Synthese (2021) 198:4573–4601
https://doi.org/10.1007/s11229-019-02357-9

# Two kinds of explanatory integration in cognitive science

**Samuel D. Taylor**[1]

Received: 14 December 2018 / Accepted: 3 August 2019 / Published online: 30 August 2019


**Abstract**
Some philosophers argue that we should eschew cross-explanatory integrations of mechanistic, dynamicist, and psychological explanations in cognitive science, because, unlike integrations of mechanistic explanations, they do not deliver genuine, cognitive scientific explanations (cf. Kaplan and Craver in Philos Sci 78:601–627, 2011; Miłkowski in Stud Log 48:13–33, 2016; Piccinini and Craver in Synthese 183:283–311, 2011). Here I challenge this claim by comparing the theoretical virtues of both kinds of explanatory integrations. I first identify two theoretical virtues of integrations of mechanistic explanations—unification and greater qualitative parsimony—and argue that no cross-explanatory integration could have such virtues. However, I go on to argue that this is only a problem for those who think that cognitive science aims to specify one fundamental structure responsible for cognition. For those who do not, cross-explanatory integration will have at least two theoretical virtues to a greater extent than integrations of mechanistic explanations: explanatory depth and applicability. I conclude that one's views about explanatory integration in cognitive science cannot be segregated from one's views about the explanatory task of cognitive science.

**Keywords** Cognitive science · Integration · Cross-explanatory · Mechanistic explanation · Dynamicist explanation · Psychological explanation

## 1 Introduction

According to Piccinini and Craver (2011, p. 284), we can attain a unified science of cognition "by showing how functional analyses of cognitive capacities can be and in some cases have been integrated with the multilevel mechanistic explanations of neural systems." The problem, however, is that we do not yet have a working account of what

✉ Samuel D. Taylor
sam.taylor@hhu.de

[1] Department of Philosophy, Heinrich-Heine-University Düsseldorf, Universitätsstraße 1, 40225 Düsseldorf, Germany

explanatory integration entails (cf. Miłkowski 2016). While it is clear that explanatory integration demands that there be some constraints on the combination of two or more explanations—for instance, constraints on the space of possible (representations of) mechanisms supported by the explanations—, it is unclear if any given constraint or set of constraints is necessary and sufficient (Craver 2007; Thagard 2007). As a result, open questions remain about the viability of integrating all explanations formulated in the cognitive sciences into one coherent account of the states and processes responsible for cognition (Newell 1990).

Miłkowski (2016) has argued that the integration of mechanistic explanations is of paramount importance in cognitive science. This view aligns with Kaplan and Craver's (2011) claim that integrations of mechanistic explanations deliver some of the only genuine, cognitive scientific explanations.[1] On the standard formulation, mechanistic, cognitive scientific explanations are those explanations that explain by identifying, through analysis, the component parts of the mind/brain (e.g. neurons, modules, etc.) and their principles of interaction, before showing how these component parts causally interact to generate some phenomena. (cf. Glennan 1996; Horst 2007; Machamer et al. 2000). An integration of two mechanistic, cognitive scientific explanations, therefore, will show how two sets of causally efficacious parts and interactions—e.g. two mechanisms—are co-organised to generate the cognitive phenomena for which both are co-responsible. Such integration can occur, for example, when one mechanism is shown to itself be a part of another mechanism that contributes to the working of the latter by producing the phenomena for which it is responsible.

Here I compare the integration of mechanistic explanations with another kind of integration in cognitive science: the cross-explanatory integration of mechanistic, dynamicist, and psychological explanations. In contrast to mechanistic explanations, dynamicist explanations explain not by decomposing the mind/brain into entities and interactions, but by identifying the critical variables characterising the state of the system and constructing laws—that is, sets of differential equations—to account for changes to the system's state (Chemero 2009; Varela et al. 1991).[2] Like mechanistic explanations, psychological explanations explain by capturing causal structures between components, but do so by representing abstract relationships between functional/intentional components that lack spatio-temporal organisation (Weiskopf

[1] Kaplan and Craver (2011, p. 603) accept that there are "domains of science in which mechanistic explanation is inappropriate." However, the examples they give are of "certain areas of physics [...] that do not involve decomposing phenomena into component parts (Bechtel and Richardson 2010; Glennan 1996)" and of "mental phenomena, such as belief and inference, [that] are fundamentally normative and so demand noncausal forms of explanation (McDowell 1996)." The first is not clearly an explanation of cognition, because physical systems are just as likely non-cognitive. And the second is not clearly a cognitive scientific explanation, because noncausal explanations of normative phenomena like belief and inference need not be informed by empirical data about the exercise of cognitive competences. Of course, Kaplan and Craver may think that there are non-mechanistic explanations in physics that are explanations of cognition; and that noncausal explanations of belief and inference are informed by empirical data about the exercise of cognitive competences. But we cannot be sure. This ambiguity—given that they are supposed to be demonstrating that they "do not intend [...] to rule out nonmechanistic explanation generally"—is worth noting. However, it is not necessary for my argument to defend the stronger claim that Kaplan and Craver take integrations of mechanistic explanations to deliver the *only* genuine, cognitive scientific explanations.

[2] Formally, the set of differential equations that can be solved to characterise the changing state of a system as a trajectory through a state space.

2017). Neither of these two kinds of explanations are (necessarily) mechanistic, although advocates of cross-explanatory integration will think that they should be integrated with mechanistic explanations to give a more "complete" explanation of the brain/mind.

In Sects. 2 and 3 of this paper, I consider an example of an integration of mechanistic explanations and identify two theoretical virtues of such integrations: unification and greater qualitative parsimony. In Sect. 4, I introduce dynamicist and psychological explanations to illustrate that no potential case of cross-explanatory integration could have the theoretical virtues of unification and greater qualitative parsimony. However, in Sect. 5 I argue that this only undermines cross-explanatory integration if we adopt a fundamentalist attitude towards cognitive scientific explanation, which conceives of cognitive science as aiming to specify the fundamental structure responsible for cognition. In Sect. 6, I work from an anti-fundamentalist basis to show that cross-explanatory integrations can be taken to have two theoretical virtues to a greater extent than do integrations of mechanistic explanations: explanatory depth and increased applicability. I conclude that one's views about explanatory integration in cognitive science cannot be segregated from one's views about the explanatory task of cognitive science.

## 2 Integrating mechanistic explanations: an example

A mechanism need be thought of as nothing more than "a structure, responsible for one or more phenomena, that performs a function in virtue of its component parts, component operations, and their organization" (Bechtel and Abrahamsen 2005). A mechanism, therefore, need not be *deterministic* (it's components may be stochastic) (Bogen 2005, 2008); nor *reductionistic* (it may be, e.g., a multilevel explanations spanning a range of spatio-temporal levels of grain) (Bechtel 2009); nor *sequential* or *linear* (it may include feedback loops wherein the output of the mechanism or components in turn influences the input of the mechanism or components in a subsequent iteration) (Bechtel 2011); nor *localisable* (components of mechanisms might be widely distributed (as are many brain mechanisms) and might violate our intuitive sense of the boundaries of objects (as an action potential violates the cell boundary) (cf. Craver and Tabery 2017, for a thorough account of what mechanisms are and are not). It need only be a collection of "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).

Broadly speaking, any integration of mechanistic explanations aims to arrive at the best set of cohering, mechanistic explanations of its explananda. With respect to cognitive science and in the maximally explanatory case, an integration of mechanistic explanations would hope to account for much of the phenomena associated with (human) cognition. It is an open question how much can be explained by an integrated mechanistic explanations. Craver and Kaplan (2018), for example, argue that the "completeness" of any mechanistic explanation will depend on how many "relevant details" of "explanatory knowledge" it "stores." Still, we can assume that those in favour of mechanistic explanations will, at the very least, take integrations of these

explanations to be able to explain cognitive competencies such as language production and comprehension, memory, perception, problem solving, categorisation, and reasoning; but also general, flexible behaviours and real-time performance, as well as the processes of learning and development that are characteristic of the human cognitive system.

Concerns about how to integrate mechanistic explanations in cognitive science arise whenever we are uncertain how to "fit together" two or more mechanistic explanations in a way that gives us insight into a cognitive system's organisation (Weiskopf 2017). Piccinini and Craver (2011, p. 307), however, argue that cognitive science has "advanced to the point that these pursuits can meaningfully come together, and there are tremendous potential benefits from affecting such integration." According to Craver (2007), integrating mechanistic explanations involves integrating three perspectives: the "isolated perspective (level 0)" that characterises the mechanism with respect to the causal processes (input-output relations) that the mechanism is meant to explain; the "contextual perspective (level + 1)" that locates the mechanism as a contributing part of another mechanism; and, finally, the "constitutive perspective (level − 1)" that breaks down the mechanism into its constitutive parts and interactions to make perspicuous how the interactions of these parts give rise to the causal story told at level 0.

It is clear, therefore, that the integration of mechanistic explanations will entail the telling of an "inter-level" story that relates, in one way or another, two or more mechanistic explanations (Miłkowski 2016). An almost canonical example of this kind of integration is given in the discussion of explanations of Long-Term Potentiation (LTP) and spatial memory in Bechtel (2009), Craver (2005), and Craver (2007). Marraffa and Paternoster (2013, p. 14) provide a clear summary of Craver's (2007) account as follows:

> Craver (2007) examines the development of the explanations of Long-Term Potentiation (LTP) and spatial memory. He distinguishes at least four levels. At the top of the hierarchy (the behavioral-organismic level) are memory and learning, which are investigated by behavioral tests. Below that level is the hippocampus and the computational processes it is supposed to perform to generate spatial maps. At a still lower level are the hippocampal synapses inducing LTP. And finally, at the lowest level, are the activities of the molecules of the hippocampal synapses underlying LTP (e.g., the N-methyl Daspartate receptor activating and inactivating). These are "mechanistic levels" or "levels of mechanisms": the N-methyl D-aspartate receptor is a component of the LTP mechanism, LTP is a component of the mechanism generating spatial maps, and the formation of spatial maps is a part of the spatial navigation mechanism. Integrating these four mechanistic levels requires both a "looking up" integration, which will show that an item (LTP) is a part of a upper-level mechanism (a computational-hippocampal mechanism); and a "looking down" integration, which will describe the lower-level mechanisms underlying the higher-level phenomenon (the molecular mechanisms of LTP) (See Fig. 1).

