investigations at both levels describe the same component parts, operations, and/or organization of a particular mechanism, they differ with respect to the kinds of properties being described: Abstract mathematical properties on the one hand, and concrete physical properties on the other. Put differently, although the questions being asked at each level are fundamentally the same—they are how-questions in each case—the answers being given differ because they are articulated at different levels of abstraction.

Although this view is relatively widespread in the literature, it faces an important challenge: Why should it be necessary to answer a how-question twice? Since a mechanism's abstract mathematical properties are instantiated by its concrete physical properties, why should the former need to be cited in an explanation, in addition to the latter? Polger (2004) argues that mechanisms can compute information-processing functions or state-space trajectories just in virtue of their physical properties, regardless of whether these properties are also said to instantiate any particular abstract mathematical properties. In the same vein, Bickle (2015) advances a conception of mechanistic explanation in which a characterization of a mechanism's physical properties suffices to explain a wide variety of behavioral and cognitive phenomena; whether or not a cognitive mechanism can also be characterized in abstract mathematical terms at the algorithmic level is explanatorily irrelevant.

There is reason to be wary of this conclusion, however. For one, it contradicts Marr's highly influential claim that all three levels of analysis are needed to "completely understand" a cognitive system's behavior. For another, it calls into question the explanatory relevance of cognitive modeling, a practice that is widely considered to be the centerpiece of cognitive scientific research (see statements to this effect by e.g. Busemeyer & Diederich 2010; Cummins 1983; 2000; Luce 1995; Stinson 2016; Weiskopf 2011). Unfortunately, several previous attempts to avoid this conclusion fall short. Consider, for example, Bechtel & Shagrir's recent claim that cognitive models facilitate the identification of *design principles* that "produce the same results across a wide range of different implementations" (Bechtel & Shagrir 2015: 318). Although Bechtel & Shagrir show that the description of design principles can render a particular mechanism's organization intelligible and generalizable (see also Levy & Bechtel 2013), they do little to argue that these principles actually *produce* anything over and above the physical properties in which they are instantiated. In other words, design principles are subject to Kim's (1993) *causal exclusion* argument, according to which realized properties possess no causal powers over and above their realizers. If the abstract design principles identified at the algorithmic level cannot be thought to possess causal powers over and above the concrete properties that instantiate them, it is unclear why

cognitive models are needed to answer questions about *how* a mechanism does what it does. A similarly unsuccessful response is given by Craver, who argues that the algorithmic level describes "realized properties [that] figure in unique causal relevance relations"—relations that, on his manipulationist account of constitutive relevance, "are true of realized properties and are not true of their realizers" (Craver 2007: 220). Recent challenges to Craver's manipulationist account suggest that interventions always affect realized and realizing properties simultaneously (Baumgartner & Gebharter 2015), thereby calling into question the claim that there can actually be any such unique causal relevance relations. Thus, the explanatory relevance of the algorithmic level remains uncertain.

These unsuccessful responses to the challenge of explanatory irrelevance are weighed down by the view of cognitive models as descriptions of a mechanism's abstract mathematical properties. But there is an alternative view that has yet to receive serious consideration in the literature: Rather than consider cognitive models as describing a subset of a mechanism's properties, they might instead be thought to describe a subset of its internal features.[6] Specifically, cognitive models can be thought to describe a mechanism's component operations as well as their functional organization, rather than its component parts and their structural organization. Indeed, insofar as cognitive models are developed by functionally analyzing the behavioral or cognitive capacity being explained, it should come as no surprise that they are concerned primarily or even exclusively with a mechanism's component operations. Piccinini & Craver appear to have this in mind when they observe that cognitive models are mechanism sketches "in which *some* structural aspects of a mechanistic explanation are omitted" (Piccinini & Craver 2011: 284, emphasis added). Nevertheless, they do not go far enough: Cognitive models may in fact be mechanism sketches in which *all* structural aspects are left out. For sure, many cognitive models contain constructs that are labeled with nouns such as 'filter', 'channel', or 'representation'. Nevertheless, these constructs are nearly always defined functionally: A filter is something that *filters*, a channel is something that *channels*, and a representation is something that *represents*.[7] Notably, on this view the fact that a cognitive model's constructs are specified in mathematical or computational terms is irrelevant to its ability to explain; in line with the 3M constraint discussed previously, the production rules, symbols, equations or variables that typically feature in such a model have just as much explanatory relevance as any other description—abstract or concrete—of the relevant mechanism's component operations. That is, it is misleading to align Marr's levels of analysis with levels of abstraction.

