
THE UNIVERSITY
*of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

# Wide Computation:

## A Mechanistic Account

Luke Kersten


PhD in Philosophy

University of Edinburgh

2020

# Contents

# Declaration of Authorship

I, Luke Kersten, declare that this thesis is my own work and has not been submitted for any other degree or professional qualification.

Luke Kersten

4 August 2020

# Notes on Publication

Parts of this thesis have been published, or are forthcoming, in the following articles:

(Chapter 1). Kersten, L. (2017). A Mechanistic Account of Wide Computationalism. *Review of Psychology and Philosophy*, 8(3): 501-517. doi: 10.1007/s13164-016-0322-3.

(Chapter 3). Kersten, L., Dewhurst, J., & Deane, G. (2017). Resolving two tensions in 4E cognition using wide computationalism. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of Cognitive Science Society* (pp.2395-2400). Austin, TX: Cognitive Science Society.

(Chapter 4). Kersten, L. (2020). How to be concrete: mechanistic computation and the abstraction problem. *Philosophical Explorations*, 23(3): 251-266. doi: 10.1080/13869795.2020.1799664.

# Abstract

This Ph.D. thesis explores a novel way of thinking about computation in cognitive science. It argues for what I call 'the mechanistic account of wide computationalism', or simply *wide mechanistic computation*.

The key claim is that some cognitive and perceptual abilities are produced by or are the result of computational mechanisms that are, in part, located outside the individual; that computational systems, the ones that form the proper units of analysis in cognitive science, are particular types of functional mechanisms that, on occasion, spread out across brain, body, and world.

Wide mechanistic computation is the result of bringing together two distinct strands of thinking about computation: (i) 'wide' views, which hold that computational systems can, on occasion, include parts of the surrounding environment; and (ii) 'mechanistic' views, which hold that computational explanation is a species of mechanistic explanation, and that computational mechanisms are a special type of functional mechanism.

I argue that wide mechanistic computation draws support from several sources. First, I examine research on animal and human psychology and show that several organisms' behaviours are properly treated as being the output of wide computational mechanisms. Second, I defend the view from several philosophical charges, including worries about its explanatory parsimony and empirical testability. Finally, I argue for the view's theoretical credentials by showing that it can help resolve specific problems that have recently troubled 4E cognition. The result is an argument for not only the coherence but also empirical plausibility of wide mechanistic computation.

On route to its main objective, the thesis also accomplishes a number of related tasks, including: (i) providing a framework for organising and conceptualising different views of

computation, (ii) securing the conceptual foundations of mechanistic computation by addressing an outstanding challenge called the 'abstraction problem', (iii) sounding a cautionary note about recent predictive processing accounts of extended cognition and (iv) arguing against a particular conception of levels often used within cognitive science, what is labelled the 'hierarchical correspondence view of levels'.

# Lay Summary

The concept of computation is central to the interdisciplinary study of the mind, what is known as cognitive science. This is because computation offers a method for explaining how purely physical systems, such as humans, can perform extremely complex tasks and activities, such as acquiring language. The guiding idea is that, like the familiar digital computer, humans also perform some form of 'mental' computation. This Ph.D. thesis explores a new way of thinking about computation in cognitive science.

There are two key ideas that underpin the thesis. The first is that the concept of computation, when used in cognitive science, applies not only to individuals, as is sometimes supposed, but also to how individuals interact with their environments. This is known as the 'wide' theory of computation. The second idea is that the concept of a mechanism is key to figuring out when a physical system is computing. So, for example, to explain why brains, laptops and calculators are all computing systems, the mechanistic view says that they are all certain types of mechanisms – that is to say, they are all systems which are organised so as to compute some function, such as addition or multiplication.

When we combine these two ideas, we get a novel approach to computation. This is the view that some cognitive abilities, such as our ability to navigate spaces, are produced by or are the result of computational mechanisms that are, in part, located outside the individual. I call this the 'mechanistic account of wide computationalism', or wide mechanistic computation for short – 'wide' because it claims that some computational systems include parts of the environment and 'mechanistic' because it focuses on the role of mechanisms in computation.

To support this view, I look at research in animal and human psychology. What this research shows, I suggest, is that in order to explain various organisms' behaviours it makes

sense to treat such behaviours as if they were caused by computational mechanisms spread out between brain, body and world. So, for example, I look at research on bat echolocation and argue that the bat's sensory system is often so tightly coupled with its environment that it actually satisfies several of the requirements we would normally have for digital computing.