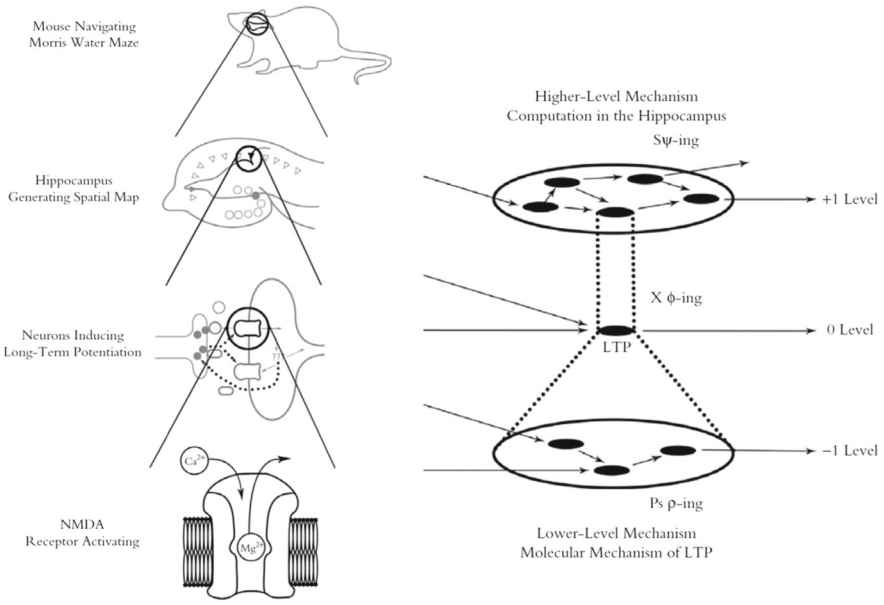


**Fig. 1** Levels of spatial memory (left). Integrating levels of mechanisms (right). . Source: Craver (2007)

For Craver (and others), therefore, the integration of mechanistic, cognitive scientific explanations entails giving a causal explanation of a cognitive phenomenon (e.g. cognitive competency or cognitive system) that spans various "levels."[3] For example, that spans behavioural capacities (memory and learning), brain regions (hippocampus), and neural structures (synapses). In this sense, integrations of mechanistic explanations demonstrate how mechanisms have other mechanisms as their interacting parts and so are "intrinsically organized in multilayers" of mechanisms. In other words, integrations of mechanistic explanations show how two or more mechanisms are "hierarchically organised" in such a way that organisations of lower level entities and activities are the component parts of higher level organisations of entities and activities (Craver 2001).[4] Integrations of mechanistic explanations will, therefore, be "multilevelled" in the sense that they account for the way that a "hierarchically organ-

[3] Note that it would be incorrect to say that *phenomena* at different levels of organization are integrated within a mechanistic explanation. Mechanistic explanations are epistemic products of human ingenuity, which purport to represent component entities/parts, their organisation, and their interactions. Specifying a mechanism is part and parcel of what it means to give a mechanistic explanation, but it is a further issue as to whether or not a mechanistic explanation accurately or truthfully represents reality. Here, I confine myself to a discussion of integrations of mechanistic *explanations* as explanations given for some phenomena; leaving aside any discussion of how or if they represent what they purport to represent.

[4] One may suppose that there is a distinction to be made here between, on the one hand, integrations of mechanisms within the *same* mechanistic explanation and integrations of mechanisms from *different* mechanistic explanations. But this distinction cannot hold water. Consider Craver's example again. Where, exactly, should we draw the line between the "same" and "different" mechanistic explanations? For sure, the multileveled, mechanistic explanation of LTP *and* spatial memory is a single, mechanistic explanation. But we can still give different—albeit less "complete" in Kaplan and Craver's (2018) sense—mechanistic explanations of LTP *or* spatial memory: one in terms of computational mechanisms in the hippocampus

ised mechanism" produces a certain cognitive competency or behaviour in virtue of itself being composed of causally efficacious mechanisms.[5]

## 3 Two virtues of integrating mechanistic explanations

### 3.1 Unification

Craver's account supposes that an integration of mechanistic explanations takes all the mechanisms specified by the explanations being integrated and locates them in a single, hierarchically organised mechanism. Suppose, then, that $e_1$ were a mechanistic explanation of the visual processes responsible for edge detection, $e_2$ were a mechanistic explanation of depth perception, and $e_3$ were a mechanistic explanation of colour perception. According to Craver's account, $e_1$, $e_2$, and $e_3$ would only be integrated when the mechanisms specified by all three explanations were located in a single, hierarchically organised mechanism that accounts for edge detection, depth perception, and colour perception.[6] The upshot is that the integration of mechanistic

Footnote 4 continued
(cf. Knierim and Neunuebel 2016, as one of many examples); the other in terms of molecular mechanisms of the hippocampul synapses (cf. Bliss and Collingridge 1993, as one of many examples).

One may retort that these "two" mechanistic explanations are not "different," because they can be accommodated in a single, integrated mechanistic explanation with the same (unified) explananda; e.g. LTP *and* spatial memory. But how are we to know *a priori* where integration is and is not possible? Might it not be the case that another mechanistic explanation—say, of the activities of molecules of the central nervous system (e.g. the $\alpha$-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptor activating and inactivating)—could also be integrated with the mechanistic explanation of LTP *and* spatial memory? Suppose that it could be, but has not yet been, integrated (which is not beyond the realms of possibility). Would this imply that the mechanistic explanations of LTP *and* spatial memory, and the mechanistic explanation of the activities of molecules of the central nervous are the "same" explanation even prior to integration? To answer "yes" here is plainly absurd.

The point, therefore, is that the entire distinction between the "same" and "different" mechanistic explanations is relative to the kinds of explanations we have developed. There is no sensible question about whether two mechanistic explanations are the "same" or "different" until after the fact of (un)successful integration, when it is shown that two or more mechanistic explanations were or were not part of the "same" explanation all along. This, of course, makes the language of "same" and "different" mechanistic explanations superfluous. What we have, rather, are those mechanistic explanations that are integrated and those that are not. Those that are not will by necessity address different explananda (e.g. LTP *or* spatial memory), but those that are will by necessity address the same explananda (e.g. LTP *and* spatial memory). There is, therefore, no such thing as integrations of the "same" or "different" mechanistic explanations; there are only integrated and not-integrated mechanistic explanations.

[5] Note that every mechanistic explanation that is integrated must do some relevant explanatory work. This could be achieved if that explanation helps to explain a previously unexplained explanandum, thereby increasing the number of explananda accounted for by the integrated mechanistiic explanation specifying a hierarchically organised mechanism (e.g. makes the integrated explanation more complete); or if it contributes to an existing explanation of some explanandum, thereby consolidating and/or furthering the explanatory power attained by specifying the hierarchically organised mechanism (e.g. makes the integrated explanation more deep).

[6] One may worry that the integrated mechanistic explanation $\{e_1, e_2, e_3\}$ lacks a clear explanandum, but this is arbitrary. Why not simply say that the explanandum of $\{e_1, e_2, e_3\}$ is the visual processes responsible for edge detection *and* depth perception *and* colour perception? Is this not an explanandum of cognitive science? It seems clear that it could be. How are we to know *a priori* where the boundaries between "different" mechanistic explanations of "different" explananda lie? The answer, again, is that we do not

explanations will be a process that takes a set of explanations $\{e_1 \ldots e_x\}$ specifying different mechanisms and unifies them within a single, multilevel explanation specifying one, hierarchically organised mechanism. It seems, therefore, that on Craver's account *unification* is an unmistakable virtue of the integration of mechanistic explanations.

Miłkowski (2016, pp. 26–27), however, has recently argued that "unification should be considered to be an epistemological virtue of scientific representations rather than of mechanisms described by these representations." I make no claim about the virtues of "mechanisms." Rather, I am concerned with the virtues of integrations of mechanistic *explanations*, which purport to represent some objective mechanism and so to explain (the generation of) some phenomenon. It would seem, therefore, that there is no disagreement between Miłkowski and I. But Miłkowski goes on to argue that mechanistic explanations are not (always) "unified explanations," because they do not (always) possess the properties of "simplicity, invariance or unbounded scope, and non-monstrosity."[7] Simplicity, he argues, will not necessarily be a property of mechanistic *explanations*, because "models of mechanisms should be simple and parsimonious only as far as it aids their uses" (ibid., p. 27). Neither will invariance or unbounded scope, because the mechanisms specified by mechanistic explanations may account for "local" phenomenon that occur "only in certain spatiotemporal locations." And neither will non-monstrosity, because "structures may exist that are composed of relatively independent subsystems," and so any mechanistic explanations specifying a "totally interdependent" mechanism "would be at best an idealization."

It follows, for Miłkowski, that *integrated* mechanistic explanations and "*unified*" mechanistic explanations should be differentiated, because integrated mechanistic explanations "need not be simple, beautiful, or general." Rather, integrated mechanistic explanations need only satisfy certain constraints on their being combined "in a coherent manner" (ibid., pp. 17–18). For example, constraints that regulate "the boundaries of the space of plausible mechanisms" or the "probability distribution over that space."[8] So, in cases of the integrated mechanistic explanations, Miłkowski argues that:

> Even if mechanistic constraints are preserved, the resultant representation may be quite disconnected; for example, one can integrate the account of the cognitive map in the hippocampus (Derdikman and Moser 2010) with, say, Baddeley's account of working memory (Baddeley and Hitch 1974). Both models refer to working memory but as (Baddeley 2000) notes, they use the notion to mean

Footnote 6 continued
know until after the fact of (un)successful integration, when it is shown either that two explanations are "different" and do not share the same explananda or that they are the "same" and share a (unified) explananda.

[7] Miłkowski (2016, 19–20) conceives of simplicity as "The classical principle of ontological parsimony," which holds "that entities should not be multiplied beyond necessity." He conceives of invariance or unbounded scope as either having "unlimited scope" to explain any phenomena (as with explanations based exclusively on natural laws) or as having the "maximal scope possible." And he endorses the definition of non-monstrosity given by Votsis (2015), whereby a monstrous explanation is an explanation with a "lack of shared relevant deductive consequences" in the sense that it contains "isolated islands" that are confirmationally disconnected, i.e., where what these "islands" imply is completely disjoint.