On this alternative view, "filling in" is a matter of *localizing* the production rules, symbols, equations or variables of a cognitive model by identifying them with a particular mechanism's

component parts (Bechtel & Richardson 1993). Thus, implementational-level questions about the way in which certain mathematical or computational constructs are "realized physically" are not questions about *how*, but are in fact questions about *where* a mechanism's component operations are carried out. A cognitive mechanism's component parts may be situated at several different levels of organization—from the level of molecules to the level of organisms (Bechtel 2008; Craver 2007)—and they may be highly distributed within any particular level—spanning whole neural populations, the brain as a whole, but also spanning the physical boundaries between brain, body and world (Kaplan 2012; Zednik 2011). Like a mechanism's component operations, its component parts can be described in abstract mathematical or concrete physical terms: Molecules, nerve cells, neural populations, bodily limbs and tools in the environment can all be described in full anatomical or physical detail, but may also be characterized schematically or abstractly, e.g. as lattice-like structures, geometric shapes, or graphical topologies. In general, no matter whether they emphasize abstract mathematical properties or concrete physical ones, descriptions of operations and their functional organization answer how-questions at the algorithmic level of analysis, while descriptions of parts and their structural organization answer where-questions at the implementational level.

In closing, it is worth highlighting an important corollary of this alternative view of the realization-relationship that obtains between the algorithmic and implementational levels: It is wrong to assume that cognitive models at the algorithmic level are *explanatorily autonomous* (compare Feest 2003; Weiskopf 2011). Investigators have recourse to a plethora of analytic and experimental techniques with which to characterize the component operations of a cognitive mechanism, as well as to test the accuracy of any particular characterization. For sure, some of these techniques are driven by purely behavioral methods, and may be independent of the possibility of localizing operations in a mechanism's component parts (Busemeyer & Diederich 2010; Cummins 1983). Nevertheless, the fact that cognitive models are subject to norms that do not involve localization has no bearing on the issue of whether localization—or more generally, the ability to identify the physical structures that are involved in the production of a cognitive or behavioral phenomenon—is *also* important (compare Stinson 2016). Moreover, it is important not to exaggerate the ability to describe a mechanism's component operations independently of its component parts. There are many historical examples in which the successful characterization of a mechanism's component operations was greatly facilitated or even enabled by the prior identification of its component parts (Bechtel 2008; Bechtel & Richardson 1993; Craver 2013). Indeed, as technological advances amplify investigators' ability to individuate structures in the brain and to characterize their functional properties, it seems reasonable to expect that how-questions at

the algorithmic level and where-questions at the implementational levels will become increasingly intertwined (see also Boone & Piccinini 2015).

5. Conclusion

The primary aim of this chapter has been to resolve ambiguities in Marr's account of explanation in cognitive science by subsuming it under the framework of mechanistic explanation. To this end, it was argued that Marr's three levels can be individuated by the different types of questions that are typically asked about a cognitive system, and that ambiguities concerning the way in which these questions are posed by Marr can be resolved by aligning each question with a specific aspect of mechanistic explanation. Computational-level questions about a system's computational goals are in fact what-questions that can be answered by describing a mechanism's behavior, and questions about these goals' appropriateness are in fact why-questions that can be answered by situating the mechanism in a containing environment. At the algorithmic level, questions about how a certain computational goal is achieved are in fact questions about a mechanism's component operations. Finally, at the implementational level, questions about the way in which algorithms and representations are physically realized are in fact where-questions that can be answered by identifying the relevant mechanism's component parts. Construed in this way, no single level of analysis bears sole responsibility for delivering mechanistic explanations of behavior and cognition; all three levels are involved in the discovery and description of cognitive mechanisms.

A secondary aim of this chapter has been to show that several different research programs—not just the computationalist approach in which Marr was himself embedded—seek to deliver answers to questions at all three levels of analysis. Thus, the scope of Marr's account is much wider than traditionally assumed, encompassing many different areas and traditions in contemporary cognitive scientific research. Insofar as researchers across cognitive science answer questions about the what, why, how and where of behavior and cognition, they are all in the business of mechanistic explanation.

Notes

References

Abrahamsen, A. & Bechtel, W. (2006). Phenomena and mechanisms: Putting the symbolic, connectionist, and dynamical systems debate in broader perspective. In R. Stainton (Ed.), *Contemporary debates in cognitive science*. Oxford: Basil Blackwell.