Additionally, I also apply my view to a specific debate within a sub-area of cognitive science called 4E cognition. 4E cognition is an area of research that focuses on the unique way the body and world shape our minds – the 4Es stand for embodied, embedded, extended, and enactive. My aim is to resolve some specific tensions that are thought to arise from trying to unify the different Es under one banner. The goal of this discussion is not only to help out those sympathetic to 4E cognition but also demonstrate the wider usefulness of my view within philosophy and cognitive science.

Finally, the thesis examines three related issues in cognitive science. The first is a particular problem that arises for the mechanistic view of computation because of some of its core assumptions. In resolving this problem, I attempt to secure the foundations on which wide mechanistic computation is built. The second is an examination of two recent proposals that try to connect 'predictive processing' views of cognition, the idea that the mind/brain is a multi-layered prediction machine, and 'extended' views of cognition, the idea that cognitive systems and processes extend beyond the boundary of the individual. I take a critical view of these proposals, arguing that neither successfully marries the two views. The third is an examination of a particular application of the concept of 'levels' within cognitive science. So, for example, it is quite common to say that cognition can be analysed at three different 'levels', e.g., the computational, the algorithmic, and the implementational. I argue that there are problems with how the concept of levels is sometimes developed.

The key contribution of this thesis is that it not only combines two important ideas about computation but it also reshapes how we should think about applying the concept of computation within cognitive science.

# Acknowledgements

There are a number of people to which this thesis owes a considerable debt.

First and foremost, to my supervisors Mark Sprevak, Dave Ward and, at the beginning, Jesper Kallestrup, I owe an enormous debt of gratitude. Each has put their stamp on the thesis in their own unique way, and both the content and my thinking has been made the better for it. This thesis would not be what it is if not for their thoughtful and generous contributions.

Many thanks are also owed to the post-graduate students at the University of Edinburgh. To my fellow PhD students, Anna-Katrina Page, George Deane, Giles Howdle, Fausto Carcassis, James Brown, Joe Dewhurst, Johnny Lee, and Nina Poth, I owe a significant intellectual and personal debit. Their willingness to engage with me on an any and all topics, philosophical or otherwise, helped to foster a rich and stimulating environment in which to work.

From my life outside philosophy, I owe a large debt to Chris van Wyk, Chris Tung and the Edinburgh Ice Hockey Team. My time in Edinburgh would have not been as exciting or enjoyable without this colourful cast of characters.

A special thanks is also owed to the University of Edinburgh, and the staff and professors of the School of Philosophy, Psychology and Languages Sciences. Without their generous academic and financial support, I would not have had the many academic and personal opportunities that I did.

To my family, particularly Dianna, Jeff, Matthew, and Simon, I want to express my sincerest gratitude. Their unwavering support throughout my studies has always been a constant source of confidence and encouragement.

And finally, I would like to express my deep appreciation and affection to Katie Hay. Without her lively company, my life in Edinburgh would not have been as rich and exciting as it was.

# Introduction

Computational thinking has been at heart of cognitive science for some time now. Comprising a diverse array of techniques and methods, computational approaches have touched on everything from language learning and visual processing to skill acquisition and problem solving. The computational sciences have helped to expand and deepen our understanding of the mind, offering an unprecedented level of detail and precision to researchers.

While often driven by technical developments in computer science and engineering, philosophers, in their own unique way, have sought to contribute to advances in computational thinking. In large part, this has been done by clarifying foundational concepts, questioning orthodox thinking, and suggesting novel avenues for future research. Philosophical investigation has often sought to push computational thinking in new and interesting directions by teasing out important assumptions and charting fruitful links between different approaches.

To take one illustrative example, it was sometimes suggested that certain features of mentality, such as productivity (our ability to produce novel strings of sentences from a finite set of elements) and compositionality (our ability to combine and recombine syntactic structures with fixed contents), could only be explained by appealing to rules operating over proposition-like structures (Fodor and Pylyshyn, 1988). The computational implication was that only models appealing to language-like 'rules' and 'symbols' were said to be capable of explaining how and why human thought was recursive, compositional and inferential, usually referred to as 'symbolic' or 'classical' models.