[8] This idea is taken straight from Craver (2007, p. 247).

different things; hence, even if rats have both kinds of memory, no explanatory unity is observed here (Miłkowski 2016, p. 19).

Miłkowski's differentiation of integrated mechanistic explanations and "unified explanations" depends on a particular type-identification of the latter as explanations that are simple, have an invariant or unbounded scope, and are non-monstrous. There will, of course, be many tokens of this kind of explanation, but Miłkowski is at pains to argue that developing explanations of this type is not a norm of mechanistic, cognitive science, even if it is often a "useful heuristic" (ibid., p. 26). Instead, Miłkowski argues that mechanistic explanations that are non-unified, but integrated, can still be "genuine or satisfying." For my purposes here, however, there is an open question: is Miłkowski's criteria for type-identifying "unified explanations" sufficient for type-identifying mechanistic explanations possessing the virtue of unification? And, hence, does Miłkowski's differentiation of "unified" and integrated mechanistic explanations entail a differentiation of integrated mechanistic explanations and explanations possessing the virtue of unification?

In his "systematisation of theoretical virtues," Keas (2018, p. 2775) defines the virtue of unification as an "aesthetic virtue" that an explanation $E$ possesses when $E$ "explains more kinds of facts than rivals with the same amount of theoretical content" (cf. Thagard 1978, for an earlier elaboration of the same point). Similarly, Mackonis (2013, p. 987) argues that an explanation possesses the virtue of unification when it "explain[s] more facts with [the] same resources." Neither Keas nor Mackonis argue that explanations possessing the virtue of unification must have the properties of invariance or unbounded scope, or non-monstrosity. Moreover, both make it clear that the virtue of unification differs from the virtue of simplicity, because the former, unlike the latter, will be attributed to an explanation $E$ only when a "comparative evaluation" establishes that $E$ helps to explain "more kinds of facts" and/or "different kinds of data" than its rivals.[9] The upshot is that if we endorse Keas' and Mackonis definition of the theoretical virtue of unification, it does not follow that Miłkowski's type-identification of "unified explanations" is sufficient for type-identifying explanations possessing the virtue of unification.[10]

My concern is with the virtue of unification in Keas' and Mackonis' sense. The question, then, is whether or not integrations of mechanistic explanations allow us to do more explanatory work with the same theoretical content; where "theoretical content" is measured in terms of the number of "entities postulated by the theory" (Keas 2018, p. 2775). Integrations of mechanistic explanations always represent a relative decrease in the number of entities postulated, because they decrease the number of *independent* mechanisms postulated by subsuming every mechanism specified by an individual, un-integrated explanation within a single hierarchically organised mechanism.[11] Still, integrations of mechanistic explanations do the same explanatory work as the set of un-integrated explanations and, hence, more explanatory work than rival

[9] Mackonis (2013, p. 987, my italics) makes the same point when he argues that an explanation possesses the virtue of simplicity when it "explain[s] *[the] same facts* with fewer resources."

[10] In fact, there is nothing to say that the virtue of unification—in Keas' and Mackonis' sense—could not be possessed by both unified and integrated mechanistic explanations in Miłkowski's sense.

[11] I consider this topic in detail in my discussion of the virtue of greater qualitative parsimony below.

explanations postulating a single mechanisms. Thus, integrated mechanistic explanations will always do more explanatory work than any rival explanation with the same theoretical content.

To see why this is the case, consider again the integrated mechanistic explanation $E = \{e_1, e_2, e_3\}$ mentioned above. This integrated mechanistic explanations would explain the visual processes responsible for edge detection *and* depth perception *and* colour perception by specifying one hierarchically organised mechanism that unifies the (causal) descriptions of entities and interactions given by each explanation separately. Any rival (set-of) mechanistic explanation(s) would either have increased theoretical content (e.g. would specify three separate mechanisms: one to explain edge detection, another to explain depth perception, and another still to explain colour perception) or would explain only some of the phenomena explained by $E$ (e.g. edge detection *or* depth perception *or* colour perception). Thus, the integrated mechanistic explanation $E$ represents a "comparative increase in the different kinds of data that get explained" by a single explanation. The same will hold for every integrated mechanistic explanation, because they will do at least the same amount of explanatory work as a set of un-integrated explanations, but will do this work with less theoretical content in virtue of specifying only one hierarchically organised mechanism.

No consideration of unification as this kind of theoretical virtue is apparent in Miłkowski's discussion. Rather, he is concerned with showing that "unified explanations"—as a general and elegant type of explanation—differ from integrated mechanistic explanations. This claim has no repercussions for my argument. There is, of course, a further question as to whether Miłkowski is correct. Answering this question will require close examination of whether or not Miłkowski's account of the success conditions for integrated and "unified" mechanistic explanations are correct and whether or not these conditions overlap.[12] But the question of whether or not Miłkowski's claims turn out to be correct is irrelevant for my purposes here. What matters is only that Miłkowski's differentiation of integrated and "unified" explanations does not impugn the claim that integrations of mechanistic explanations have the virtue of unification. This much cannot be denied, regardless of whether one thinks that integrated mechanistic explanations and "unified" mechanistic explanations are the same thing.

### 3.2 Greater qualitative parsimony

Integrations of mechanistic explanations also have the virtue of *greater qualitative parsimony*. Qualitative parsimony concerns the number of types (or kinds) of thing postulated by an explanations; whereas quantitative parsimony concerns the number

[12] Miłkowski "understand[s] integration in terms of constraints." He gives two examples of relevant constraints: one in terms of an "adequate" "representation of mechanisms" that "changes the boundaries of the space of plausible mechanisms or changes the probability distribution over that space" (Craver 2007, p. 247); and another that different explanations muse be "true at the same time" (Miłkowski 2016, pp. 17–18). The question, then, is, firstly, whether or not it is correct to say that integrations of mechanistic explanations satisfying these constraints are not, in fact, simpler, more general, and less-monstrous; and, secondly, whether or not it is correct to define "unified explanations" as explanations that have the properties of "simplicity, invariance or unbounded scope, and non-monstrosity" in the first place.

of individual things postulated. For example, the explanation that the damage to my car was caused by 10 children is more qualitatively parsimonious, but less quantitatively parsimonious, than the explanation that it was caused by 1 child, 1 bear, and 1 dog. The idea that qualitative parsimony is theoretical virtue is well-established in the literature and in the history of philosophy (cf. Quine 1964; Sober 1994, as examples of why qualitative parsimony (e.g. Occam's razor) is a theoretical virtue).[13] Therefore, to say that integrations of mechanistic explanations have greater qualitative parsimony is just to say that such integrations have the virtue of doing the same explanatory work by positing fewer kinds of things.

*Prima facie*, it seems that the number of kinds of mechanisms specified by a set of mechanistic explanations will not be affected by whether or not that set is integrated. For example, it seems that mechanistic explanations of, say, Long-Term Potentiation *and* spatial memory will always specify at least four mechanisms: a behavioural-organismic mechanism, a hippocampus-computational mechanism, a hippocampus-synapses mechanism, and molecular mechanism. However, we can see that this conclusion is mistaken when we factor in Craver's account of the integration of mechanistic explanations introduced above. For then we see that it is only in the case of an integration of mechanistic explanations that those mechanisms are mereologically subsumed as parts in one hierarchically organised mechanism. Therefore, integrations of mechanistic explanations postulate only one *superordinate kind* of mechanism that subsumes all other mechanisms as its parts.

All integrations of mechanistic explanations will postulate only kind of thing: a hierarchically organised mechanism that has other mechanisms as its parts. Such part may include, for instance, spatial mechanisms (cf Wimsatt 1997), temporal mechanisms (cf Bechtel 2013), stable and ephemeral mechanisms (cf Glennan 2009) neural mechanisms, or computational mechanisms (cf Miłkowski 2013), etc.[14] One could argue that there are other kinds of things postulated in such cases; namely, "bottoming-out" types of entities and activities that are the parts of the lowest level mechanisms. In cognitive science, such 'bottoming-out" entities and activities may include, for instance, the "descriptions of the activities of macromolecules, smaller molecules, and ions" provided by neurobiology (Machamer et al. 2000, pp. 13–15). However, these entities and activities must be accepted as "fundamental" in the sense that they demarcate where the "field stops when constructing mechanisms" (ibid., pp. 13–15). Thus, for all sets of mechanistic explanations in cognitive science—whether integrated or not—these entities and activities must be presupposed and so will not vitiate the increase in qualitative parsimony following integration.

There are open questions about the "independence" or "objecthood" of the parts of mechanisms. Simon (1996) argues that the parts of a mechanism have stronger and more abundant causal relations with other components in the mechanism than they do with items outside the mechanism, and that the decomposition of mechanisms into

[13] Lewis (1973, p. 87), for instance, subscribed "to the general view that qualitative parsimony is good in a philosophical or empirical hypothesis." For historical discussion of the theoretical virtue of qualitative parsimony see Sober (2015).

[14] These different kinds of mechanisms are individuated as classes by their different entities and interactions (cf. Miłkowski 2013, for an illuminating discussion of this idea with respect to computational mechanisms in particular).

parts will depend, in some way, on the intensity of interaction among components. Others—such as Craver (2007)—argue that a part of a mechanism is only defined relative to what one takes the mechanism to be doing. In any case, all agree that an integration of mechanistic explanations will result in the specification of a single hierarchically organised mechanism that has other mechanisms as its parts. This is just what Craver (2007) means when he talks about the "levels of a hierarchy of mechanisms" following integration. Moreover, it is clear that a mechanism that subsumes others will be of a superior order within the classification of mechanisms. The idea here, then, is just that the class of hierarchically organised mechanisms is a superordinate class, because the kind of mechanisms grouped in that class are mechanisms of mechanisms.

