Synthese (2021) 199 (Suppl 1):S19–S41
https://doi.org/10.1007/s11229-020-02568-5

SI: EXPLANATIONS IN COGNITIVE SCIENCE: UNIFICATION VS PLURALISM

The methodological role of mechanistic-computational models in cognitive science

Jens Harbecke[1]

Received: 2 November 2018 / Accepted: 6 February 2020 / Published online: 17 February 2020

Abstract

This paper discusses the relevance of models for cognitive science that integrate mechanistic and computational aspects. Its main hypothesis is that a model of a cognitive system is satisfactory and explanatory to the extent that it bridges phenomena at multiple mechanistic levels, such that at least several of these mechanistic levels are shown to implement computational processes. The relevant parts of the computation must be mapped onto distinguishable entities and activities of the mechanism. The ideal is contrasted with two other accounts of modeling in cognitive science. The first has been presented by David Marr in combination with a distinction of "levels of computation". The second builds on a hierarchy of "mechanistic levels" in the sense of Carl Craver. It is argued that neither of the two accounts secures satisfactory explanations of cognitive systems. The *mechanistic-computational* ideal can be thought of as resulting from a fusion of Marr's and Craver's ideals. It is defended as adequate and plausible in light of scientific practice, and certain metaphysical background assumptions are discussed.

Keywords Cognitive science · Explanations · Mechanistic approach · Computational explanations · Levels of explanation

This work has been supported by the Deutsche Forschungsgemeinschaft (DFG), reference number HA 6349/5-1, Project Number 413568662; project: "Model-Development in Neuroscience: Simplicity and Generalizability in Mechanistic Explanations".

✉ Jens Harbecke
jens.harbecke@uni-wh.de
https://www.jensharbecke.com

[1] Witten/Herdecke University, Witten, Germany

1 Introduction

Researchers working in cognitive neuroscience have sometimes expressed different views on which research strategies optimize explanation, prediction, and understanding given limited research resources. Some scientists have championed the integrated reconstruction and simulation of detailed neural mechanisms for the explanation, understanding, and prediction of behavior on the system level. Others have emphasized the importance of the principles and algorithms of information-processing by behavioral modules which are then used as constraints on the search for the relevant neural mechanisms.

Some of these views have surfaced, for example, in exchanges over the Human Brain Project (HBP) and computational approaches to brain research. One of the declared aims of the HBP is a reconstruction of the human brain, including its electrical and chemical functions. The co-director of the HBP, Henry Markram, has expressed hopes "(...) to learn a great deal about brain function and dysfunction from accurate models of the brain. (...) There is no fundamental obstacle to modeling the brain and it is therefore likely that we will have detailed models of mammalian brains, including that of man, in the near future." (2006, p. 158) The view expressed here is that modeling brain structures allows for the simulation of cognitive functions, which in turn provides explanations of actual cognitive functions implemented by the human brain.

As a critical response to this approach, renown neurocomputationalist Haim Sompolinsky has emphasized "(...) the crucial relevance of theory in providing the necessary analytical and synthetic mathematical tools to interpret the experimental data and computational simulations, particularly as they are applied to natural behavior." (Shepherd et al. 2011, p. 15) Sompolinsky considers theoretical models as crucial for the interpretation of simulations, which otherwise remain uninformative and non-explanatory. Genuine explanations are primarily supplied by an explication of the information processes realized by the brain, and secondarily by a description of the brain's physiological structures and mechanisms.[1]

In other words, whereas the explanatory strategy favored by Markram at the time considers the identification and simulation of the brain mechanisms underpinning different cognitive phenomena of primal importance, Sompolinsky sees the primary key for explaining cognitive phenomena in an analysis of the computational processes that are at the heart of these phenomena.

The different views that have surfaced at times in methodological exchanges among neuroscientists about the HBP and other research projects can be interpreted as implicitly, and perhaps unintentionally, reflecting a longstanding debate on the norms for explanation in the philosophy of cognitive science. It is not plausible that this debate within the world of philosophy is actually known to practicing neuroscientists. Nevertheless the correspondence may not be entirely accidental as there may be a real question to be answered here about how modeling and explaining of the brain and its cognitive capacities should look like.

[1] Sompolinsky has restated his pessimistic assessment of the Human Brain Project on various occasions; cf. also his statements quoted by the New York Times on 18 March, 2013 ("Bringing a Virtual Brain to Life" by Tim Requarth).

In order to grasp the essence of this philosophical debate that has seen a large number of contributions especially over the last 30 years, it is instructive to look at two ideals of modeling and explanation that have played a prominent role in the discussion. The first is represented, in particular, by David Marr (1982) who has emphasized the importance of "levels of computation" in an explanation of cognitive phenomena. The second is reflected in the work of Peter Machamer, Lindley Darden, and Carl Craver (2000) and in subsequent publications by Craver and Darden (2001) and Craver (2001, 2007).[2] It stresses a distinction of "levels of mechanism" as they appear in models of experimental neurobiology and neuropsychology—disciplines that often use an explanatory strategy based on an integrated reconstruction and simulation of detailed neural mechanisms.

In this paper, I am concerned with a specific question within this philosophical debate, which may be relevant for practice in contemporary cognitive neuroscience as well, given the implicit analogy between the methodological exchanges hinted at above holds. More concretely, I am concerned with the general question of explanatory adequacy for cognitive systems and the relationship between the two philosophical ideals of modeling and explanation in cognitive neuroscience mentioned above. As part of my argument, I show that both explanatory ideals often associated with Machamer et al. (2000) and Marr (1982) have important limitations. Neither corresponds to what should ultimately be considered a comprehensive and satisfactory model and/or explanation of a phenomena in cognitive neuroscience.

Building on this conclusion, my main claim in this paper is that a more adequate normative ideal results from a fusion of the two modeling ideals. Such a "mechanistic-computational" or "MC-account of modeling and explanation" emphasizes that a satisfactory explanation of the human brain as a cognitive system will ultimately have to satisfy a fairly comprehensive set of explanatory norms. More specifically, adequate explanations of cognitive phenomena in the human brain will typically have to involve both, a specification and modeling of the various mechanistic levels within the brain as targeted by neuroscience, as well as a description of the computational principles and relations that these levels implement. Neither Marr and Machamer et al. might have ultimately disagreed with this overarching claim. However, as will be shown below, the synthesized explanatory ideal goes substantially beyond what these and other authors have explicitly defended within the existing literature, and it is actually novel in this sense.

The paper proceeds as follows. I begin with a general characterization of the "MC-account of modeling and explanation" (Sect. 2), and I offer some hints as to why this view differs from other prominent contributions and is distinctive and novel. Subsequently, I present the central features of Marr's and Craver's normative frameworks (Sects. 3 and 4), which I interpret as roughly corresponding to the different methodological views that sometimes surface in cognitive neuroscience as mentioned above. I then discuss some limitations of the frameworks, and I investigate into their conceptual connections, which naturally leads into a defense of the MC-ideal of modeling and explanation (Sect. 5). In a short digression, I discuss the potential redundancy of computational explanations vis-à-vis mechanistic explanations (Sect. 6). The final

[2] Some of its precursors are found in Bechtel and Richardson (1993), Glennan (1996), and Skipper (1999).

section summarizes the results and lists some further questions relevant for future research on the topic (Sect. 7).

2 Mechanistic-computational explanations

The normative ideal of modeling that I defend in this paper can be characterized as a "mechanistic-computational" or "MC-theory of modeling". It claims that a satisfactory neuroscientific explanation of a cognitive system "must bridge phenomena at multiple [mechanistic] levels" (Craver 2007, p. 10), such that at least several of these mechanistic "hardware and implementation levels" are analyzed in their information-processing dimensions in the form of a "computational level" and a "representational and algorithmic level" (Marr 1982, pp. 24–25). The background assumption is that explaining a cognitive system such as the human brain essentially requires saying something about (i) what information processes it implements and (ii) what its physiological structure is. Additionally, (iii) it will have to provide details on which aspects of the physiological structure realize which part of the implemented information processes. Figure 1 is an attempt to illustrate the general idea.