In opposition to such views, a number of philosophical interlocutors, in step with technical developments in modelling, made in-roads on behalf of connectionist approaches to computation. Clark (1993), for instance, argued that the whole framing of the problem was

coloured by symbolic thinking. The question of productivity and compositionality should not be seen as one of trying to discover the inner structure of fixed mental contents, but rather as one of trying to understand how the representational elements of thought change over time to form such content, a role better suited to the process-oriented approach of connectionism.[1] While only one brief example from the history of cognitive science, such episodes are illustrative of the important role philosophers can have in thinking about computation.

In its own modest way, the current thesis attempts to make a similar contribution. My goal is to advance a novel vision of computation by pulling together two distinct strands of computational thinking.

The first is what I call 'wide' views of computation, or *wide computation* for short (Wilson, 1994, 1995; Losonksy, 1995; Hutchins, 1995; Wells, 1998).[2] These views hold that computational systems, the ones that form the basis of cognitive systems, can, on occasion, include parts of the surrounding or immediate environment. These accounts emphasise the location neutrality of computational analysis, and make much of the tight causal coupling between informationally rich environments and the internal functional machinery of individuals. As we will see, these views draw on a wide range of examples, everything from animal cognition and developmental psychology to Turing machines and ship navigation.

The second is what has more recently come to be known as 'mechanistic' views of computation (Piccinini, 2007, 2015; Fresco, 2014; Milkowski, 2013, 2015; and Dewhurst, 2018). According to these views, computational explanation is a species of mechanistic explanation, and computational mechanisms a special type of functional mechanism – a

[1] In a similar spirit, Dawson (1998) argued that when seen through Marr's three levels, connectionist and symbolic models largely turned out to be explanatorily equivalent. Connectionist networks, when properly interpreted, he argued were capable of accomplishing computational level achievements akin to those of symbolic models, and admitted of interesting algorithmic level analysis – in one such case, for example, 'wire-tapping' methods were used to reveal several inference rules used by networks to solve logic problems, such as *modus ponens* or *modus tollens*.

[2] The moniker originally comes from Wilson (1994). While I adopt Wilson's term in this thesis, several others would have been just as serviceable, including Losonsky's (1995) 'computational environmentalism' or Shagrir's (forthcoming) 'computational externalism'. I stick with Wilson's mostly out of loyalty, as it was the first one I came across, but also for consistency.

functional mechanism is an organised system of component parts, whose joint activity, when properly operating, constitute the capacity of the mechanism. Physical systems are said to implement computations when they involve functional mechanisms processing medium-independent vehicles.[3]

In bringing these two strands of computational thinking together, I aim to address the following question:

*What are the boundaries of computational mechanisms?*

I call this the 'boundary question'. Its answer, I submit, is that computational systems, the ones that form the proper units of analysis in cognitive science, are certain types of distributed or world-spanning mechanisms; mechanisms which process medium-independent vehicles in accordance with at least one abstract rule such that, on occasion, they extend into the surrounding or immediate environment.

The key insight is that computational analysis not only applies to distributed or world-spanning mechanisms but that such a conclusion follows from the internal logic of computational thinking. If computational analysis is, at core, location neutral, then it applies, in principle, equally to distributed or world-spanning mechanisms as it does to individual-bound mechanisms. I label this 'the mechanistic account of wide computationalism', or *wide mechanistic computation* for short.

But why answer such a question? Why worry about the boundaries of computational mechanisms?

One reason is that it has implications for how we think about computational theories of mind more generally. If psychological processes and states are computational in structure, then any re-drawing of the boundaries of computational mechanisms has knock on effects for the boundaries of psychological states and processes. As Wilson and Clark (2009) succinctly put

[3] I will have more to say about the term 'medium-independence' in Chapter 4, but suffice it to say, it hasn't always been the clearest what mechanists mean by this term.

the point: “If the kinds of computation that at least parts of cognition involve are extended, then those parts of or aspects to cognition will also be extended” (p.61). Depending on how we answer the boundary question, the nature and location of psychological states and processes might also change.