With this in mind, we can consider the following toy example to see how integrations of mechanistic explanations have the virtue of greater qualitative parsimony. Suppose that we have a set of mechanistic explanations specifying four mechanisms—e.g., a behavioural-organismic mechanism, a hippocampus-computational mechanism, a hippocampus-synapses mechanism, and molecular mechanism—accounting for a kind of categorisation judgement; say, the judgement of whether or not individual $c$ belongs in category $C$ in terms of similarity between the properties of $c$ and typical members of $C$. Now, if we compare this set of mechanistic explanations both before and after their integration (supposing that integration is possible), we find that following integration the set of explanations is more qualitatively parsimonious, because it explains the relevant kind of categorisation judgement by specifying only one kind of mechanism: a hierarchically organised mechanism that has the behavioural-organismic, hippocampus-computational, hippocampus-synapses, and molecular mechanisms as its parts.

Thus, I maintain that the virtue of greater qualitative parsimony is part and parcel of what makes the integration of mechanistic explanations valuable. Consequently, I argue that we can identify greater qualitative parsimony as a second theoretical virtue of the integration of mechanistic explanations; which is just to say that such integrations of mechanistic explanations have the virtue of being more qualitatively parsimonious than any un-integrated set of mechanistic explanations with equivalent explanatory power. It is possible to give formal rendering of this virtue as follows. First let $IM(x)$ stand for an integrated set of mechanistic explanations $x$ and let $UM(y)$ stand for a set of un-integrated mechanistic explanations $y$. Then let $EP(e, y)$ stand for a function that delivers the explanatory power of $y$ with respect to some explanandum or set of explananda $e$. Finally, let $QP(x, e)$ be a function which delivers the qualitative parsimony of $x$ with respect to its explanation of $e$, such that:

$$\forall x \forall y (IM(x) \wedge UM(y) \rightarrow \forall e (EP(e, x) \equiv EP(e, y) \rightarrow QP(x, e) > QP(y, e))) \quad (1)$$

### 3.3 Virtues of integrating mechanistic explanations

My claim is that the virtues of unification and greater qualitative parsimony are two of the theoretical virtues of integrations of mechanistic explanations (although there

are likely many others). Thus, I hold that each of these virtues can be appealed to as reasons for enacting an integration of mechanistic explanations in cognitive science. In fact, I think that any defence of integrations of mechanistic explanations will have to presuppose that such integrations possess the two virtues discussed in this section. In the next section, however, I introduce two different kinds of explanations—dynamicist and psychological respectively—in order to consider a different kind of explanatory integration entirely: cross-explanatory integration. Then, in the following section, I argue that no case of this kind of integration could hope to have the theoretical virtues of unification and greater qualitative parsimony. This leads to a discussion about whether or not a failure to possess these virtues is disqualifying for all kinds of explanatory integration in cognitive science.

## 4 Cross-explanatory integration

### 4.1 Dynamicist explanations

Dynamicist explanations posit variables that are "not low level (e.g., neural firing rates) but, rather, macroscopic quantities at roughly the level of the cognitive performance itself" (Van Gelder 1998, p. 619). As an example of dynamicist explanation, Chemero and Silberstein (2008) cite the HKB model of the dynamics involved in human bimanual coordination (that is, the model developed by Haken et al. (1985)). The HKB model "accounts for behavioral data collected when experimental subjects are instructed to repeatedly move their index fingers side to side in the transverse plane in time with a pacing metronome either in phase (simultaneous movements toward the midline of the body) or antiphase (simultaneous movements to the left or right of the body midline)" (Kaplan and Craver 2011, p. 614). To do this, the HKB model—as with all dynamicists explanations—introduces a differential equation, which describes the coupled dynamics of these cognitive performances:

$$\phi = -a \sin \phi - 2b \sin 2\phi \tag{2}$$

where "$\phi$ is the so-called collective variable representing the phase relationship (relative phase) between the two moving index fingers (when $\phi = 0$, the fingers are moving perfectly in phase), $a$ and $b$ are coupling parameters reflecting the experimentally observed finger oscillation frequencies, and the coupling ratio $b/a$ is a control parameter since relatively small changes in its value can have a large impact on system behavior" (Kaplan and Craver 2011, p. 614)

Dynamicist explanations were inspired by developments in the modelling of continuum systems. Modelling an object as a continuum involves assuming that the object is continuously distributed (e.g. non-discrete) and fills the entire region of space it occupies. Examples of objects that can be modelled as continuum include gases, liquids, crowds, and car traffic. Continuum mechanics relies on a number of governing equations, which account for "relations of dependency" in the system being modelled. For example, for sufficiently dense and relatively slow moving continuum (e.g. Newtonian fluids) the Navier-Stokes equations account for the linear relation of dependency

between stress and other pressures (e.g. gravity, inertial accelerations, etc.) with respect to the continuum's "flow velocity."[15] Dynamacist explanations do not explain why these dependencies hold, but do show how the behaviours of the relevant continuum systems depend on these dependencies. The HKB model, for instance, "exemplifies a law of coordination that has been found to be independent of the specifics of system structure" by "captur[ing] the coordination between behaving components of the same system" (Bressler and Kelso 2001, p. 28).

The central difference between mechanistic and dynamicist explanations concerns how they carry explanatory force. Everyone agrees that mechanistic explanations—like models in "lower-level" neuroscience—carry explanatory force "to the extent, and only to the extent, that they reveal (however dimly) aspects of the causal structure of a mechanism" (Kaplan and Craver 2011). Dynamicist explanations, in contrast, are taken to carry explanatory force not by respecting the underlying causal structures that give rise to system-level dynamics, but rather by characterising the behaviour of systems in terms of emergent or higher-level variables describing (changes to) the global state of the system (cf. Chemero and Silberstein 2008; Van Gelder 1995, 1998). That is, by representing cognitive states as points/regions in a state space and employing differential equation to account for cognitive processes as trajectories though that space. This allows dynamicist explanations to "abstract away from causal mechanical and aggregate micro-details to predict the qualitative behavior of a class of similar systems" (Chemero and Silberstein 2008, p. 12).

Integrating dynamicist and mechanistic explanations is a highly prized long-term goal for those who recognise both kinds of explanations. As an example of an attempt at this kind of cross-explanatory integration, consider the work of Bechtel (2008, 2011). In a series of papers, Bechtel argues that the best cognitive scientific explanations will introduce a continuum between fully decomposable (or highly modular) systems that are apt for mechanistic explanations and holistic, un-decomposable systems that are apt for dynamicist explanations (Bechtel 1998, 2008, 2011). The idea here is that cognition be thought of as a "functionally integrated system" with mechanistic parts (subsystems) that are constantly interacting and influencing one another in the form of dynamic feedback loops and other non-linearities. Thus, he claims that there will be a division of labour between mechanistic and dynamicist explanation reflecting the division between two explanatory tasks: explaining interactions *within* and *between* subsystems, and explaining the feedforward, feedback, and collateral connections that characterise the dynamic behaviour of the system as a whole.[16]

The problem with Bechtel's picture is that "it is by no means obvious how to link the output of modules to the relevant dynamical variables of the whole system" (Marraffa and Paternoster 2013, p. 34). While Bechtel claims that mechanistic explanation at the level of subsystem provide the foundation for dynamicist explanation, the two kinds of explanations still do independent explanatory work. For instance, mechanistic explanations explain interactions between and within subsystems; whereas dynamicist

[15] For further information about the Navier-Stokes equations and their role in continuum mechanics see Acheson (1990) and Smits (2000).

[16] The outcome of these two tasks can then be "tightly coupled together" as an integrated "dynamic mechanistic explanation" (DME) (cf. Bechtel and Abrahamsen 2010, for the canonical formulation of DME's).

explanations account for patterns of dynamic organisation characterising the total state of the cognitive system as a whole. As Marraffa and Paternoster (ibid., p. 34) point out, this means that Bechtel's cross-explanatory integration remains incomplete, because no account is given of how to connect the states and processes described by mechanistic explanations with the global states described by dynamicist explanations. Thus, Bechtel's attempted cross-explanatory integration seems to be nothing more than a 'tacking together' of mechanistic and dynamicist explanations.[17]

### 4.2 Psychological explanations

Psychological explanations are "defined in terms of the functional coupling of their components," which is neutral with respect to the physical (e.g. spatio-temporal) organisation of those components. (Weiskopf 2017). The central difference between mechanistic and psychological explanations, therefore, concerns how they capture the causal organisation of cognitive systems. In contrast to mechanistic explanations, psychological explanations are taken to capture the causal structure of a "relatively restricted aspect or subsystem of the total cognitive system" by employing "relatively few variables or factors" (e.g. representations and operations over representations). Moreover, they are taken to individuate their explanatory targets—their explananda—in a way that is neutral with respect to the physical structure of the system that realises them (ibid., pp. 10–11). The idea of psychological explanations, then, is that they explain by abstracting away from decomposable aspects of the biological and neural architecture to capture the "causal organization of a psychological system by representing it in terms of abstract relationships among functional components" (ibid., p. 37).

Aside from their neutrality with respect to the underlying physical, biological, and neural architecture, another distinctive feature of psychological explanations is their positing of contentful representations and interactions over these representations. Psychological explanations stipulate that the functional components represented must be intentionally interpreted states. In this way, psychological explanation is committed to "intentional internals" in the sense of Egan and Matthews (2006). Their account of how "cognitivist" explanation (read: psychological explanations) featuring intentional internals works runs as follows:

> The cognitive capacity to be explained—e.g., recovering the three dimensional structure of the scene, recognizing faces, understanding speech—is typically decomposed into a series of subtasks, each of which is itself characterized in

[17] Issad and Malaterre (2015) try to make sense of Bechtel's account by arguing that mechanistic explanations and dynamic mechanistic explanations can be subsumed under a new category of explanation: "Causally Interpreted Model Explanations" (CIME's). CIME's are taken to explain "neither in virtue of displaying a mechanism nor in virtue of providing a causal account, but in virtue of mathematically showing how the explanandum can be analytically or numerically derived from a model whose variables and functions can be causally interpreted" (ibid., p. 288). However, this forces Issad and Malaterre to admit that "supplying a causal-story is no longer seen as central in providing explanatory force" and so "providing a mechanism *per se* is also not so central when it comes to explanatory force" (ibid., p. 289). This view, then, does not seem like a case of cross-explanatory integration at all, but, rather, a reduction of mechanistic explanation to dynamicist explanation.