The vertical plane in Fig. 1 is intended to represent (albeit in a highly abstract way) the mechanistic hierarchy that underlies certain cognitive functions of cognitive systems such as the human brain. For simplicity, the mechanistic hierarchy in the illustrated case involves only three mechanistic levels: mechanistic level n, level $n-1$ and (implicitly) level $n-2$. The notion of a "mechanistic level" applied here is roughly the one characterized by Machamer et al. (2000), Craver and Darden (2001), Craver (2001, 2007) and others as the physical processes involving physical components and their natural activities that relate a set of initial conditions to termination conditions at a particular grain.

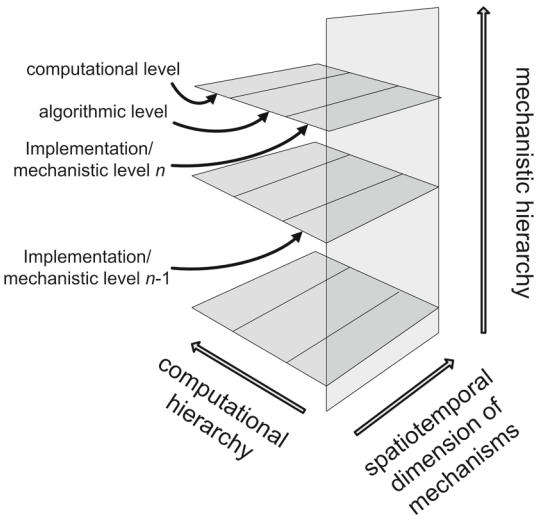

Fig. 1 The MC-ideal for explanations of cognitive neuroscience

The diagonal dimension of this vertical plane symbolizes the "spatio-temporal dimension" of mechanisms (as indicated in the figure). According to the standard definition (cf. Machamer et al. 2000, p. 3), mechanisms involve a time span from an initial state to a termination state. The diagonal arrow running parallel to the plane emphasizes this temporal evolution of the various mechanisms within the hierarchy from initial states to termination states. So in short, according to the MC-theory, adequate and explanatory models of cognitive systems as the fictional one depicted by Fig. 1 essentially comprise an explication of the physical mechanisms underpinning the phenomenon in question at several grains.

To make the idea of the abstract plane representing a mechanistic hierarchy with several levels and a spatio-temporal dimension concrete, consider the recent neuroscientific research by Wallis (2012), Payzan-LeNestour et al. (2013), Polanía et al. (2014), Gluth et al. (2017) and others on the mechanisms underlying value-based decision making in humans including boundary phenomena such as the attraction effect. As these researchers have shown, the computational process of value-based decision making as applied by the human brain in large set of different environments is physically realized in several specialized brain regions including the orbitofrontal cortex, the nucleus accumbens and the amygdala ("level n"). Frank and Claus (2006) have indicated that the phenomenon of value-based decision making is in principle analyzable also at the level of cell networks and their interaction in time ("level $n-1$"), or even at the level of single cell interactions ("level $n-2$"). The overall explanatory project that emerges from these studies conforms well with the MC-account and its demand represented by the vertical plane in Fig. 1 that at least some mechanistic levels underpinning a given cognitive phenomenon must be explicated for the explanation to be adequate and satisfactory (cf. Sect. 4 below).

Importantly, however, the MC-model goes beyond this demand for an explication of the mechanisms constituting the phenomena. The MC-theory makes the additional normative demand that a computational analysis must be developed as well for several levels in the mechanistic hierarchy. This further demand is represented (again, in a highly abstract way) by the three planes "docking onto" the three mechanistic levels n, $n-1$, $n-2$ in Figure. As made explicit in the illustration, the lines running over the three planes represent a "computational level" and a "algorithmic level", which are systematically mapped onto the mechanistic or "implementation" levels. This additional demand captures the core of a specific theory of computational explanations that has been prominently defended by Marr (1982) (cf. Sect. 3 below).

The resulting bookshelf structure is essentially the explanatory norm that the MC-account promotes. Its background claim is that adequate and explanatory models of cognitive systems as the hypothetical one depicted by Fig. 1 essentially comprise an explication of the physical mechanisms underpinning the phenomenon in question at several grains as well as an explication of the computational processes and algorithms that the system realizes at each one of these mechanistic levels.

As mentioned in Sect. 1, my main claim in this paper is that the MC-account is the adequate norm for explanations of cognitive and intelligent systems. As minor claims, I contend that there remains an important distinction between mechanistic and computational explanations, and that the latter cannot be reduced to the former, nor are the latter kinds of explanations mere "mechanism sketches".

The MC-view converges with positions defended by Bechtel (1994, 2009), Bechtel and Shagrir (2015), Shagrir and Bechtel (2017), and Harbecke and Shagrir (2019). These authors have emphasized the extrinsic and contextual dimension of computational explanations, which is a different kind of contextuality as, for instance, the causal contextuality of air pressure outside of the brain that influences the internal physiological mechanisms and processes of the brain.

At the same time, the MC-view diverges from authors such as Bickle (2015), who defends a causal-mechanical ideal of explanation in cognitive science in which computational models and analyses play a secondary role only. The MC-view also diverges from authors such as Milkowski (2013a, b) and Piccinini (2006, 2007a, b), who acknowledge the importance of computational modeling in cognitive science, but tend towards the view that computational models are mechanism sketches awaiting to be filled with mechanistic detail as scientific discovery continues.

Where the MC-model goes beyond what has been explicitly defended by all of these authors is its far-reaching demand for satisfactory and complete[3] explanations in cognitive neuroscience. The claim that computational and mechanistic explanations are essentially different kinds of explanation, and that yet both are required for satisfactory explanations in cognitive science in a multi-level coordinated fashion is a novel position not found in the literature so far. The following sections develop in detail the adequacy of the MC-account of explanation.

3 Levels of computation

In his book on vision, neuroscientist David Marr develops a normative ideal of explanation that is based on the contention that there are "[t]hree different levels at which an information processing device must be understood before it is understood completely." (1982, p. 24) In his view, the required levels of understanding are the "computational level", the "representational and algorithmic level", and the "hardware and implementation level" (*op. cit.*, 25).[4] The computational specification provides answers to questions, such as "*What* is the goal of the computation?", "*What* function is being computed?", "*Why* is it appropriate for the completion of the cognitive task?" (1982, p. 22) The algorithmic level provides answers to questions such as: "*How* does the system compute its function?", "What are the inputs and outputs?" (1982, p. 23) Finally,

[3] It should be noted that completeness as a virtue of explanation is not uncontroversial. Some authors have argued that an explanation that abstracts or idealizes is preferable to one that is complete. Without being able to dive deeper into this controversy due to limits of space, for the present paper I presuppose that the virtue of completeness remains an important criterion in the assessment of scientific modeling and explanation. However, if one takes particular issue with the term, perhaps the notion of "satisfaction" is sufficient to characterize the way that MC-models supersede purely mechanistic or computational models.

[4] The idea that the mind can be viewed as an information processing system and that computational descriptions are the right tools for explaining it had entered mainstream cognitive science and parts of theoretical neuroscience long before Marr. Warren McCulloch and Walter Pitts (1943) first formulated the view that is now usually referred to as "computationalism", namely the contention that cognitive capacities of the human brain can be explained by computational processes. Their speculations were crucially influenced by the discovery that neural impulses are all-or-none affairs. Authors such as Fodor (1975), Miller et al. (1960), and Newell et al. (1972) later argued that computational explanation could be developed by psychology in relative autonomy from the neurosciences (cf. Piccinini 2006).

the implementation level concerns the question how the information and the algorithm can be realized physically.