Another reason is that our response to the boundary question may have consequences for how we think about the relation between different views of cognition. For example, to anticipate the discussion of Chapter 3, enactive and extended views have sometimes been thought to offer competing accounts of cognition (Clark, 2008; Clark and Kiverstein, 2009). Extended views, it is claimed, construe cognition as a fundamentally representational and functional process, one which is indifferent to realising medium. Enactivist views, on the other hand, conceive of cognition as a dynamic, entity-bound sense-making activity; cognition is viewed as only one point on life-mind continuum (Thompson, 2007; Di Paolo & Thompson, 2014). When sat opposite, these views appear to pull in opposite directions. If sense-making is an entity-bound process, then it is unclear how it can also be extended into the environment. However, if, as I later argue, extended cognitive systems turn out to be wide computational systems, and wide computational systems are autonomous systems – a central concept of the enactivist views – then there may be room for both extended and enactive views under the 4E umbrella. Worrying about the boundaries of computational mechanisms may have consequences for how various views of cognition fit together.

Finally, a loftier but no less important reason to pursue the boundary question is its intrinsic interest to cognitive science. Computational thinking is, if not the only, then at least one of the most exciting games in town. Exploring its conceptual commitments seem as valuable a philosophical task as any. It offers a chance, as Wilfred Sellars (1963) famously says, to see how “things in the broadest possible sense of the term hang together in the broadest possible sense of the term.”

Yet even granting the coherence of the boundary question, there is still a question as to why combine wide and mechanistic views of computation. Why put these particular strands of computational thinking together? There are two reasons to do so, I think.

The first is that it offers a chance to resolve a number of specific issues that trouble previous formulations of wide computation. For example, as we will see in Chapter 1, traditional accounts of wide computation often rely on 'causal mapping' accounts of implementation (Chrisley, 1995; Chalmers, 1994; Scheutz, 2001).[4] These views maintain that physical systems compute when there are isomorphic mappings between states of the physical system and states of a computational model. The trouble is that wide accounts inherit the problems of causal mapping accounts from this alliance.

One such trouble, for example, is that causal mapping accounts are unable to adequately taxonomise computing systems. In virtue of underspecifying the mapping conditions between computational and physical systems, causal mapping accounts often fail to exclude non-paradigmatic cases from their taxonomies, such as solar systems or grain sieves (Milkowski 2013; Piccinini, 2015). Simply put, they are too liberal. Insofar as we want our account of computational implementation to be extensionally adequate, wide views appear guilty by association.

A second reason is that pursuing such a union allows wide computation to incorporate a number of advantages unique to mechanistic views of computation. The first, as previously hinted at, is their ability to satisfy several desiderata on a theory of implementation. For example, unlike their more liberal, causal mapping cousins, mechanistic views are able to adequately taxonomise different computing systems. While digestive systems and solar systems may exhibit certain degrees of causal organisation, they nonetheless fail to compute an abstract function in virtue of manipulating medium-independent vehicles. Mechanistic views place more stringent conditions on computational implementation. This allows them to

[4] Computational implementation refers to the conditions under which it is true or false to say of a physical system that it computes.

simultaneously exclude non-paradigmatic cases, such as solar systems or grain sieves, while including more paradigmatically classic cases, such as Turing machines or digital computers.

Another desirable feature is their ability to handle cases of *miscomputation* – that is, instances of where computing systems fail to fulfil their function. These can be explained, according to Piccinini (2015), by either appealing to a breakdown in the mechanism's component parts (i.e. hardware failures) or by describing how errors in the design specification of the mechanism's program result in execution errors (i.e. software errors). Explaining why computing goes wrong involves describing the functional or organisational structure of the underlying mechanism. Of course, these are only two desiderata amongst others, but they provide a sense of the strength of incorporating mechanistic talk into an account of computational implementation.

In addition to implementation, integration of the two views also provides an opportunity to incorporate a number of useful concepts from mechanistic thinking. Because mechanistic views of computation borrow a good deal of their explanatory apparatus from mechanistic thinking, they inherit a number of notions, such as constitutive explanation, functional analysis, and functional contextualisation, which might be a boon to wide computation. In particular, as I argue in Chapter 2, such notions offer useful tools for extending the scope and substance of wide computation. One reason for this is that, at least traditionally, wide accounts have tended to limitedly focus on implementational stories; that is, they have tended to focus on showing how a particular phenomenon satisfies a particular theory of implementation, e.g., the causal mapping account's. We will see some examples of this in Chapter 1. In connecting wide computation to mechanistic thinking, the aim is to gain access to a suite of conceptual tools that help flesh out the view.