> intentional terms. The intentional internals posited by the cognitive theory are presumed to be distally interpretable, i.e., to represent such external objects and properties as the orientation of surfaces, facial features, spatial locations, etc. It is thought that if these intentional internals are not distally interpretable, then the account is unlikely to yield an explanation of the organism's successful interactions with its environment. Moreover, cognitive processes must preserve certain epistemic and semantic relations defined over these representations. The outputs of these processes should *make sense*, should be *rational*, given the inputs. This rich cognitive structure constrains theorizing at the lower levels. Cognitive theorists then look for computational and neural states to realize the intentional internals. The outcome, if things go well, will be a mapping between the causal structure of the mind and the causal structure of the brain (Egan and Matthews 2006, p. 382).

Psychological explanations may represent systems in terms of, e.g., verbal descriptions, diagrams and graphics, or computational models or simulations. Verbal descriptions give a rough characterisation of simple cognitive models. For example, to elaborate the levels of processing framework in memory modelling as in Cermak and Craik (1979). Diagrams or graphics—such as boxological models (see Fig. 2)—provide pictorial representations of the relationships between functional components, typically in terms of schematic representations of informational exchange. Such representations may help to make sense of the functional organisation of cognitive systems by providing a decomposition of the functional components responsible for cognitive behaviours. Computational models or simulations investigate "the implications of ideas, beyond the limits of human thinking"; that is, they "allow [for] the exploration of the implications of ideas that cannot be fully explored by thought alone" (McClelland 2009, p. 16). In all cases, however, the language (whether verbal or not) will be couched in representational terms and so will focus on the manipulation of intentional states without concern for the underlying physical structure.

Some who defend the explanatory role of psychological explanations argue for their autonomy from mechanistic explanations (cf. Fodor 1974). All, however, accept that a complete understanding of the mind/brain will involve "perfecting cognitive models and coordinating them with neurobiological ones" (Weiskopf 2017, p. 37). It is not yet clear what this "coordination" should look like, but we can suppose that it will entail a fitting together of mechanistic and psychological explanations to give us greater insight into a cognitive system's operation. Such an account would likely include a specification of the causal relation between structures at the level of neural-biology and intentional structures "that hover at some remove from the neural organization of the mind/brain" (ibid., p. 33). Such an integration would, therefore, connect different kinds of causal explanation to show how functionally characterised elements of psychological explanations relate to neural structures and processes.[18]

[18] Weiskopf (2017) makes as start on providing this taxonomy by subsuming both mechanistic and psychological explanations under a single kind of explanation: componential causal explanation.

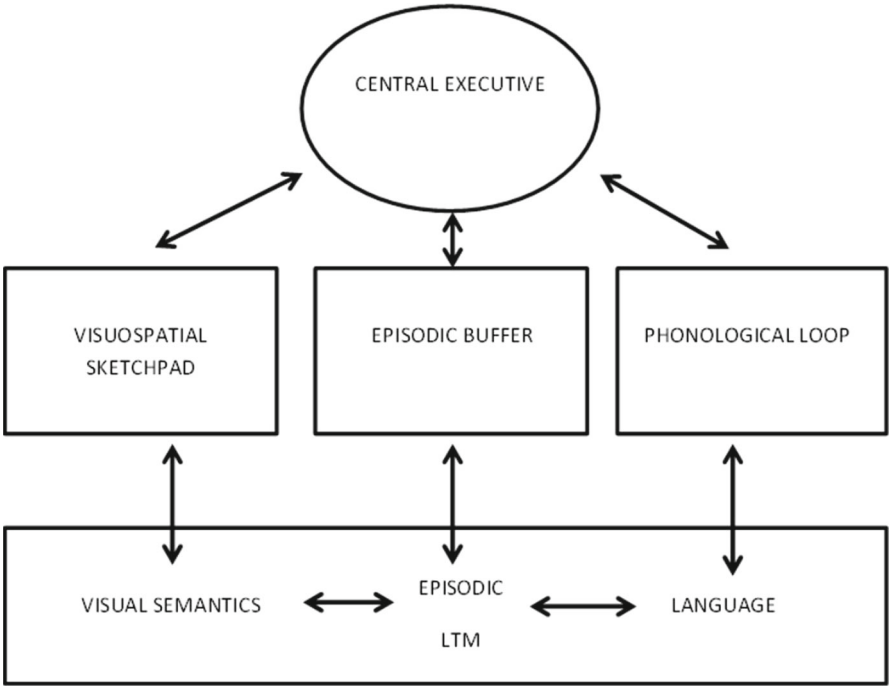


**Fig. 2** Model of working memory. Source: Baddeley (2000)

### 4.3 Cross-explanatory integration

Having now introduced dynamicist and psychological explanation, it is possible to give a brief definition of what I mean by cross-explanatory integration. A cross-explanatory integration will be any integration that somehow fits together two or more different kinds of explanation. In cognitive science, therefore, a cross-explanatory integration will somehow fit together mechanistic and dynamicist explanations (as Bechtel (2008) attempts to), mechanistic and psychological explanations (as Weiskopf (2017) alludes to), dynamicist and psychological explanations, or mechanistic, dynamicist, and psychological explanations. Of course, there are open questions about how such an integration is enacted, just as there are open questions about how any integration in cognitive science is enacted (Miłkowski 2016). I will not engage with these questions, but, rather, will consider whether or not cross-explanatory integrations can be said to be worthwhile at all.

## 5 Two views of cognitive scientific explanations

Kaplan and Craver (2011) argue that dynamicist explanations do not have a role to play in cognitive science. They defend "a mechanistic approach to thinking about explanation at all levels of explanation in neuroscience," by arguing that "Dynamical

models do not provide a separate kind of explanation subject to distinct norms," because "the explanatory force of dynamical models, to the extent that they have such force, inheres in their ability to reveal dynamic and organizational features of the behavior of a mechanism" (Kaplan and Craver 2011, p. 623). Similarly, Piccinini and Craver (2011) argue that "there is no functional analysis that is distinct and autonomous from mechanistic explanation because to describe an item functionally is, *ipso facto*, to describe its contribution to a mechanism." The upshot of this view is that there is no added benefit from recognising either dynamicist or psychological explanations in cognitive science, because insofar as such explanations are explanatory they are merely "elliptical or incomplete mechanistic explanation" (ibid.).

One consideration in favour of Craver, Kaplan, and Piccinini's view is that by introducing non-mechanistic explanations we are confronted with the problem of cross-explanatory integration. To some, cross-explanatory integrations seem futile, because they cannot have the virtues of unification and greater qualitative parsimony. This is case because cross-explanatory integrations must integrate explanations that postulate inconsistent kinds; e.g. spatio-temporally organised components and activities (mechanistic explanations), non-decomposable global states (dynamicist explanations), and non-spatio-temporally organised intentional internals (psychological explanations). Therefore, unification and greater qualitative parsimony will not be virtues of cross-explanatory integrations, because no reduction or subsumption of the kinds postulated is possible without impugning the explanatory postulates of at least one kind of explanation.

Thus, there seems to be a good case against non-mechanistic explanation: the fact that cross-explanatory integrations lack the virtues of integrations of mechanistic explanations. It is important to recognise, however, that only explanatory integrations of certain kinds of explanations will have the virtues of unification and greater qualitative parsimony. Mechanistic explanations are perfect in this respect, because the posit "mechanism" can be unified with other postulated mechanisms via subsumption in one superordinate kind of mechanism: a hierarchically organised mechanism. However, things become more difficult when integrating explanations that do not have straightforwardly subsumable postulates. In this way, the view which rejects cross-explanatory integrations because they lack the virtues of unification and greater qualitative parsimony appears biased towards a certain kind of explanation from the start. That is, explanations whose postulates are consistent and enough alike in kind.

The motivation for this bias, I think, is an attitude towards cognitive scientific explanation that is *fundamentalist* in Weiskopf's (2017) terms. Fundamentalists assume that the end-goal of explanation is the specification of one fundamental structure, which unifies and subsumes all other structures postulated in explanation. My claim is that those who take issue with cross-explanatory integration do so because they think they have identified a kind of explanation that specifies such a structure in cognitive science: mechanistic explanation. They think this because they assume that a hierarchically organised mechanism is all that is needed to make sense of the connection between, say, higher-level computational cognition and lower-level, implementational cognition. Notably, however, a fundamentalist attitude towards cognitive scientific explanation is not the only option on the table. According to an *anti-fundamentalist*

view, cognitive science is not in the business of specifying a single fundamental structure, but of capturing a variety of structures via a diversity of explanatory strategies.

Those who defend an exclusively mechanistic approach to cognitive scientific explanation are fundamentalists in the sense defined above. However, it is important to recognise that there is one good reason for supposing that this view is flawed: because mechanistic explanations do not seem to be able to do the work of explaining cognition on their own. To see why, Weiskopf (2017, p. 1) asks us to:

> [...] consider protein folding, a process which starts with a mostly linear native state of a polypeptide and terminates with a complexly structured geometric shape. There does not appear to be any mechanism of this process: for many proteins, given the initially generated polypeptide chain and relatively normal surrounding conditions, folding takes place automatically, under the constraints of certain basic principles of economy. The very structure of the chain itself plus this array of physical laws and constraints shapes the final outcome. This seems to be a case in which complex forms are produced not by mechanisms but by a combination of structures and the natural forces or tendencies that govern them.

Weiskopf's example demonstrates that not all aspects of cognition can be explained by mechanistic explanations, since the process by which a protein structure assumes its shape or conformation will determine whether or not it functions as, say, a structural element, a receptor, or and enzyme in a neuron (cf. Dill and MacCallum 2012; Sweeney et al. 2017). This, in turn, undermines the claim that hierarchically organised mechanisms are the fundamental structure of cognition, because it demonstrates that "All mechanistic explanations come to an end at some point, beyond which it becomes impossible to continue to find mechanisms to account for the behavior of a system's components" (Weiskopf 2017, p. 31). At this point, the "description of lower-level mechanisms would be irrelevant" and other explanatory strategies must be found (Machamer et al. 2000, p. 13).