Marr's well-known theory of edge detection in the human visual system is often quoted to illustrate these distinctions. According to this proposal, the function of the retina can be described as a conversion of the retinal image consisting of an array of points representing light intensities into a "raw primal sketch", in which point neighborhoods corresponding to object edges are value 1 and all other points are value 0. The conversion is characterizable by the function $f(x, y) = \nabla^2 \mathrm{G} * I(x, y)$, where "$I$" refers to the array of light intensities, "$*$" is the convolution operator, and "$\nabla^2 \mathrm{G}$" the filtering operator consisting of a Gaussian and a Laplacian. In Marr's view, implementing this function "is (...) what the retina does." (1982, p. 337). The reason *why* it implements this function is the fact that, in normal environments of humans, sudden changes in light intensities correspond to edges of objects.[5]

The computation performed by the retina is realizable by several different algorithms listing the steps in which the operators are applied (cf. Sharifi et al. 2002). Finally, and despite the fact that Marr warns against an exclusive focus on the neural activity when developing the explanation, the complete theory of vision specifies also the neural circuits and processes in the eye, the lateral geniculate nucleus (LGN), and the visual cortex, through which the algorithm is implemented.

One feature of this general theory of vision is that some of its levels of description involve reference to mathematical entities such as functions and algorithms (the "what?"-element). A computational function such as $f(x, y) = \nabla^2 \mathrm{G} * I(x, y)$ "receives" values assigned to points on a coordinates plane as inputs and "emits" values assigned to points on a coordinates plane as outputs. Other levels directly refer to concrete entities such as neural processes in actual cognitive systems consisting of a collection of entities and activities in space and time. This difference between mathematical and physical objects gives a first hint that Marr's levels fall under different ontological kinds.[6]

A second feature concerns the relationship that Marr's levels bear to one another. Marr expresses his view as follows:

> [T]here is usually a wide choice of representation [= syntax for a given physical computational system]. [The] same algorithm may be implemented in quite different technologies. (...) [The] three levels are coupled, but only loosely. The choice of an algorithm is influenced (...) by the hardware in which it must run. But there is a wide choice available at each level, and the explication at each level involves issues that are rather independent of the other two. (Marr 1982, pp. 23–25)

[5] In other words, which of the syntantic structures implemented by the retina is its actual computation is partially determined by the environment. For the role of the environment for computational explanations, cf. Harbecke and Shagrir (2019).

[6] The interpretation of mathematical expressions as referring to abstract or fictional objects reflects mainstream metaphysics of mathematics. A different view might suggest that the referents of mathematical expressions are in fact physical objects or states of affairs. On such an account, all of Marr's level descriptions refer to concrete entities.

It is not immediately apparent what Marr means with the "wide choice", the "looseness" and "independence" of the computational, algorithmic, and implementation levels. Some authors have taken these expressions to imply that Marrian levels are discrete and fail to belong to a hierarchy defined by any clear relationship (cf. Dawson 1998).[7] Hence, it may be misleading to talk about "levels" in this context in the first place. There have also been proposals to add more levels to the Marrian hierarchy that might glue the computational and the algorithmic level closer together (cf. O'Hara 1994; Peacocke 1986).

Whilst the diagnosis of a certain metaphysical discreteness of Marr's levels may actually be correct (cf. also Sect. 6), the discreteness as such does not exclude a systematic ordering of the elements included in the levels. A weaker interpretation sees Marr as rejecting the claim the the implementation level "composes" the algorithmic and the computational level, but as suggesting a many-many relationship that still allows for order: The same kind of hardware can run different algorithms and implement different functions, the same algorithm can run on different kinds of hardware, and the same computational function can be solved by different algorithms.

That such a many-many structure is in fact what Marr had in mind has been shown by Shagrir (2010b) with reference to Marr's statements that "(...) the most abstract is the level of what the device does and *why*." (Marr 1982, p. 22; emphasis added). The "why"-element of computational descriptions essentially refers to the task that the system solves within a contingent environment. Such a task is not exhaustively characterizable by the system's causal and mechanistic structure and its immediate inputs from, and outputs to, the environment. Rather, it refers to a more general ecology within which the system acts. The computations are still performed by and "within" the system. However, which of the many syntactic structures, also called "automata", that are simultaneously implemented by the system form(s) the actual computation(s) performed by the system depends on the tasks that a system actually solves within its environment.

For instance, the fact that the colocated zero crossings of $\nabla^2 \mathrm{G} * I$ detect edges is due to the fact that, in our world, sudden changes in light intensities are usually found at the edges of objects. If seated in a radically different environment − perhaps one in which changes in light intensities are entirely random −, the retina may no longer compute the function $f(x, y) = \nabla^2 \mathrm{G} * I(x, y)$. That is, it would be fundamentally misleading to characterize it as performing such a computation. Computing colocated zero crossings of $\nabla^2 \mathrm{G} * I$ would solve no task relevant to the retina and the brain in such a surrounding. At the same time, the retina may compute a different function there since, as it happens, one of its many syntactic structures solves a new task now relevant for the system. Shagrir summarizes his interpretation as follows (cf. also the discussion in Sect. 6):

> As Marr says, it does not immediately follow that if the term $I(x, y)$ refers to the array of light intensities in the visual field, then the colocated zero crossings of the different-scale filtering of $\nabla^2 \mathrm{G} * I$ stand for physical edges. That they do is a contingent fact about our visual environment. (Shagrir 2010b, p. 492)

[7] I thank the editors for bringing this point to my attention.

If this interpretation is correct, then in Marr's view the physical processes and mechanisms of the cognitive system are not sufficient for the computed function, even if the physical hardware strongly constrains ("influences") the set of functions potentially computed by the system.[8] Only when the broader environment is taken into account, the system's computations are fully determined.

4 Levels of mechanism

In his contributions to the mechanistic debate on the norms of research in experimental neuroscience, Carl Craver has emphasized that "(...) an adequate explanation of many phenomena in the central nervous system must bridge phenomena at multiple [mechanistic] levels." (Craver 2007, p. 10).[9] More concretely, in neuroscience "(i) explanations describe mechanisms; (ii) explanations span multiple levels; and (iii) explanations integrate findings from multiple fields." (Craver 2007, p. 2). Mechanisms, in this context, are taken to be "(...) entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions." (Machamer et al. 2000, p. 3).

As an example, the cognitive phenomenon of spatial memory generation in rats (A) is explained, roughly, by certain activities of neural assemblies in hippocampal area CA1 (B), which in turn is explained by a long-term potentiation of synapses of pyramidal cells in this region (C), which in turn is explained by an increase of calcium ion influx into postsynaptic dendrites after activation of *N*-methyl-D-aspartat (NMDA) receptors gating Ca^{2+} (D).

Phenomena and processes A to D are interpreted as occupying different "mechanistic levels", which Craver takes to be real aspects of the world (cf. Craver 2007, p. 177). In other words, the descriptions specifying the various mechanistic processes on different levels are taken to refer to concrete entities and activities. The entities and activities themselves are local features of a given system and its causal surroundings, i.e. they cannot be changed merely by relocating the system into a different environment in which its physical inputs and outputs remain the same but where the overall ecological structure is radically different.[10]

[8] The constraining assumption shows that Marr would probably have rejected the claim that any physical system can implement any algorithm whatsoever (cf. Putnam 1988; Searle 1992). Such a belief can be upheld only if one sets no constraints on which mappings of sequences of states individuated by a computational description onto sequences of states individuated by a physical description of the system are acceptable (cf. Piccinini 2007a, b).

[9] I focus on Craver here because of the leading mechanists, he has probably formulated the most well-known version of the mechanistic ideal of modeling as captured in this and the following quotes. It seems to me that he thereby reflects well the implicit approach of many researchers in the field of neuroscience. One should however say that other mechanistic philosophers disagree with his basic picture and his ideals. See, for instance, the debate on constitutive inference as represented by Harbecke (2015, 2019), Baumgartner and Gebharter (2016), Baumgartner and Casini (2017), and Gebharter (2017).