A final, more general consideration in favour of the project is that it offers a chance to chart the logical space of thinking on computation. As I suggest later, given our interest in explaining the mind computationally, it seems useful to have the full array of positions available for consideration. While, for various reasons, talk of wide computation has dropped out of the

conversation lately, it nonetheless constitutes a distinct and interesting approach to computation. On route to articulating an updated version of wide computation, there is an opportunity to map the general space of views on computation.

## A Potted History

I want to provide a bit of background about wide computation at this point. I do this for two reasons: first, given the central role of wide computation in the thesis, it will help to have a clearer sense of what the view amounts to; second, it should provide a bit more insight into some of the view's core commitments and motivations. I consider here what I take to be the four main formulations of the view.[5]

The first two I consider come from Rob Wilson (1994) and Michael Losonksy (1995). Despite being formulated independently of one another, Wilson and Losonksy's views share a number of striking similarities. Both views, for example, are formulated in response to *individualism*, the view that the mind is best understood and studied in isolation from the environment.[6] Both views attempt to undermine what they see as an illicit connection often made between individualism and computational theories of mind.

Fodor's (1980, 1987) 'formality condition' and his 'methodological individualism', in particular, receives both authors' ire. The formality condition says that mental processes are formal processes driven by the syntactic or physical properties of mental representations. Since referential relations do not make a difference to the syntactic or physical properties of mental states, mental operations only operate on the internal states of individuals. The formality condition implies that mental states and processes operate without regard to external features of the environment.

[5] There are, scattered throughout the literature on computation, other gestures in the direction of wide computation, such as Clark (1989) and Rowlands (1999). However, for present purposes, I focus on only those treatments which are the most explicit and sustained in their development of the view.

[6] A classic statement of individualism can be found in Stephen Stich's 'principle of autonomy', which says that "[h]istorical and environmental facts will be psychologically relevant only when they influence an organism's current, internal physical state" (1983, p.165).

Wilson and Losonksy attempt to sever the link between the formality condition and individualism by showing that at least some computational systems include elements outside the individual. The importance of this is that "[i]f this were so, then the computational states of such a cognitive system would not supervene on the intrinsic, physical states of the individual; likewise, the resulting computational psychology would involve essential reference to the environment beyond the individual" (Wilson, 1994, p.352). The existence of wide computational systems demonstrates that the formality of cognition does not entail individualism.

To motivate the existence of wide systems, Wilson and Losonksy draw on a range of examples from cognitive psychology. Wilson appeals to research on form and spatial perception by Seluker and Blake (1990) and Gallistel (1989), while Losonksy appeals to work on infant and child development by Karmiloff-Smith (1992) and Rutkowska (1993).[7] The first set of studies involve cases of perceptual sensitivity to formal features of the environment – for example, in the case of form perception, these include certain parameters in the visual pathway, such as spatial frequency, contrast, and spatial phase. The second set of studies explore developmental behaviours, such as grasping or problem solving. These behaviours involve tightly integrated action-perception cycles. Such subject-environment feedback loops highlight the important ways in which representational thinking can be scaffolded onto environmental structures, and how action systems can help to simplify and extend computational abilities during development.[8]

What these studies suggest to Wilson and Losonksy is that there is no necessary distinction within computational analysis between the contributions of parts internal and external to the individual. As Wilson (1994) puts the point: "within a wide computational system much of the processing that takes place may well be instantiated fully within the boundary of the individual, but what makes it a wide system is that not all of the computational processes that

[7] In latter work, Wilson appeals to Baldwin's (1994) theory of animate vision.
[8] I will have say more about each of these examples in Chapters 1 and 2.

make the up the system are so instantiated" (p.353). Losonksy (1995) draws a similar lesson: "we can no longer treat behavior such as grasping or fixating on an object as a molar entity without an internal structure. Instead, we will have to analyze it into sequences of steps that involve internal as well as environmental structures" (p.364). For both authors, there is no important link between the internal states and activities of an individual and the resulting location of the computational system as a whole.

The third view I will consider is Edward Hutchins' (1995). Hutchins' aim is to broaden the unit of cognitive analysis from the individual to the group. He does so in order to combat what he calls the 'attribution problem', the tendency among researchers to over ascribe complex representations and information processing to individuals (pp.355-6).