A mechanistic fundamentalist can respond that "To accept as an explanation something that need not correspond with how the system is in fact implemented at lower levels is to accept that the explanations simply end at that point" (Piccinini and Craver 2011, p. 307). But this is just to say that explanations are only explanations if mechanistic; which is a claim that finds its justification in the doctrine of mechanistic fundamentalism itself. From an anti-fundamentalists perspective, this argument is circular and should be rejected out of hand. Accordingly, anti-fundamentalists will argue that different kinds of explanations should have the taxonomic autonomy to define their own range of entities, states, and processes as target explananda; and the explanatory autonomy to develop independently sufficient and adequate explanations of the target explananda they identify (Weiskopf 2017).

One's choice between fundamentalism and anti-fundamentalism will directly influence one's views about the legitimacy of different kinds of cognitive scientific explanations. If one sides with the fundamentalist, then one will assume that at the endpoint of cognitive scientific inquiry the distinctions between different kinds of explanations will be dissolved and a fundamental structure will be specified. For example, that we will come to recognise that as well as having mechanistic explanations of

neural structure and representational operations we are able to have mechanistic explanations of, say, psychological, social, and even ecological dimensions of cognition as well. On this view, ironing out the different perspectives instituted by the different kinds of explanations will be a matter of homogenising our explanatory practices to reflect the real ontological state of affairs.[19]

Conversely, if one sides with the anti-fundamentalist, then one will assume that different kinds of explanations are responsible for explaining different aspects of cognitive systems; e.g. neuronal cognition, psychological cognition, and environmentally-embedded cognition. On this view, cognitive systems are multi-dimensional, but not in the mechanistic sense where we have the embedding of lower-order mechanism within higher-order mechanisms. Rather, cognitive systems are such that one dimension (say, the dimension of functional states) cannot be reduced to another dimension (say, the dimensions of interactions between spatio-temporally organised components) even if what exists in one dimension is in some way dependent on what exists at another.[20] The different kinds of explanation would, then, each be tasked with explaining one of these dimensions on their own terms and according to their own standards of success.

Choosing between fundamentalism and anti-fundamentalist will influence the interpretation one gives to the predicates deployed in different kinds of explanations in cognitive science. All will agree that predicates—such as 'is a mechanism,' 'is an intentional internal,' or 'is a global state'—are deployed under the *assumption* that they designate genuine properties possessed by (some aspects of) cognitive systems. But if one takes a fundamentalist view, then one will deny that some of the predicates deployed designate genuine properties. For instance, one could deny that a predicate such as 'is an intentional internal' designates genuine properties or ague that it only designates causally operative and spatio-temporally organised properties, which could in fact be best designated by other predicates (e.g. 'is a mechanism'). The diametrically opposite view is that all predicates deployed in all kinds of explanations in cognitive science designate genuine properties possessed by (some aspects of) cognitive systems. If one takes this anti-fundamentalist view, then the different kinds of explanations in cognitive science are much more than mere methodological distinctions; they each explain a different ontological dimensions of cognitive systems with their own irreducible properties.

In summary, the fundamentalist argument that cross-explanatory integration is disqualified because it lacks virtues such as unification and greater qualitative parsimony is not decisive. Anti-fundamentalists need not assume that theoretical virtues are uniform across cognitive science, because they endorse the autonomy of different kinds of explanations. Fundamentalists, however, cannot share this view, because they will be convinced that the virtues of explanatory integration in cognitive science are indexed

[19] Note here that I have been discussing mechanistic fundamentalism in order to critically examine the claim that cross-explanatory integrations should be judged according to the standards of integrations of mechanistic explanations. However, one could equally espouse 'dynamicist fundamentalism' or 'psychological fundamentalism,' whereby the fundamental structure of cognition is, say, some un-decomposable system or a collection of functional/intentional states.

[20] This second view is analogous to the kind of "non-reductive" view endorsed in the philosophy of science/physics (cf. Poland 1994, for discussion about "non-reductive physicalism"). Thus, this view would entail a rejection of "crass scientistic reductionism" and the endorsement of the ontological autonomy of all dimensions of a cognitive systems recognised by cognitive scientific explanations (Heil 2003).

to the explanatory specification of one fundamental structure. From this perspective, explanatory legitimacy obtains only when we make progress in specifying a fundamental structure, which *ipso facto* mandates that all genuine explanations must be apt for specifying such a structure. The surest route to reaching such a specification is by ensuring that all cognitive scientific explanations postulate kinds that are, in principle, subsumable within a superordinate kind. Kinds, that is, like "mechanism." It is important to note, however, that the fundamentalist has not won the day yet. It follows that, at this time at least, any repudiation of cross-explanatory integration on the grounds that it lacks the virtues of unification and greater qualitative parsimony is premature. Another possibility is still available: that cross-explanatory integrations have virtues in their own right.

## 6 Virtues of cross-explanatory integration

### 6.1 Explanatory depth

Note first that those who accept cross-explanatory integration and endorse anti-fundamentalism will recognise diverse kinds of cognitive scientific explanations, but will also conceive of cognition as a somehow unified explanandum. If this were not true, then integration would be entirely without purpose. In line with this way of thinking, Weiskopf (2017, p. 14) argues that different kinds of explanations do not have a privileged evidential bases (whether neurophysiological, behavioural, introspective etc.); what is always being explained is our evidence for cognitive competencies, even if explanations differ in kind. For its advocates, therefore, cross-explanatory integration can be understood as an attempt to give a "multi-dimensional" explanation of the same thing—cognition—without prioritising one dimension or another.[21]

When we understand this point, we can recognise that for anti-fundamentalists cross-explanatory integrations will exhibit one theoretical virtue to a higher degree than integrations of one kind of explanation: *explanatory depth*. According to Keas (2018, p. 2766), an explanation exhibits explanatory depth "when it excels in causal history depth or in other depth measures such as the range of counterfactual questions that its law-like generalizations answer regarding the item being explained." Clearly, a cross-explanatory integration will not excel in causal history depth, because it may integrate dynamicist explanations which do not aim to capture causal structure at all.[22] However, anti-fundamentalists can argue that cross-explanatory integrations will exhibit a higher level of "law-focused" explanatory depth than will integrations of mechanistic explanations, because they allow for a greater "generality with respect to

[21] Multi-dimensional explanation should not be confused with multilevel explanation, since the idea of levels may be relevant from one perspective (mechanistic explanations), but not from another (dynamicist explanations).

[22] Keas (2018, p. 2766) says that "Causal history depth is often characterized in a causal-mechanical way by how far back in a linear or branching causal chain one is able to go." Evidently, then, this is not the kind of explanatory depth that cross-explanatory integrations could have as a virtue; so integrations of mechanistic explanations will definitely have the virtue of *causal history depth explanatory depth* to a higher degree than cross-explanatory integrations.

other possible properties of the very object or system that is the focus of explanation" (Hitchcock and Woodward 2003, p. 182).

The idea of "law-focused" explanatory depth is complex. Put simply, it holds that an "explanation is deeper insofar as it makes use of a generalization that is more general" (ibid., p. 181). Hitchcock and Woodward (2003, 182) argue that the "right sort of generality is generality with respect to other possible properties of the very object or system that is the focus of explanation." Such generality can be identified by undertaking "testing interventions," which probe the "counterfactual dependencies" of an object or system by intervening to manipulate—perhaps in an idealised way—the system's behaviour under various conditions. The counterfactual dependencies of an object or system, therefore, are just the manipulable dependencies that are constitutive of the behaviour of the system. To make this clear, consider Hitchcock and Woodward's helpful example:

> suppose that the height ($Y$) of a particular plant depends upon the amount of water ($X_1$) and fertilizer ($X_2$) it receives according to the following formula:
>
> $$Y = a_1 X_1 + a_2 X_2 + U \tag{3}$$
>
> where $U$ reflects unknown sources of error [...and...] for some change $\Delta X_1$ and $\Delta X_2$ [(3)] correctly 'predicts' that if $X_1$ and $X_2$ had been changed by those amounts, then the height of the plant would have changed by (approximately) the amount $a_1 \Delta X_1 + a_2 \Delta X_2$.
> [...]
> the low-level generalization [(3)] relating water and fertilizer to plant height strikes us as explanatory, but only minimally so: the explanations in which it participates are shallow and relatively unilluminating. If we had a theory—call it ($T$)—describing the physiological mechanisms governing plant growth it would provide deeper explanations. Such a theory would presumably be invariant under a wider range of changes and interventions than [(3)]; that is, we would expect ($T$) to continue to hold in circumstances in which the relationship between height, fertilizer and water departed from the linear relationship [(3)] (Hitchcock and Woodward 2003, pp. 183–184).

"Law-focused" explanatory depth, therefore, should be understood in terms of the "range of invariance of a generalization," where "Explanatory generalizations allow us to answer what-if-things-had-been different questions: they show us what the value of the explanandum variable depends upon" (ibid., p. 182). This idea was further elucidated by Keas (2018) by means of the following example:

> Newton's account of free fall possessed more explanatory depth than Galileo's. Newton explained not just free fall very near earth's surface (the restricted range of Galileo's theory), but also free fall toward earth starting from any distance. Furthermore Newton could explain free fall toward a hypothetically "altered earth"—perhaps if there is a change in its mass and radius, or if one works with another planet or a star that has such an alternative mass and radius. So the Newtonian explanation of free fall remains invariant through a larger range of

investigator interventions. In short, Newton's "free fall" account is explanatorily deeper than Galileo's because it handles a larger range of counterfactual (what-if-things-had-been-different) questions about the same kind of phenomena (free fall in various circumstances).

The multi-dimensional explanations engendered by cross-explanatory integrations will handle a range of counterfactual (what-if-things-had-been-different) questions about cognition. For example, counterfactual cases involving differences in neural structure (brought about, for instance, via brain lesions), differences in the global state of the cognitive systems (brought about, for instance, as the result of environmental contingencies), and differences in the intentional states of the cognitive system (brought about, for instance, as a result of the contingent availability of external objects to be represented). Supposing, then, that the "explanandum variable" for cognitive science is the set of cognitive behaviours—call it $C$—it is evident that cross-explanatory integrations will make use of a generalisation that is very general indeed: the generalisation that cognition is multi-dimensional.