[10] It should be mentioned that Craver (2007, p. 141) as well as other mechanists such as Bechtel (2009) and Piccinini (2008a) have pointed out that mechanisms often "transgress compartmental boundaries" or even the boundaries of an organism. However, this has no bearing on the local character of a mechanism in Craver's sense in the sense that any change of the mechanism will have to be brought about by some causal

Craver uses a modified version of the interventionist definition of causation in order to explicate the relation of constitution that the various mechanisms on different levels bear to another. A mechanism, e.g. x's ϕ-ing is constitutively relevant for a phenomenon, e.g.. y's ψ-ing, if, and only if:

"(i) [x] is part of [y];
(ii) in the conditions relevant to the request for explanation there is some change to [x]'s ϕ-ing that changes [y]'s ψ-ing; and
(iii) in the conditions relevant to the request for explanation there is some change to [y]'s ψ-ing that changes [x]'s ϕ-ing." (Craver 2007, p. 153)

Condition (i) is based on the relation of mereological parthood, which underlines the assumption that Craver's levels are concrete and in space and time. The locality and concreteness of levels suggest that all levels of mechanism belong to the same ontological kind.[11]

A different question concerns the relationship of higher and lower mechanistic levels. Craver suggests the following about the manipulability of S's ψ-ing and X's ϕ-ing:

> Note that this is not a supervenience or identity claim. (...) S's ψ-ing does not supervene on X's ϕ-ing. Rather, it supervenes on the organized activities of all of the components in the mechanism. (Craver 2008, p. 15; cf. also Craver 2007, fn. 26)

In other places, Craver also says that "[l]evels of mechanisms (...) are a variety of partwhole relation." (Craver 2007, p. 165) What both supervenience relations[12] and part-whole relations among levels have in common (at least under **GEM**; see Varzi 2019) is that the lower levels are sufficient for the higher levels. Or in other words, if a mechanism that constitutes a phenomenon P is present, then it is inevitable that P is present as well. Or, once you fix the lower level sufficiently broadly understood, the higher level is fixed, too.

To conclude, in Craver's view, satisfactory explanations of phenomena in neuroscience describe in detail the mechanisms that constitute the phenomenon. Mechanisms are local to the system and to its causal inputs and outputs, and they inhabit different levels. Levels are real features of the world, and lower levels are systematically sufficient for higher levels.

Footnote 10 continued
influence. A relocation of the mechanism into a different environment that maintains the same inputs and outputs of the mechanism does not change the mechanism as mechanism.

[11] An alternative view of mechanistic constitution has been proposed by myself (cf. 2010) and Couch (2011), which is based on a regularity analysis of the constitution relation between mechanisms and phenomena.

[12] It should be noted that other mechanists such as Bechtel (2009) and Milkowski et al. (2018) have actually rejected a supervenience relation among levels. However, this issue is not entirely resolved among proponents of the mechanistic approach. As I have shown elsewhere (cf. 2014), for instance, regularity constitution between particular mechanistic types is logically independent from any strong or weak supervenience claim about the levels. Nevertheless, higher levels as a whole (and understood as sets of mechanisms) can supervene on lower levels.

5 Limits, incomparability, and identity

I now argue that the two modeling ideals promoted by Marr and Craver (cf. Sects. 3 and 4) are neither incommensurate nor the same as it has sometimes been suspected in the literature. Or in other words, they are not distinct, but also not identical. I then demonstrate that the two modeling ideals by themselves do not secure fully satisfactory and complete explanations of cognitive systems. These two arguments support the basis for the formulation and justification of the MC-theory of explanation as it has already been presented in Sect. 2 and as it is discussed further in Sect. 6.

A presumed distinctness of the modeling ideals defended by Marr and Craver amounts to the claim that the two level hierarchies widely cross-classify systems and phenomena, and that the norms associated with the models do not match up in any systematic way. This was perhaps the view of Churchland and Sejnowski when they insinuated that "(...) when we measure Marr's three levels of analysis against levels of organization [or mechanism] in the nervous system, the fit is poor an confusing at best" (1992, p. 19).

The disunity in the application of terms between the two normative ideals by Marr and Craver may be counted as evidence for this interpretation. In particular, it is not obvious that Craver (2007) makes a clear-cut distinction between the notions of a "mechanism", a "representation", and a "computation". Instead, he sometimes seems to interpret computational descriptions as (preliminary versions of) mechanistic descriptions.[13] Marr, in contrast, emphasizes the relative autonomy of the computational level and the algorithmic from the implementation level,[14] and he stresses that the discovery of neural correlates to cognitive functions has contributed little to an actual understanding of cognition.[15] These points may suggest a distinctness of the level hierarchies and the normative theories associated with them.

[13] The following quote from Craver's reconstruction of the received explanation of spatial memory in rats suggests that, in his view, computational properties are *part* of the mechanistic hierarchy: "*At a lower level*—the level of spatial map formation—*are the computational properties* of neural systems, including brain regions such as the hippocampus and other areas in the temporal and frontal cortex. (2007, p. 167; *emphasis added*). Another example is a figure on p. 257 of his (2007), which characterizes the top-layer of the mechanistic hierarchy as the "higher-level mechanism: computation in the hippocampus". The joint paper with Piccinini (2011) states a closely related view, according to which computational descriptions are a different, perhaps more abstract kind of mechanistic description: "[F]unctional analyses [− of which computational analyses are a specific kind −] are *sketches of mechanisms*, in which some structural aspects of a mechanistic explanation are omitted. Once the missing aspects are filled in, a functional analysis turns into a full-blown mechanistic explanation. By this process, functional analyses are seamlessly integrated with multilevel mechanistic explanations." (284) In short, in some of his works, Craver has characterized computational levels as mechanistic levels, and he has sometimes considered computational descriptions as mere functional analyses having little to no explanatory value.

[14] The belief in autonomy is present when Marr emphasizes the "loose connection" and "independence" of the levels and the "wide choice" that one has available when developing them (cf. Sect. 3, and the quotes from Marr 1982, pp. 23–25). As before and in line with Shagrir (2010b), I interpret these statements as stipulating a many-many relationship between the levels.

[15] As Marr reconstructs the case of vision, the increasing number of electrophysiological studies of the 1970s did not provide an understanding of the function of the visual cortex: "[F]or a long time, the best hope seemed to lie along another line of investigation, that of electrophysiology. (Marr 1982, p. 11) (...) But somewhere underneath, something was going wrong (...) No neurophysiologist had recored new and clear high-level correlates of perception. (14) (...) None of the new studies succeeded in elucidating the

On the other hand, the examples of neuroscientific theorizing briefly reviewed in Sects. 3 and 4 offer some reasons to believe that the hierarchies have fruitful applications in explanations of the same systems and behaviors. After all, both Marr and Craver are primarily interested in explaining the functions and dysfunctions of the human central nervous system. Moreover, the two hierarchies have at least one clear touch point, namely the "implementation states" referred to by Marr's level 1 and Craver's levels of mechanism. According to Marr, a satisfactory explanation of a cognitive capacity essentially involves an understanding of how the capacity or function is implemented in the processes at the neural level. This suggests that, even though the terminology of Marr's and Craver's level hierarchies is substantially different, the hierarchies overlap conceptually. More specifically, both theories involve descriptions of neural processes, or "mechanisms", on at least one level.

On the opposite side of the conceptual spectrum, an identity assumption about the two normative approaches would amount to the claim that, despite initial appearance, the descriptive terms within the hierarchies simply corefer in the prototypical explanations of edge detection and other cognitive phenomena. They describe the same structures through different linguistic forms. Or in other words, they are notational variants expressing roughly the same propositions and truth conditions. Consequently, the sets of norms associated with the hierarchies overlap substantially as well. They simply demand an explication of the mechanisms constituting a phenomenon at different levels. This view is in line with an interpretation of computational descriptions as "mechanism sketches". On this particular interpretation, the direct referents of computational descriptions are neural processes, activities, and concrete objects. They describe the latter as mechanistic descriptions do, though in less detail (cf. also Piccinini and Craver 2011; Kaplan and Craver 2011 and Sect. 6 below).

From the features summarized in Sect. 3, it is clear that Marr was skeptical about a classification of the computational level as another mechanistic level. In his view, the implementation level describing the physical mechanisms of a cognitive system is not sufficient for a full computational explanation that answers "why?"-questions on top of "what?"-questions (cf. also fn. 15 above). Craver, in contrast, points out that "lower levels" understood as sets of mechanisms are sufficient for "higher levels" either in the sense of supervenience or mereology (cf. Sect. 4).