Hutchins grounds his account in the approach of David Marr (1982). For Marr, cognitive analysis operates at three distinct levels: the computational, algorithmic and implementational. The computational level addresses the function or task being performed by a system; it answers 'what'-questions. The algorithmic level addresses the procedures and representations by which a system carries out its function or task; it answers 'how'-questions. Finally, the implementational level addresses physical instantiation; it addresses how a system is realised in a given medium.

Hutchins attempts to show that group level activities, such as ship navigation, are properly treated as distributed at each of Marr's three levels. He anchors his account on several case studies of ship piloting from Western and Micronesian naval traditions. Such activities, he argues, involve the use of specific representations and algorithms, implemented in an array of socio-technological materials, to accomplish specific computational tasks, such as dead reckoning.

Many of the computational tasks associated with navigation, Hutchins argues, are not reducible to the actions or activities of individual crew members, such as the fathometer operator or pelorus operator. Rather, they are the result of the coordinated, de-centralised activities of the navigation team as a whole. Even unskilled members, when embedded in the

right social organisation, can learn and contribute to computationally significant activities. As Hutchins (1995) writes:

> [O]rganised groups may have cognitive properties that differ from those of the individual who constitute the group. These differences arise from both the effects of interactions with technology and the effects of a social distribution of cognitive labour. The system formed by the navigation team can be thought of as computation in which social organisation is computational architecture. (p.228).

The final view I wish to consider is A. J. Wells' (1998). Wells offers a slightly different approach to wide computation. Rather than arguing via empirical research in cognitive psychology or anthology, Wells opts to approach wide computation via the concept of Turing machines.

To be specific, Wells is interested in arguing against what he calls the 'internalist' interpretation of Turing machines. The internalist interpretation is what has largely been espoused by classic computational theorists, such as Newell and Simon (1976), Pylyshyn (1984), and Fodor (1987). The internalist interpretation says that the principal parts of the Turing machines (the finite state controller/central and memory/tape mechanism) are instantiated squarely inside the individual. He writes, for instance: "[s]ymbol systems theorists pay particular attention to the memory/tape mechanisms because they are hypothesized to contain structured symbolic representations of the world. The manipulation and transformation of these representations constitute cognitive activity" (p.271).

Against the backdrop of the internalist picture, Wells suggests an alternative. While the finite state controller/central processor may be instantiated in the brain, Wells suggest that the memory/tape mechanism should be re-envisaged as residing in the environment. According to this 'interactive interpretation', cognitive computations are the structured interaction of a control architecture inside the individual and an input architecture outside the individual. As Wells puts it: "[c]ognitive computation, in other words, is irreducibly world involving; the cognizer is embedded in an environment which constitutes part of the cognitive architecture" (1998, p.280).

Wells suggests a number reasons for moving to the interactive interpretation. One is that it resolves thorny evolutionary questions about how a computational system could undergo structural change while retaining flexibility. Programmability, a key feature of the internalist picture, imposes a constraint on the structure of computational systems, one that doesn't fit neatly with the modifiability requirement imposed by evolution. The interactive interpretation solves this problem by giving up on the assumption that the structural architecture of cognitive systems has to be the same as that of digital computers. Because of this, structural change is explained as a function of the evolving interaction between input and control architectures; the internal control architecture is not imagined as general purpose, and so, in principle, it is evolvable with respect to its input environment.

Another reason is that it side-steps difficulties about transduction. The internalist interpretation has to explain how sensory transducers filter only the cognitively relevant information into cognitive systems. The interactive interpretation has no such problem. Because it gives up on the idea of internal symbolic encodings, the "function of a representation may have more to do with modifying the system's sensitivity to a class of inputs rather than storing a detailed inventory of its features" (p.280).

I think a number of interesting points emerge from the preceding discussion, so it will be worth consolidating what has been said so far.

First, the different formulations make it clear that wide computation has been motivated by a variety of problems and issues over the years. Whether it is abstract debates about methodological assumptions in psychology or technical debates about specific computational formalisms, such as Turing machines, a number of different considerations have motivated thinking on wide computation. There is no 'one' motivation for wide computation.

Second, the various views show that wide computational thinking has been applied to a number of different types of entities, at two distinct levels of analysis. While the majority of views have focused on individuals and their surrounding environment, wide computational analysis has also been applied to artificial agents and groups of individuals (Losonksy, 1995,

pp.365-9). The scope of wide analysis is not restricted to individual-world hybrids. It can, in principle, apply to any phenomena that exhibit the requisite formal structure.