Working from Hitchcock and Woodward's example, we can say that a multi-dimensional explanations engendered by cross-explanatory integration will have the following form:

$$C = a_1 M_1 + a_2 D_2 + a_3 P_3 \tag{4}$$

where $M_1$ is some components and activities specified by a mechanistic explanation, $D_2$ is some global state specified by a dynamicist explanation, and $P_3$ is some functional/intentional states specified by a psychological explanation; and for some change $\Delta M_1$, $\Delta D_2$, and/or $\Delta P_3$, (4) correctly 'predicts' that if $M_1$, $D_2$, and/or $P_3$ had been changed, then the behaviours of the cognitive systems would be different. Thus, (4) just says that explaining the explanandum 'cognitive behaviours' ($C$) depends on identifying some dependencies between whatever is explained by mechanistic, dynamicist, and psychological explanations.[23] From this perspective, it is clear that (4) will have greater explanatory depth than any exclusively mechanistic, dynamicist, or psychological explanation, because it will make use of a generalisation that is more general: that cognitive behaviours depend on the dependencies between the various dimensions explained by different kinds of cognitive scientific explanations.

A fundamentalist could respond that we have no good reason to think that the dependencies expressed by (4) are constitutive of the behaviour of cognitive systems. This amounts to the same thing as arguing that only one kind of explanation (typically, mechanistic explanation) is needed for a complete explanation of cognitive behaviours. But the anti-fundamentalist will deny that this fundamentalist response has force. And this denial is at least plausible, since it is obvious—at least from the perspective of folk psychology—that cognitive behaviours can be affected equally by changes to the components and activities involved in cognition (e.g. from the destruction or

[23] It is important to recognise that the nature of such dependencies is not necessarily linear. We should not expect a change to, say, $M_1$ to affect $D_2$ or $P_3$; just as we would not expect a change in the amount of water to affect the amount of fertiliser in Hitchcock and Woodward's example. This is true even if we would expect changes to either the amount of water or the amount of fertilizer to affect plant height; and if we would expect changes to whatever is explained by either mechanistic, dynamicist, or psychological explanations to affect cognitive behaviours.

degeneration of brain cells), the global state of cognition (e.g. in twin earth cases where environmental contingencies matter), or the functional/intentional states of cognition (e.g. when we rationally hold two distinct singular beliefs about one and the same object; say, Venus). For sure, difficult questions remain about how these dependencies work and about the scope of such dependencies. But trying to give an answer to these questions just is the reason for doing cognitive science in the first place.

### 6.2 Applicability

Applicability is a diachronic virtue in Keas' (2018, pp. 2780–2787) sense, which is to say that it "can only be instantiated as a theory is cultivated after its origin." Some—e.g. Strevens (2008)—think that "Successful scientific theories constitute knowledge of the world (knowing that), not control over the world (which is mainly knowing how) for practical (non-theoretical) purposes." I will not directly engage with Streven's arguments here. However, it is important to note that even he would not endorse the view that practical applicability *detracts* from the value of a theory, even if such applicability is said to depend on the understanding or knowledge that theory provides. In accord with this sentiment, Douglas (2014, p. 62) argues that "With the pure versus applied distinction removed, scientific progress can be defined in terms of the increased capacity to predict, control, manipulate, and intervene in various contexts." For the anti-fundamentalist, cross-explanatory integrations will be taken to increase this capacity to a greater extent than integrations of only one kind of explanation.

We are best able to understand the virtue of applicability by considering Keas' (2018, p. 2785) introduction of the virtue. He says:

> Applicability refers to when a theory is used to guide successful action (e.g., prepare for a natural disaster) or to enhance technological control (e.g., genetic engineering). High degrees of the virtue of applicability obtain when a theory that is used to guide such action or control provides more effective outcomes than what is possible in the absence of the theory.

The idea, then, is that cross-explanatory integrations will result in "theories" that are better placed to guide successful action or to enhance technological control than are the "theories" resulting from integrations of a single kind of explanation.

It is an open question how a pragmatic virtue like applicability relates to epistemic concerns about knowledge or understanding. Keas points to Agazzi's (2014) claim that:

> the existence of technological applications is the last decisive step that assures that [theories] have been able to adequately treat those aspects of reality they intended to treat.

This leads Agazzi to the conclusion that—with a mechanistic perspective assumed—theories:

> contain not only prescriptions as to the way of realising the structure of the machine but also as to its functioning. This functioning is something that hap-

> pens; it is a state of affairs that constitutes a confirmation of the theories used in projecting the machine Agazzi (2014, pp. 308–310).

Although Keas takes Agazzi's position to afford an "inflated epistemic role for applicability," he notes that this view is consistent with Hacking's (1983) loudly italicised argument that:

> *We are completely convinced of the reality of electrons when we regularly set out to build—and often enough succeed in building—new kinds of device that use various well understood causal properties of electrons to interfere in other more hypothetical parts of nature* (Hacking 1983, p. 265).

Whether or not concrete applications of cognitive scientific explanations really are planned in advance and have epistemic import, one thing is clear: anti-fundamentalists will take any technological innovation inspired by cross-explanatory integrations in cognitive science to be a validation of such integrations. And from the anti-fundamenatalist perspective there is good reason to think that cross-explanatory integrations will have a high degree of applicability with respect to technological innovations. This follows because cross-explanatory integrations will be taken to better explain cognitive behaviours by postulating a wider variety of "properties"—both causal and non-causal—of cognitive systems. And better explanations will, in turn, result in a more predictively powerful body of knowledge or understanding, which will be expected to guide more successful actions or to further enhance technological control (Vincenti 1990). This story, however, is only convincing if we can identify technological innovations inspired by cross-explanatory innovations in cognitive science.

But identifying such technological innovations does not seem far-fetched. In fact, one could argue that the evidence of the greater applicability of cross-explanatory integrations is already manifest. To underscore this point, consider the development of certain AI-robotics systems; for example, the development of self-steering robotics such as Tesla's cars equipped with autopilot systems. Systems such as these certainly do involve many parts and components, but also operate over functional/intentional states (for example, representations of locations) and will transition between global states according to governing equations accounting for relevant dependencies (for example, equations that account for the angle of turning as a relation of dependency between, say, speed, radial load, and axial load) (Yang et al. 2013). For the anti-fundamentalist, then, the successes of such technologies demonstrate that we should have sufficient confidence in the application of cross-explanatory integrations "as the basis for a new or improved technology" (Keas 2018, p. 2785).

When compared with integrations of a single kind of explanation, the fundamentalist will dispute the claim that cross-explanatory integrations are a *better* guide to successful action and are *better* able to enhance our technological control. Once again, however, this criticism is grounded in an assumption about the aim of cognitive scientific explanation—i.e. to specify one fundamental structure—and the failure of cross-explanatory integrations to contribute to that aim. It is clear that one will not think that cross-explanatory integrations can function as the basis for better technology if one also thinks that some of the putative explanations being integrated are not

explanatory at all. But since the anti-fundamentalist takes the opposite view—namely, that all kinds of explanations are explanatory and so cross-explanatory integrations deliver more predictively powerful bodies of knowledge—, we find once again that the deciding factor is the attitude one takes towards the explanatory ambitions of cognitive science. Given an anti-fundamentalist viewpoint, the applicability of cross-explanatory integrations will far outstrip the applicability of integrations of a single kind of explanation.

### 6.3 Virtues of cross-explanatory integration

The preceding discussion of the virtues of cross-explanatory integrations illustrates that such integrations can be taken to have some virtues to a greater extent than integrations of a single kind of explanation; e.g. explanatory depth and applicability. For sure, this claim depends upon the adoption of an anti-fundamentalist perspective, but there is no *a priori* reason that such a perspective could not be correct. Thus, I have shown that any evaluation of different kinds of explanatory integration in cognitive science in terms of their respective virtues will depend on the perspective one adopts towards the explanatory task of cognitive science. A more detailed study could be undertaken to show which other theoretical virtues align with which perspective. However, this first requires that we have agreement about which theoretical virtues exist and are relevant. This task is beyond the scope of this paper. It is enough, however, to have shown that the importance and weight of *some* theoretical virtues of explanatory integrations—e.g. unification, greater qualitative parsimony, explanatory depth, and applicability—will depend on one's views about what cognitive science works to explain.

## 7 Two kinds of explanatory integration in cognitive science

The debate about what is required from cognitive scientific explanation is long and convoluted. As Weiskopf (2017) points out, modelling cognition can involve various abstractions and idealisation as we, say, "neglect the brain's intricate internal organization and treat it simply as a suitably discretized homogeneous mass having certain energy demands (Gaohua and Kimura 2009)"; or focus on "detailed structural and dynamical properties" revealed by "the distribution of various neurotransmitter receptor sites (Zilles and Amunts 2009)." There is an open question, however, about how to reconcile these different explanatory programs and so integrate the various models of cognition into one coherent picture of the operation and organisation of the mind/brain. In this paper, I have argued that one's view about the scope of explanatory integration in cognitive science cannot be conveniently segregated from one's view about the explanatory task of cognitive science, because one's view about the explanatory task of cognitive science determines the theoretical virtues one favours.

The general thrust of this idea has been nicely formulated by Cat (2017) in his discussion of unification. He says:

> Philosophically, assumptions about unification help choose what sort of philosophical questions to pursue and what target areas to explore. For instance,

> fundamentalist assumptions typically lead one to address epistemological and metaphysical issues in terms of only results and interpretations of fundamental levels of disciplines. Assumptions of this sort help define what counts as scientific and shape scientistic or naturalized philosophical projects. In this sense, they determine, or at least strongly suggest, what relevant science carries authority in philosophical debate.

In much the same way, perspectives on explanatory integration are shaped and refined by reference to assumptions about what questions to pursue and what target areas to explore. But we must be careful not to transpose such assumptions into the meta-discussion about which kinds of explanatory integrations are to be admitted into cognitive science. That is, we must be vigilant against evaluations of explanatory integrations that are biased from the start.