Moreover, Marr's computational level involves mathematical functions that require a systematic mapping onto physical states to describe a physical system (cf. Chalmers 2011).[16] Craver's mechanistic descriptions directly describe concrete processes that causally "map" physical inputs onto physical outputs. Hence, at least from Marr's perspective, there is a mathematical/concrete disparity between some of the levels in

Footnote 15 continued
function of the visual cortex." (15) More polemically, Marr claims that "[t]he discoveries of the visual neurophysiologists left one in a similar situation. Suppose, for example, that one actually found the apocryphal grandmother cell. Would that really tell us anything much at all? It would tell us that it existed (...) but not *why* or even *how* such a thing may be constructed from the outputs previously discovered cells." (15)

[16] Note that it makes little sense to say that Marr's computational level as a whole maps onto physical levels.

the two hierarchies. These observations suggest a non-identity of Marr's and Craver's explanatory ideals.

If the two normative models of explanation are neither distinct nor identical, the question arises whether one of the theories is more adequate and plausible than the other in light of scientific practice. I will now argue that, if taken literally, neither of the two modeling ideals actually secures satisfactory explanations of cognitive systems and of the cognitive phenomena realized by these.

As pointed out in Sect. 3, Marr believed that there are "[t]hree different levels at which an information processing device must be understood before it is understood completely", namely on the "computational level", the "representational and algorithmic level", and the "hardware and implementation level" (1982, pp. 24–25). All of these levels were essential for a "complete" explanation in his view. At the same time, it appears that he saw the computational and algorithimic levels to carry the main explanatory weight (cf. Marr 1982, pp. 11–15).[17] But also Marr would probably have conceded that having understood how and why the visual system transforms values representing light intensities into edge representations, and having mapped the computational principles onto the propagation of electric signals from the retinal receptors through LGN into V1 still makes a relatively limited explanation of the phenomenon only.

First, such a mapping does not by itself answer the question why it is these regions out of all that are involved in the visual task. Second, it does not specify what the cell assemblies within V1 actually do, how they actually manage to store the relevant information, and how they communicate with other cells. Third, knowledge of such a general projection pattern in the brain in a specific task environment does not help much in predicting how the same organism will behave in different tasks or even in the same task but under the influence of certainc drugs or substances. That is simply because a coarsely grained mechanistic theory is mute about activity patterns of cell networks and singular cells. In short, a mapping of the computational level onto a coarsely grained implementation level may simply not be very informative.

These observations do not imply that the original theory is not useful. If one's explanatory and manipulative goals are suitably modest, Marr's three levels are often all that is required. For instance, when the explanandum is a consistent misidentification of edges of certain transparent objects by humans, or when it is a general breakdown of visual perception in a patient with severe lesions in the thalamus, Marr's model may constitute a satisfactory explanans. It offers correct predictions about both cases, and it scores high on simplicity.

Notwithstanding, rigorous explanation in science is usually expected to aim, among other things, at a high predictive accuracy and a broad generalizability of the selected model (cf. Forster 2000). As Craver and others have long emphasized, there is an important asymmetry between explanation and prediction (cf.Craver 2007, ch. 2.4). At the same time, Craver's expectation has been that detailed mechanistic models

[17] This reading of Marr's "polemic" chapter 1 is perhaps not the only one. If Marr did take the implementation level at different grains to carry the same amount of explanatory weight, then perhaps he defended already a first version of the MC-acount of explanation. Marr's own contributions do not offer decisive textual evidence in either direction, as far as I can see.

will allow for better predictions in novel circumstances than purely functional models designed to fit a set of past observations.

Consequently, even if Marr gets the predictions right in some contexts, in many other contexts his three-level computational model of edge detection is likely to fall behind models that dive deeper into the mechanistic hierarchy. The reason is that the latter models usually offer more accurate predictions about the visual system in more circumstances and that they allow more successful generalizations to similar systems and their visual capacities.[18] In particular, detailed information about the conversion of light quanta into electric signals in the photoreceptor layer of the retina allows making predictions about the impact of certain genetic defects, lesions, drug infusion etc. on edge detection in human vision. Morever, the model can serve to generalize to the computational capacities of other parts of the brain, or even other kinds of brains, that happen to display the same kind of mechanistic structure. It is due to this increase in predictability and generalizability that the model specifying the mechanisms underlying the propagation of electric signals from the retinal receptors through LGN into V1 etc. typically offers a more satisfactory explanation of human vision.

In the opposite direction, it is questionable that neuroscience is primarily about a specification and modeling of mechanistic levels, and that computational descriptions have a secondary or merely preliminary role to play in the discovery of full-fledged mechanistic descriptions. Marr's theory of edge detection is an example in support of this claim. Knowing which brain regions are involved in the completion of a certain cognitive task, and how the various mechanisms underlying this activity look like, is certainly interesting and informative to some extent.

For instance, knowledge about the propagation of electric signals from the retinal receptors through LGN into V1 in the case of edge detection enables researchers to make certain general predictions about the effect of certain brain lesions in V1 onto vision. Moreover, it allows researchers to make predictions about the brain activity of other mammals in comparable tasks. However, knowing where visual information is processed in the brain provides almost no understanding of the kinds of information that are being processed, of how the information is transformed, and of how aspects of the visual environment are actually represented in the brain.

As Shagrir (2001) has shown, a physical system with cognitive capacities can sometimes be "carved up" in various incompatible ways, such that different aspects of the system become the vehicles of computation and the system as a whole computes different functions. Only when the task that the system completes is specified, it becomes possible to determine which of these different functions the system *actually* computes. In this sense, satisfactory explanations in neuroscience usually go beyond a purely mechanistic model of explanation.

To be sure, this demand for a computational dimension in certain explanations does not imply that all explanation should involve computational descriptions. For some systems, a full mechanistic description seems to be sufficiently explanatory.

[18] It should be pointed out that Marr does not explicitly commit himself to the claim that the implementation level should always take on a macroscopic grain. However, in his own examples it is not clear that a differentiation of the implementation level into several mechanistic levels is advisable and required for a satisfactory explanation.

For instance, an explanation of the human heart might be complete once the heart's pumping mechanism is explicated in detail and its place in the organism (its "causal function") is specified.[19] In contrast, to fully explain the working of a cognitive system, it does not suffice to describe its physiological structure and its causal-mechanistic connections. Without a specification of the system's representations and information processing strategies, it remains unclear why nature has equipped the system with these mechanisms and not others. The mechanistic description can only identify what is there; it cannot explain why it is there. Only when it is shown which tasks the system has to solve within its specific environment and what role its representations play in this respect, this important piece of information about the system becomes available.

As an example, consider the conversion of an array of light quanta of different intensities (the input of the mechanism) into a cascade of action potentials passing through LGN into the primary visual cortex (V1), which is locatable and describable in purely mechanistic or physiological terms, at least in principle. Such a mechanistic description does not help to understand what the system actually does, namely identifying edges of objects in its environment. It is not even clear which of the various outputs of the retina are actually relevant for the case of vision. Only a computational theory involving semantical and computational descriptions achieves such. In contrast, it seems much more plausible that the comparable mechanistic description would hep to understand what the heart does: it pumps blood.

To conclude, the MC-picture emphasizes the non-redundancy of computational descriptions, and it embraces Marr's distinction between computational entities such as functions, variables, values, algorithms etc., and implementation states (cf. Sect. 3). Neural mechanisms are considered as ontologically different from computations. Moreover, the mechanistic processes as such are described as insufficient for the implementation of a specific computational level description including the algorithm, the computed function, and the "why?"-element. At the same time, Craver's metaphysical hypothesis about the objective existence of a hierarchy of mechanistic levels and the normatively proclaimed integration from different fields into any complete explanation of a given cognitive phenomenon (cf. Sect. 4) are integrated into the MC-hypothesis. Moreover, within this framework, Craver's claim about the sufficiency of lower mechanistic levels for higher levels is consistent with Marr's denial of an intrinsicness of computational levels. It may actually come as a natural extension of Craver's framework that, according to the MC-hypothesis, computational functions can be sensibly associated with many mechanistic levels. There is not *the* computational level even within a single theory of a particular cognitive task such as edge detection. Rather, there are several, and potentially nested, ones. Neither is there *the* implementation, or mechanistic, level underlying a given computational task. Rather, there are several hierarchically ordered levels. Also Craver's definition of the constitution relation between mechanistic levels fits well with the MC-framework and the explanatory norms associated with it.