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences* 14: 471-517.

Baumgartner, M. & Gebharter, A. (2015). Constitutive Relevance, Mutual Manipulability, and Fat-Handedness. *The British Journal for the Philosophy of Science. doi: 10.1093/bjps/axv003.*

Bechtel, W. (2008). *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience. London: Routledge.*

Bechtel, W. (2009). Looking down, around, and up: Mechanistic explanation in psychology. *Philosophical Psychology* 22: 543-564.

Bechtel, W. & Abrahamsen (2010). Dynamic mechanistic explanation: Computational modeling of circadian rhythms as an exemplar for cognitive science. *Studies in History and Philosophy of Science Part A* 1: 321-333.

Bechtel, W. & Richardson, R. C. (1993). *Discovering complexity: Decomposition and localization as strategies in scientific research.* Princeton: Princeton University Press.

Bechtel, W. & Shagrir, O. (2015). The Non-Redundant Contributions of Marr's Three Levels of Analysis for Explaining Information Processing Mechanisms. *Topics in Cognitive Science* 7: 312-322.

Bickle, J. (2015). Marr and Reductionism. *Topics in Cognitive Science* 7: 299-311.

Boone, W. & Piccinini, G. (2015). The cognitive neuroscience revolution. *Synthese 193: 1509. doi:10.1007/s11229-015-0783-4.*

Bowers, J.S., & Davis, C.J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin* 138: 389–414.

Buckner, C. (2015). Functional kinds: A skeptical look. *Synthese* 192: 3915. doi:10.1007/s11229-014-0606-z.

Busemeyer, J.R. & Diederich, A. (2010). *Cognitive Modeling.* SAGE.

Busemeyer, J.R. & Townsend, J. (1993). Decision Field Theory: A Dynamic-Cogntive Approach to Decision Making in an Uncertain Environment. *Psychological Review* 100: 432-459.

Chemero, A. & Silberstein, M. (2008). After the Philosophy of Mind: Replacing Scholasticism with Science. *Philosophy of Science* 75: 1–27.

Craver, C.F. (2007). *Explaining the Brain*. Oxford: Oxford University Press.

Craver, C.F. (2013). Functions and Mechanisms: A Perspectivalist View. In P. Huneman (ed.), *Functions: Selection and Mechanisms* (pp. 133–158). Dordrecht: Springer.

Cummins, R. (1983). *The Nature of Psychological Explanation*. Cambridge: MIT press.

Cummins, R. (2000). 'How Does It Work?' versus 'What Are the Laws?': Two Conceptions of Psychological Explanation. In F. Keil and R. Wilson (eds.), *Explanation and Cognition* (pp. 117–44). Cambridge, MA: MIT Press.

Danks, D. (2008). Rational analyses, instrumentalism, and implementation. In N. Chater & M. Oaksford (eds.), *The Probabilistic Mind: Prospects for Bayesian Cognitive Science* (pp. 59-75). Oxford: Oxford University Press.

Darden, L. & Craver, C.F. (2012). *In Search of Mechanisms.* Chicago, IL: The University of Chicago Press.

Feest, U. (2003). Functional Analysis and the Autonomy of Psychology. *Philosophy of Science* 70: 937-948.

Fodor, J.A. (1980). Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and Brain Sciences* 3: 63-73.

Glennan, S. (2010). Mechanisms, Causes, and the Layered Model of the World. *Philosophy and Phenomenological Research* LXXXI: 362-381.

Hummel, J.E. & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review* 99: 480-517.

Kaplan, D.M. (2011). Explanation and Description in Computational Neuroscience. *Synthese* 183: 339-373.

Kaplan, D.M. (2012). How to Demarcate the Boundaries of Cognition. *Biology and Philosophy* 27: 545-570.

Kaplan, D.M. & Craver, C.F. (2011). The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective. *Philosophy of Science* 78: 601–627.

Kelso, J.A.S. (1995). *Dynamic Patterns: The Self-Organization of Brain and Behavior.* Cambridge, MA: MIT Press.

Kim, J. (1993). *Supervenience and Mind: Selected Philosophical Essays.* Cambridge: Cambridge University Press.

Levy, A. & Bechtel, W. (2013). Abstraction and the organization of mechanisms. *Philosophy of Science* 80: 241-261.

Luce, R.D. (1995). Four tensions concerning mathematical modeling in psychology. *Annual Review of Psychology* 46: 1-26.