One's view about the explanatory task of cognitive science will inform different theoretical frameworks with different theoretical virtues. In turn, these frameworks will inform different empirical hypothesis about the operation and organisation of cognition. Where a fundamentalist view is adopted, a theoretical framework will be introduced which informs the empirical hypothesis that a fundamental structure is responsible for cognition. From this perspective, cross-explanatory integrations seem devoid of theoretical virtues and so integrations of a single kind explanation will be preferred. However, where an anti-fundamentalist view is adopted, a theoretical framework will be introduced which informs the empirical hypothesis that a number of irreducible structures are responsible for cognition. From this perspective, cross-explanatory integrations will have some theoretical virtues to a greater extent than do integrations of a single kind explanation. Thus, we find that the virtues of different kinds of explanatory integration are, in an important sense, view-dependent.

The open question is whether or not we can arrive at a consensus view about the explanatory task of cognitive science. This paper leaves that question open. However, I want to stress that a researcher's choice between fundamentalist and anti-fundamentalist views will often be informed by a range of heuristic and ideological factors; for instance, their experience with and preference for distinct (mathematical, computational, or psychological) concepts and tools; or their institutional embedding and the degree to which they interact with colleagues sharing the same philosophical inclinations. No compelling arguments about how we should view the explanatory task of cognitive science can be made by appeal to such factors. More objective arguments in favour of one view or another must wrestle with difficult problems about the demarcation, reduction, and explanatory capacities of science. These problems have a difficult and fractious philosophical history and are not likely to be resolved any time soon. Thus, the project of determining the explanatory task of cognitive science is, for now, ongoing and uncertain. The upshot is that we must make space for two kinds of explanatory integrations in cognitive science.

**Acknowledgements** I would like to thank all of the anonymous reviewers for their comments, critique, and advice about how the paper could be improved. Thanks to Ruben Noorloos, René Baston, Gottfried Vosgerau, and Markus Schrenk for their constructive comments on earlier versions of the paper. In particular, thanks to Frances Egan for the helpful comments and guidance at the start of this project, and for invaluable discussions about the topic of cognitive scientific explanation and beyond. This work was funded by the

DFG (German Research Foundation) as part of the Collaborative Research Centre 991: The Structure of Representations in Language, Cognition, and Science.

## References


Acheson, D. J. (1990). *Elementary fluid dynamics: Oxford applied mathematics and computing science series*. Oxford: Oxford University Press.

Agazzi, E. (2014). *Scientific objectivity and its contexts*. Berlin: Springer.

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *Psychology of learning and motivation* (Vol. 8, pp. 47–89). New York: Academic Press.

Bechtel, W. (1998). Representations and cognitive explanations: Assessing the dynamicist's challenge in cognitive science. *Cognitive Science*, *22*(3), 295–318.

Bechtel, W. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. Abingdon: Taylor & Francis.

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

Bechtel, W. (2011). Mechanism and biological explanation. *Philosophy of Science*, *78*(4), 533–557.

Bechtel, W. (2013). From molecules to behavior and the clinic: Integration in chronobiology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *44*(4), 493–502.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*(2), 421–441.

Bechtel, W., & Abrahamsen, A. (2010). Dynamic mechanistic explanation: Computational modeling of circadian rhythms as an exemplar for cognitive science. *Studies in History and Philosophy of Science Part A*, *41*(3), 321–333.

Bechtel, W., & Richardson, R. (2010). *Discovering complexity: Decomposition and localization as strategies in scientific research*. Cambridge: MIT Press.

Bliss, T. V., & Collingridge, G. L. (1993). A synaptic model of memory: Long-term potentiation in the hippocampus. *Nature*, *361*(6407), 31.

Bogen, J. (2005). Regularities and causality; generalizations and causal explanations. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*(2), 397–420.

Bogen, J. (2008). Causally productive activities. *Studies in History and Philosophy of Science Part A*, *39*(1), 112–123.

Bressler, S. L., & Kelso, J. S. (2001). Cortical coordination dynamics and cognition. *Trends in Cognitive Sciences*, *5*(1), 26–36.

Cat, J. (2017). The unity of science. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (Fall 2017)*. Stanford: Metaphysics Research Lab, Stanford University.

Cermak, L. S., & Craik, F. I. (1979). *Levels of processing in human memory*. New Jersey: Lawrence Erlbaum.

Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge: MIT Press.

Chemero, A., & Silberstein, M. (2008). After the philosophy of mind: Replacing scholasticism with science. *Philosophy of Science*, *75*(1), 1–27.

Craver, C., & Tabery, J. (2017). Mechanisms in science. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (Spring 2017)*. Stanford: Metaphysics Research Lab, Stanford University.

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

Craver, C. F. (2005). Beyond reduction: Mechanisms, multifield integration and the unity of neuroscience. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*(2), 373–395.

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

Craver, C. F., & Kaplan, D. M. (2018). Are more details better? On the norms of completeness for mechanistic explanations. *The British Journal for the Philosophy of Science*,. https://doi.org/10.1093/bjps/axy015.

Derdikman, D., & Moser, E. I. (2010). A manifold of spatial maps in the brain. *Trends in Cognitive Science*, *14*(12), 561–569.
Dill, K. A., & MacCallum, J. L. (2012). The protein-folding problem, 50 years on. *Science*, *338*(6110), 1042–1046.
Douglas, H. (2014). Pure science and the problem of progress. *Studies in History and Philosophy of Science Part A*, *46*, 55–63.
Egan, F., & Matthews, R. J. (2006). Doing cognitive neuroscience: A third way. *Synthese*, *153*(3), 377–391.
Fodor, J. A. (1974). Special sciences. *Synthese*, *28*, 97–115.
Gaohua, L., & Kimura, H. (2009). A mathematical model of brain glucose homeostasis. *Theoretical Biology and Medical Modelling*, *6*(1), 26.
Glennan, S. (2009). Productivity, relevance and natural selection. *Biology & Philosophy*, *24*(3), 325–339.
Glennan, S. S. (1996). Mechanisms and the nature of causation. *Erkenntnis*, *44*(1), 49–71.
Hacking, I. (1983). *Representing and intervening*. Cambridge: Cambridge University Press.
Haken, H., Kelso, J. S., & Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics*, *51*(5), 347–356.
Heil, J. (2003). Levels of reality. *Ratio*, *16*(3), 205–221.
Hitchcock, C., & Woodward, J. (2003). Explanatory generalizations, part ii: Plumbing explanatory depth. *Noûs*, *37*(2), 181–199.
Horst, S. (2007). *Beyond reduction: Philosophy of mind and post-reductionist philosophy of science*. Oxford: Oxford University Press.
Issad, T., & Malaterre, C. (2015). Are dynamic mechanistic explanations still mechanistic? *Explanation in Biology*, *11*, 265–292.
Kaplan, D., & Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of Science*, *78*(4), 601–627.
Keas, M. N. (2018). Systematizing the theoretical virtues. *Synthese*, *195*(6), 2761–2793.
Knierim, J. J., & Neunuebel, J. P. (2016). Tracking the flow of hippocampal computation: Pattern separation, pattern completion, and attractor dynamics. *Neurobiology of Learning and Memory*, *129*, 38–49.
Lewis, D. (1973). *Counterfactuals*. Oxford: Basil Blackwell.
Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*(1), 1–25.
Mackonis, A. (2013). Inference to the best explanation, coherence and other explanatory virtues. *Synthese*, *190*(6), 975–995.
Marraffa, M., & Paternoster, A. (2013). Functions, levels, and mechanisms: Explanation in cognitive science and its problems. *Theory & Psychology*, *23*(1), 22–45.
McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, *1*(1), 11–38.
McDowell, J. (1996). *Mind and world*. Cambridge: Harvard University Press.
Miłkowski, M. (2013). *Explaining the computational mind*. Cambridge: Mit Press.
Miłkowski, M. (2016). Unification strategies in cognitive science. *Studies in Logic, Grammar and Rhetoric*, *48*(1), 13–33.
Newell, A. (1990). *Unified theories of cognition*. Cambridge: Harvard University Press.
Piccinini, G., & Craver, C. F. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, *183*(3), 283–311.
Poland, J. (1994). *Physicalism, the philosophical foundations*. Oxford: Oxford University Press.
Quine, W. V. O. (1963). On simple theories of a complex world. *Synthese*, *15*(1), 103–106.
Simon, H. A. (1996). *The sciences of the artificial*. Cambridge: MIT press.
Smits, A. J. (2000). *A physical introduction to fluid mechanics*. Hoboken: Wiley.
Sober, E. (1994). *From a biological point of view: Essays in evolutionary philosophy*. Cambridge: Cambridge University Press.
Sober, E. (2015). *Ockham's razors*. Cambridge: Cambridge University Press.
Strevens, M. (2008). *Depth: An account of scientific explanation*. Cambridge: Harvard University Press.
Sweeney, P., Park, H., Baumann, M., Dunlop, J., Frydman, J., Kopito, R., et al. (2017). Protein misfolding in neurodegenerative diseases: Implications and strategies. *Translational Neurodegeneration*, *6*(1), 6.
Thagard, P. (1978). The best explanation: Criteria for theory choice. *The Journal of Philosophy*, *75*(2), 76–92.
Thagard, P. (2007). Coherence, truth, and the development of scientific knowledge. *Philosophy of Science*, *74*(1), 28–47.

Van Gelder, T. (1995). What might cognition be, if not computation? *The Journal of Philosophy*, *92*(7), 345–381.
Van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences*, *21*(5), 615–628.
Varela, F., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge: MIT Press.
Vincenti, W. G. (1990). *What engineers know and how they know it*. Baltimore: Johns Hopkins University Press.
Votsis, I. (2015). Unification: Not just a thing of beauty. *THEORIA. International Journal for Theory, History and Foundations of Science*, *30*(1), 97–114.
Weiskopf, D. A. (2017). The explanatory autonomy of cognitive models. In D. M. Kaplan (Ed.), *Explanation and integration in mind and brain science* (pp. 44–69). New York: Oxford University Press.
Wimsatt, W. C. (1997). Aggregativity: Reductive heuristics for finding emergence. *Philosophy of Science*, *64*, S372–S384.
Yang, S., Lu, Y., & Li, S. (2013). An overview on vehicle dynamics. *International Journal of Dynamics and Control*, *1*(4), 385–395.
Zilles, K., & Amunts, K. (2009). Receptor mapping: Architecture of the human cerebral cortex. *Current Opinion in Neurology*, *22*(4), 331–339.