As a potential objection to the MC-ideal, one might worry that the unification of Marr's and Craver's views creates a tension between what the authors have said

[19] This observation does not deny, of course, that mathematical models and analyses of heart functions are in fact widely developed and used.

about the completeness of explanations in neuroscience respectively. Marr was quoted above as suggesting that a "complete" (cf. Marr 1982, p. 24) explanation specifies the computational function, the algorithm, and the implementation. Hence, in some of his works he seems to suggest that the specification of one or perhaps two such implementation level(s) at some grain is sufficient to complete the explanation and nothing more is required.[20]

Craver's focus on mechanistic levels, on the other hand, could be interpreted as implying that a full description of the physical entities and their physical activities is sufficient for attaining a satisfactory explanation for cognitive phenomena. No specification of computed functions and "why?"-elements seems to be required for the explanation in his view.[21]

These more or less explicit "and that is all what is required" claims found in Marr and Craver cannot be upheld if the norms of explanation associated with the two level frameworks respectively are united. The unified ideal, in contrast, demands that complete explanations of cognitive phenomena in the human brain ought to ascribe information-processing tasks not only to macro-systems, but to various kinds of subsystems of the brain. Only when this complex picture of the originally to-be-explained cognitive phenomenon can be laid out, the explanation becomes adequate and satisfactory. In this sense, not all of what Marr and Craver have said about explanation in neuroscience is consistent with the MC-framework.

A second potential objection may challenge the need for computational explanations on top of mechanistic explanations by pointing out that physical mechanisms are often already characterized mathematically. For instance, the mechanism of stochastic resonance (cf. Benzi et al. 1981), the mechanism of heat transfer in metals (cf. Qiu and Tien 1993), and the mechanism of gene expression (cf. Chen et al. 1999) have all been modeled with differential equations. Hence, it may seem as though computational explanations of mechanistically modeled physical systems are often simply redundant. Arguments of this kind neglect the fact that not every mathematical model is a computational model. The latter, in contrast to the former, essentially involves an interpretation of the arguments and values of a mathematical function as units of information, as well as a specification of a task that the modeled system is taken to perform in its environment. The unsupplemented mathematical models of stochastic resonance, of heat transfer in metals, and of gene expression do not fulfill these broader requirements. In this sense, not every mathematical model of a given mechanism satisfies the MC-ideal of modeling in cognitive science.

A third objection may point out that a non-computational system can also perform a task. For instance, a vacuum cleaner has been designed to complete the task of cleaning floors and carpets. This is entirely correct, of course. But it does not challenge the claim

[20] An example where Marr talks about two implementation levels is in an article co-authored with Tomaso Poggio where he says that, "[f]or a machine that solves an information processing problem, there are four important levels of description. At the lowest, there are basic component and circuit analysis—how do transistors, neurons, diodes, and synapses work?" (Marr and Poggio 1976).

[21] Craver's more recently defended ontic view of explanation points into the same direction. As he has claimed, "to explain something, one might plausibly argue, just is to show how it fits into the causal structure of the world" (Craver 2009, p. 578). However, mathematical functions are not ontic in the sense that physical entities are. If they are explanatory, they are at least not explanatory in the same way as flying rocks and spiking neurons.

that, in order to explain a cognitive system, it suffices to model its mechanisms only. Whether a vacuum cleaner can be explained in a satisfactory way by a description of its physical structure only (and without reference to its sociological and hygenic function) is a question I will not go into here.

A fourth objection could perhaps add that a vacuum cleaner could be a computing system also, according to the MC-model. In particular, it could be that it is a random number generator, where the random numbers are encoded by the configuration of dust particles. From the view point of the MC-model, this is much less absurd than it might first seem. If the generation of random numbers in a certain context serves to solve a task, a vacuum cleaner can well be a cognitive system.

6 Redundancy

The MC-theory presented in Sect. 2 and defended further in Sect. 5 adapts Marr's metaphysical claim about the non-intrinsicness, and relative autonomy, of the computational levels from the implementation level. The metaphysical claim implies that computational explanations are informative and non-redundant vis-à-vis mechanistic explanations. This Section makes a short digression by discussing the question whether Marr and the MC-account are right on this metaphysical point and on the associated epistemic claim about explanations.

First, note that the MC-ideal goes beyond Marr in at least two ways: The first is its demand that a satisfactory explanation of a cognitive system or phenomenon essentially requires the identification and analysis of several levels of implementation, or of the implementation level "at different grains". This idea, adopted from the mechanistic approach, does not contradict anything Marr said. But it is a stronger demand than the position explicitly advanced by Marr in his own contributions. Marr also did not explicitly commit himself to the claim that, for many of these implementation levels at different grains, both an algorithmic and a computational level should be specifiable. This demand of the MC-account again does not contradict Marr, but it goes again beyond what Marr was ready to state.

The recent literature on computational explanation has sometimes challenged the distinction between computations and mechanisms as well as between computational explanations and mechanistic explanations. For example, Milkowski (2013a, b) has argued that there are valuable insights in Marr's approach, but that these insights are preserved in the mechanistic approach that does not need any further enrichment from Marr anymore (cf. also Bickle 2015). Furthermore, Piccinini has characterized "computational explanation [as] a specific kind of mechanistic explanation." (2006, p. 350) With respect to the metaphysical thesis, he argues in (2008b) against "computational nihilists" who hold that there is no local fact of the matter whether a physical system is a computer or not. As an example, Churchland and Sejnowski have claimed that "there is no intrinsic property necessary and sufficient for all computers, just the interest-relative property that someone sees value in interpreting a system's states as representing states of some other system (...)." (1992, pp. 65–66)

In contrast, Piccinini believes that "there is a fact of the matter: a computer is a calculator of 'large capacity' and a calculator is a mechanism whose function is

to perform a few computational operations on inputs of bounded but nontrivial size." (Piccinini 2008b, p. 33) In his view, a computation does not essentially involve extrinsic and semantic aspects such as information processing and representation (cf. Piccinini 2008a, p. 232; Piccinini and Bahar 2013, p. 477). Rather, whether a system computes is a local mechanistic matter (cf. also Piccinini 2007b, a, 2008c).

From this perspective, computational explanations are essentially re-descriptions of mechanistic explanations or simply "mechanism sketches". As Piccinini and Craver point out:

> [F]unctional analyses are *sketches of mechanisms*, in which some structural aspects of a mechanistic explanation are omitted. Once the missing aspects are filled in, a functional analysis turns into a full-blown mechanistic explanation. By this process, functional analyses are seamlessly integrated with multilevel mechanistic explanations. (Piccinini and Craver 2011 284; cf. also fn. 14 above).

If computational explanations as a specific kind of functional explanations are mechanism sketches, it follows that the referents and truthmakers of the computational descriptions are essentially the same as those of mechanistic descriptions.

Churchland and Sejnowski's contention about the observer-relativity of computation is probably too strong, and Piccinini and others must be granted a point here. Whether a frog's brain computes does not seem to depend on any actual or potential observer. It simply does.

A different argument against the intrinsicness or locality of computations has been advanced by Shagrir (2001). The author discusses a physical system **P** consisting of an input unit receiving electrical signals of different voltages, two gates, and an output unit. Shagrir demonstrates that, relative to different assignments of symbols '1' and '0' to different voltage thresholds, the gates implement AND-, OR-, or XOR-functions. In other words, the same physical system sometimes implements not only an abstract computational function $\mathcal{F}$, but several further ones $\mathcal{F}_1$, ..., $\mathcal{F}_n$ that are logically incompatible with $\mathcal{F}$.