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

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information*. San Francisco: W. H. Freeman.

McClamrock, R. (1991). Marr's three levels: A re-evaluation. *Minds and Machines* 1: 185- 196.

Milkowski, M. (2013). *Explaining the Computational Mind.* Cambridge, MA: MIT Press.

Nosofsky, R.M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology* 115: 39-57.

Oaksford, M. & Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford: Oxford University Press.

Piccinini, G., & Craver, C.F. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese* 183: 283-311.

Polger, T. (2004). Neural machinery and realization. *Philosophy of Science* 71: 997-1006.

Pouget, A., Beck, J.M., Ma, W.J., & Latham, P.E. (2013). Probabilistic brains: Knowns and unknowns. *Nature Neuroscience* 16: 1170-1178.

Pylyshyn, Z. (1980). Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences* 3: 111-169.

Ramsey, W.M. (2007). *Representation Reconsidered.* Cambridge: Cambridge University Press.

Rumelhart, D.E., McClelland, J.E. & the PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* Cambridge, MA: MIT Press.

Shagrir, O. (2010). Marr on computational-level theories. *Philosophy of Science* 77: 477- 500.

Stinson, C. (2016). Mechanisms in psychology: Ripping nature at its seams. *Synthese* 193: 1585. doi:10.1007/s11229-015-0871-5.

Strevens, M. (2008). *Depth*. Cambridge, MA: Harvard University Press.

van Gelder, T.J. (1995). What Might Cognition Be, If Not Computation? *Journal of Philosophy* 91: 345–81.

van Gelder, T.J. (1998). The Dynamical Hypothesis in Cognitive Science. *Behavioral and Brain Sciences* 21: 1–14.

Weiskopf, D.A. (2011). Models and mechanisms in psychological explanation. *Synthese* 183: 313. doi:10.1007/s11229-011-9958-9

Zednik, C. (2011). The nature of dynamical explanation. *Philosophy of Science* 78: 238- 263.

Zednik, C. & Jäkel, F. (2016). Bayesian reverse-engineering considered as a research strategy for cognitive science. *Synthese*. doi:10.1007/s11229-016-1180-3.

1 The author is indebted to Frank Jäkel, Holger Lyre, the volume editors, and conference audiences in Düsseldorf and Warsaw for helpful feedback on earlier versions of this chapter.

2 Whereas Marr sometimes speaks of explanations at individual levels of analysis, here the term 'explanation' will be reserved for a full three-level account, i.e. the kind of account that Marr deems necessary for "complete understanding" (Marr 1982: 4ff). This is not meant to be a commitment to any particular account of the relationship between explanation and understanding, however.

3 Levels of analysis must not be confused with levels of organization within a mechanism (see also Bechtel 2008; Craver 2007). Whereas the former are individuated by the kinds of questions an investigator might ask about a cognitive system, the latter are individuated by constitution-relations within a mechanism. Notably, insofar as many real-world cognitive mechanisms are hierarchical (Bechtel & Richardson 1993; Craver 2007), it may often be profitable to apply all three levels of analysis at several different levels of organization within a single mechanism.

4 Indeed, some proponents of the Bayesian approach take themselves to be delivering teleological explanations *rather than* mechanistic explanations (see e.g. Oaksford & Chater 2007). The accuracy of this characterization has been questioned, however, with some claiming that the Bayesian approach does not provide explanations at all (Bowers & Davis 2012; Danks 2008), and others arguing that it does in fact explain in a way that centers on the discovery of mechanisms (Zednik & Jäkel 2016).

5 The present discussion concerns only the realization-relationship that is relevant to Marr's account of explanation in cognitive science. No commitment will here be made regarding the realization-relationship in other contexts (see e.g. Kim 1993). Moreover, the present contribution will not explore the question of whether the realization-relationship that obtains between the implementational and algorithmic levels also obtains between the algorithmic and computational levels.

6 Intriguingly, Glennan (2010) has called it a "category mistake" to attribute a mechanism's causal powers to its properties, instead of to its component parts, operations and/or overall organization. The view advanced here of the relationship that obtains between the algorithmic and implementational levels is consistent with Glennan's.

7 The view outlined here is consistent with the proposal that what a particular representation represents may be secondary to the causal-functional role it plays in a cognitive system (see e.g. Fodor 1980). A particular representation's causal-functional role is akin to a mechanistic operation, rather than e.g. a part.