A potential reaction to this observation is to say that the system computes all of $\mathcal{F}, \mathcal{F}_1, ..., \mathcal{F}_n$ at the same time. For instance, the implementation account of computation described by Chalmers (2011) yields this result (cf. Egan 1992, p. 446 for a similar idea). According to this view, what is required for a physical system **P** to implement a computational function $\mathcal{F}$ is a mapping f that maps internal states of **P** to internal states of $\mathcal{F}$, inputs to P to input states of $\mathcal{F}$, and outputs of P to output states of $\mathcal{F}$ (Chalmers 2011, p. 327). The system then computes every function that can be mapped onto it in this way.

In Shagrir's opinion, the fact that the system **P** computes all of $\mathcal{F}, \mathcal{F}_1, ..., \mathcal{F}_n$ because it can be "carved up" into various syntactic structures is correct in some sense but widely uninteresting (cf. Shagrir 2001, p. 379). The interesting or salient computational structure of a system coincides with a single syntactic structure. To determine this structure, certain environmental aspects of the system need to be taken into account. Typically, these aspects involve the task that the system solves in a specific environment. In other words, the semantic content of a system's computational states determines which function(s) the system *actually* computes out of the vast num-

ber of functions it could in principle be seen as computing. It follows that it is generally not obvious from a local point of view what function the system actually computes (cf. also Shagrir 2006, 2010a).

Shagrir's example of the systen **P** shows that Marr made a valid point by emphasizing the "why?"-element in the computational level (cf. Marr 1982, p. 22). The "why?"-element concerns certain real-world relations that correspond to the formal relations among the computational states of a system as specified by a given computational function (cf. Shagrir 2010b, p. 493). And the partial determination of *actual* computations by environmental aspects implies their non-locality and non-intrinsicness.

The MC-account adheres to this metaphysical thesis about the relationship between the computational and mechanistic or implementation levels. It concludes that, epistemically speaking, computational explanations are not redundant vis-à-vis mechanistic explanations. They are informative even when a complete mechanistic description of the system is in place. The referents and truthmakers of the computational descriptions are not identical to those of the local mechanistic descriptions, accordingly.[22]

Note, firstly, that this general proposal offers an elegant answer to the question how it is possible that one physical system computes and another does not whilst both are characterizable by the same mathematical descriptions. Only when a system processes information about aspects of its environment to complete a specific task can it be considered as actually computing a specific function.

Secondly, note that the metaphysical non-locality thesis is compatible with Chalmers' mapping account of computation. It grants that there are physical computations in the sense of the mapping. However, it denies that all of the functions that can be mapped onto the system, some of which are logically incompatible with others, are *actually* computed by the system.

Finally, note that also mechanisms are not always strictly local in the sense that they sometimes extend over the boundaries of a system. In light of this observation, Piccinini has argued that the function a cognitive system actually computes can generally be determined in a non-semantic way. As the says, "[b]y looking at which effects of which computations are functionally significant within a context, we can identify the computation that is explanatory within that context" (Piccinini 2008a, p. 231).

The problem of this argument is that, even though mechanisms sometimes transcend a system's boundaries, this non-locality has a different quality than the extrinsicness of computations. Mechanisms, i.e. processes involving physical entities and their causal interactions, cannot be changed if only features of the remote environment with no causal impact on the mechanism are changed. Such follows from the definition of a mechanism offered by Machamer et al. (2000): The mechanisms of a cognitive system *S* are determined by their entities and activities as well as their causal structures, which include certain starting and termination conditions.

For *S*'s computations, in contrast, matters are different. If the environment of *S* is modified without affecting *S*'s causal input-output structure, it is sometimes possible

[22] See also Rusanen and Lappi (2007) for a related defense of the non-redundancy view of computational explanations vis-à-vis mechanistic explanations.

to modify which of the syntactic structures of *S* performs a computation. Or as Shagrir says, "content does determine (...) which of the implemented formal structures constitutes the system's computational structure. In this important sense, computational processes are intentional through and through." (2001, p. 298) In consequence, it is still not obvious from a local point of view what function a cognitive system actually computes. What is needed is usually something more than a specification of its mechanistic and causal structure, namely a broad characterization of the system's environment including the tasks the system completes, or could complete, in that environment. The epistemic counterpart to this metaphysical thesis states that computational explanations are informative and non-redundant vis-à-vis mechanistic explanations.

7 Conclusion

This paper was concerned with the role that models that integrate mechanistic and computational descriptions play in cognitive science. I started by mentioning a controversy in contemporary neuroscience on the adequate explanatory methodology. I suggested that the methodological dispute implicitly restates a debate in the philosophy of cognitive science represented, among others, by neuroscientist David Marr and philosopher Carl Craver. Marr grounds his normative model onto a descriptive distinction of "computational levels". Craver builds on a hierarchy of "mechanistic levels" as the descriptive foundation of the mechanistic norm of explanation. I argued that neither of the two models secures satisfactory explanations of cognitive systems.

The "mechanistic-computational" ideal can be thought of as resulting from a fusion of Marr's and Craver's accounts. It states that a satisfactory model of a cognitive system must bridge phenomena at multiple mechanistic levels, such that at least several of these mechanistic levels are shown to implement computational processes. The relevant parts of the computation must be mapped onto distinguishable entities and activities of the mechanism. I have defended this modeling ideal as adequate and plausible in light of scientific practice, and I have discussed certain metaphysical background assumptions. More specifically, I have briefly defended a metaphysical non-intrinsicness hypothesis about computations and mechanisms and a connected epistemological hypothesis about the non-redundance of computational vis-à-vis mechanistic explanations.

Due to limits of space, this paper was unable to answer certain further questions about the precise relationship between the various computational descriptions located on different levels of the mechanistic hierarchy in the sense of the MC-framework. Are these analytically related, or is there a freedom of choice in constructing the computational descriptions such that the higher computational models are not inferrable from the lower ones (where "higher" and "lower" should be understood relative to the mechanistic hierarchy)? Secondly, if the normative demands associated with the MC-framework are roughly correct, does a satisfactory explanation of a cognitive system go all the way down such that computational characterizations are required even on a molecular or atomic level in order to explain a given cognitive system? That seems implausible. But then how "deep" is the explanation supposed to go in each case?

Moreover, the MC-account of modeling and explanation predicted that a satisfactory explanation of the human brain as a cognitive system will eventually have to involve both, a modeling of the various mechanistic levels within the brain as targeted by neuroscience, as well as statements of the computational principles and relations that these levels implement. However, the MC-account itself did not answer the question which kinds of research projects are justified in light of the expected scientific gains and the required financial resources. The above investigation can only hope to have provided the conceptual background before which these further questions on explanation and research in cognitive science can be addressed.

Acknowledgements Open Access funding provided by Projekt DEAL.

References

Baumgartner, M., & Casini, L. (2017). An abductive theory of constitution. *Philosophy of Science*, *84*(2), 214–233.

Baumgartner, M., & Gebharter, A. (2016). Constitutive relevance, mutual manipulability, and fat-handedness. *The British Journal for the Philosophy of Science*, *67*(3), 731–756.

Bechtel, W. (1994). Levels of description and explanation in cognitive science. *Minds and Machines*, *4*(1), 1–25.

Bechtel, W. (2009). Looking down, around, and up: Mechanistic explanation in psychology. *Philosophical Psychology*, *22*(5), 543–564.

Bechtel, W., & Richardson, R. (1993). *Discovering complexity: Decomposition and localization as scientific research strategies*. New York: 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*(2), 312–322.

Benzi, R., Sutera, A., & Vulpiani, A. (1981). The mechanism of stochastic resonance. *Journal of Physics A: Mathematical and General*, *14*(11), L453.

Bickle, J. (2015). Marr and reductionism. *Topics in Cognitive Science*, *7*(2), 299–311.

Chalmers, D. (2011). A computational foundation for the study of cognition. *Journal of Cognitive Science*, *12*, 323–357.

Chen, T., He, H. L., Church, G. M., et al. (1999). Modeling gene expression with differential equations. *Pacific Symposium on Biocomputing*, *4*, 4.

Churchland, P. S., & Sejnowski, T. J. (1992). *The computational brain*. Cambridge, MA: MIT press.

Couch, M. (2011). Mechanisms and constitutive relevance. *Synthese*, *183*(3), 375–388.

Craver, C. (2001). Role functions, mechanisms, and hierarchy. *Philosophy of Science*, *68*(1), 53–74.

Craver, C. (2007). *Explaining the brain*. New York: Oxford University Press.

Craver, C. (2008). Constitutive explanatory relevance. *Journal of Philosophical Research*, *32*, 3–20.

Craver, C. (2009). Mechanisms and natural kinds. *Philosophical Psychology*, *22*(5), 575–594.

Craver, C., & Darden, L. (2001). Discovering mechanisms in neurobiology. In P. Machamer, R. Grush, & P. McLaughlin (Eds.), *Theory and method in the neurosciences* (pp. 112–137). Pittsburgh: University of Pittsburgh Press.

Dawson, M. R. (1998). *Understanding cognitive science*. Oxford: Blackwell.

Egan, F. (1992). Individualism, computation, and perceptual content. *Mind*, *101*(403), 443–459.

Fodor, J. A. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.
Forster, M. R. (2000). Key concepts in model selection: Performance and generalizability. *Journal of Mathematical Psychology*, *44*(1), 205–231.
Frank, M. J., & Claus, E. D. (2006). Anatomy of a decision: Striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychological Review*, *113*(2), 300.
Gebharter, A. (2017). Uncovering constitutive relevance relations in mechanisms. *Philosophical Studies*, *174*(11), 2645–2666.
Glennan, S. (1996). Mechanisms and the nature of causation. *Erkenntnis*, *44*(1), 49–71.
Gluth, S., Hotaling, J. M., & Rieskamp, J. (2017). The attraction effect modulates reward prediction errors and intertemporal choices. *Journal of Neuroscience*, *37*(2), 371–382.
Harbecke, J. (2010). Mechanistic constitution in neurobiological explanations. *International Studies in the Philosophy of Science*, *24*(3), 267–285.
Harbecke, J. (2014). The role of supervenience and constitution in neuroscientific research. *Synthese*, *191*(5), 725–743.
Harbecke, J. (2015). The regularity theory of mechanistic constitution and a methodology for constitutive inference. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *54*, 10–19.
Harbecke, J. (2019). Two challenges for a boolean approach to constitutive inference. *European Journal for Philosophy of Science*, *9*(1), 17.
Harbecke, J., & Shagrir, O. (2019). The role of the environment in computational explanations. *European Journal for Philosophy of Science*, *9*(3), 37.
Kaplan, D. M., & Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of Science*, *78*(4), 601–627.
Machamer, P., Darden, L., & Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*(1), 1–25.
Markram, H. (2006). The blue brain project. *Nature Reviews Neuroscience*, *7*(2), 153–160.
Marr, D. (1982). *Vision*. San Francisco: Freeman.
Marr, D., & Poggio, T. (1976). *From understanding computation to understanding neural circuitry*. Cambridge, MA: Massachusetts Institute of Technology, Artificial Intelligence Laboratory.
McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133.
Milkowski, M. (2013b). A mechanistic account of computational explanation in cognitive science. In N. Sebanz, M. Knauff, M. Pauen, & I. Wachsmuth (Eds.), *Cooperative minds: Social interaction and group dynamics. Proceedings of the 35th annual meeting of the cognitive science society, Austin, TX* (pp. 3050–3055). Cognitive Science Society.
Milkowski, M. (2013a). *Explaining the computational mind*. Cambridge: MIT Press.
Milkowski, M., Clowes, R. W., Rucińska, Z., Przegalińska, A., Zawidzki, T., Gies, A., et al. (2018). From wide cognition to mechanisms: a silent revolution. *Frontiers in Psychology*, *9*, 2393.
Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plansand the structure of behavior*. New York: Holt.
Newell, A., Simon, H. A., et al. (1972). *Human problem solving* (Vol. 14). Englewood Cliffs, NJ: Prentice-Hall.
O'Hara, K. (1994). *Mind as machine: Can computational processes be regarded as explanatory of mental processes?*. Ph.D. thesis, Worcester College, University of Oxford.
Payzan-LeNestour, E., Dunne, S., Bossaerts, P., & ODoherty, J. P. (2013). The neural representation of unexpected uncertainty during value-based decision making. *Neuron*, *79*(1), 191–201.
Peacocke, C. (1986). Explanation in computational psychology: Language, perception and level 1.5[1]. *Mind & Language*, *1*(2), 101–123.
Piccinini, G. (2006). Computational explanation in neuroscience. *Synthese*, *153*(3), 343–353.
Piccinini, G. (2007a). Computational modelling vs. computational explanation: Is everything a turing machine, and does it matter to the philosophy of mind? *Australasian Journal of Philosophy*, *85*(1), 93–115.
Piccinini, G. (2007b). Computing mechanisms. *Philosophy of Science*, *74*(4), 501–526.
Piccinini, G. (2008a). Computation without representation. *Philosophical Studies*, *137*(2), 205–241.
Piccinini, G. (2008b). Computers. *Pacific Philosophical Quarterly*, *89*(1), 32–73.
Piccinini, G. (2008c). Some neural networks compute, others don't. *Neural Networks*, *21*(2), 311–321.
Piccinini, G., & Bahar, S. (2013). Neural computation and the computational theory of cognition. *Cognitive Science*, *37*(3), 453–488.

Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*,. https://doi.org/10.1007/s11229-011-9898-4.

Polanía, R., Krajbich, I., Grueschow, M., & Ruff, C. C. (2014). Neural oscillations and synchronization differentially support evidence accumulation in perceptual and value-based decision making. *Neuron*, *82*(3), 709–720.

Putnam, H. (1988). *Representation and reality*. Cambridge, MA: MIT Press.

Qiu, T., & Tien, C. (1993). Heat transfer mechanisms during short-pulse laser heating of metals. *Journal of Heat Transfer (Transactions of the ASME (American Society of Mechanical Engineers), Series C; (United States)*, *115*(4), 12.

Rusanen, A.-M., & Lappi, O. (2007). The limits of mechanistic explanation in neurocognitive sciences. In *Proceedings of the European cognitive science conference*.

Searle, J. R. (1992). *The rediscovery of the mind*. Cambridge, MA: MIT Press.

Shagrir, O. (2001). Content, computation and externalism. *Mind*, *110*(438), 369.

Shagrir, O. (2006). Why we view the brain as a computer. *Synthese*, *153*(3), 393–416.

Shagrir, O. (2010a). Brains as analog-model computers. *Studies In History and Philosophy of Science Part A*, *41*(3), 271–279.

Shagrir, O. (2010b). Marr on computational-level theories. *Philosophy of Science*, *77*(4), 477–500.

Shagrir, O., & Bechtel, W. (2017). Marrs computational level and delineating phenomena. In D. M. Kaplan (Ed.), *Explanation and integration in mind and brain science* (pp. 190–214). Oxford: Oxford University Press.

Sharifi, M., Fathy, M., & Mahmoudi, M. T. (2002). A classified and comparative study of edge detection algorithms. In *Proceedings of the international conference on information technology: Coding and computing (ITCC-02)* (pp. 117–120). IEEE.

Shepherd, G. M., Rolls, E., Andreou, A., & Peitsch, M. (2011). Evaluation of the blue brain project and human brain project - epfl, lausanne. http://www.academia.edu/4769788/Evaluation_of_the_Blue_Brain_Project_and_Human_Brain_Project_EvaluationCommittee_Table_of_contents.

Skipper, R. A., Jr. (1999). Selection and the extent of explanatory unification. *Philosophy of Science*, *66*, 196–209.

Varzi, A. (2019). Mereology. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2019 ed.). Stanford: Metaphysics Research Lab, Stanford University.

Wallis, J. D. (2012). Cross-species studies of orbitofrontal cortex and value-based decision-making. *Nature Neuroscience*, *15*(1), 13.