Third, the views demonstrate that wide computation garners support from a number of distinct lines of evidence. From empirical research in cognitive psychology and cross-cultural studies in anthropology to Turing machines, the plausibility of wide computation does not rest with any specific area of research or class of phenomena. There are several converging lines of evidence that speak in favour of the view's plausibility.

Finally, and perhaps most importantly, what binds all the previous views together is an appreciation of one central insight: namely, that computational analysis is methodologically and ontologically indifferent to the location of computational systems. The possibility of wide systems follow from the location neutrality of computational analysis itself. It is this insight that largely sustains wide computational thinking. It is this insight that makes environmental states and processes apt for inclusion in wide systems.

## A Three-Dimensional Framework

I mentioned earlier that one of the benefits of pursuing the current project was that it offered a chance to map the logical space of computational views. I want to turn now to this task.

### Three Questions for Computation

There seem to be three distinct questions one can ask about any particular view of computation.

The first question is: Are only internal states and processes of the individual constitutively relevant to computational systems? Can external states and process also be included? I call this the *location question*. The location question distinguishes views according to the types and location of elements they include in their computational systems. This question takes its lead from internalist/externalist debates in philosophy of mind.

For example, there are those views which firmly ensconce computational systems within the individual. These views assume that the whole subject minus the environment "is the

largest integrated system available" for the science of mind (Segal, 1991, p. 492). Computational systems are located squarely in the brain or some submodule of the individual (Egan, 1992; Segal, 1997; Adams and Aizawa, 2008). On the other side of the equation, there are those views which look further afield for their computational units. According to these views, computational systems, in fact, spread out beyond the individual. The environment is constitutively part of at least some computational systems; the wide views previously discussed, such as Wilson (1994) or Losonksy (1995), are prime examples.

The second question is: Are representations necessary for computation? I call this the *content question*. The content question asks for a pronouncement on the significance of semantic content and representations in computation. It asks whether there can be 'computation without representation', to borrow Jerry Fodor's famous phrase.

For example, there are those accounts which view representations as part-and-parcel of computational analysis. For these views, computational theories of mind entail or require representational theories of mind. Many of the now classic views of computation, such as Pylyshyn (1984) or Fodor (1987), adopt this position. On the other hand, there are those accounts which view representational talk as only secondarily important to computational analysis. Egan (1992, 2010), for instance, maintains that the explanatory heavy lifting is primarily done by what she calls 'mathematical content', the information specified by function-theoretic descriptions. Representational talk, according to this view, only plays a secondary, heuristic role in making computational analysis more understandable to investigators.

The third question is: What conditions are necessary to individuate computations? Are extrinsic properties necessary? Semantic properties? Only intrinsic ones? I call this the *taxonomic question*. The taxonomic question asks after the conditions or requirements a view takes as key to individuating computational states and processes.

For example, there has recently been a flurry of debate about whether semantic or extrinsic properties are necessary to individuate computations. Sprevak (2010), for example, offers a

number of cases involving logic gates to suggest that semantic properties are crucial to interpreting otherwise equivalent computational descriptions. In a related vein, Shagrir (2001, 2006) devises several arguments using voltage gates to advocate for the crucial role of extrinsic properties, one he even calls the 'master' argument (Shagrir, forthcoming). On the other side of the debate, there are views which altogether eschew talk of semantic or extrinsic properties. Dewhurst (2018), for example, argues that computations can be individuated solely in virtue of their physical properties. Re-working the previous examples, Dewhurst suggests that computational differences can be accounted for simply in virtue of physical differences in the underlying systems.

**Three Dimension of Analysis**

What is interesting about the three questions is that they naturally give way to three dimensions of analysis. When translated into a yes/no format, each question forms an axis on which to analyse different views of computation. Consider three revised versions of the questions:

1) Do computational systems include features or elements of the environment? Yes/No (Location Question)
2) Do computational systems necessarily involve semantic content/representations? Yes/No (Content Question)
3) Are semantic or extrinsic properties necessary to individuate computations? Yes/No (Taxonomic Question)

Each dimension of analysis corresponds to an answer to one of the three questions. Taken together, the three dimensions create a space in which to situate various views of computation. I call this 'the three-dimensional framework for computation' (or TDF for short). Figure 1 provides an illustration.

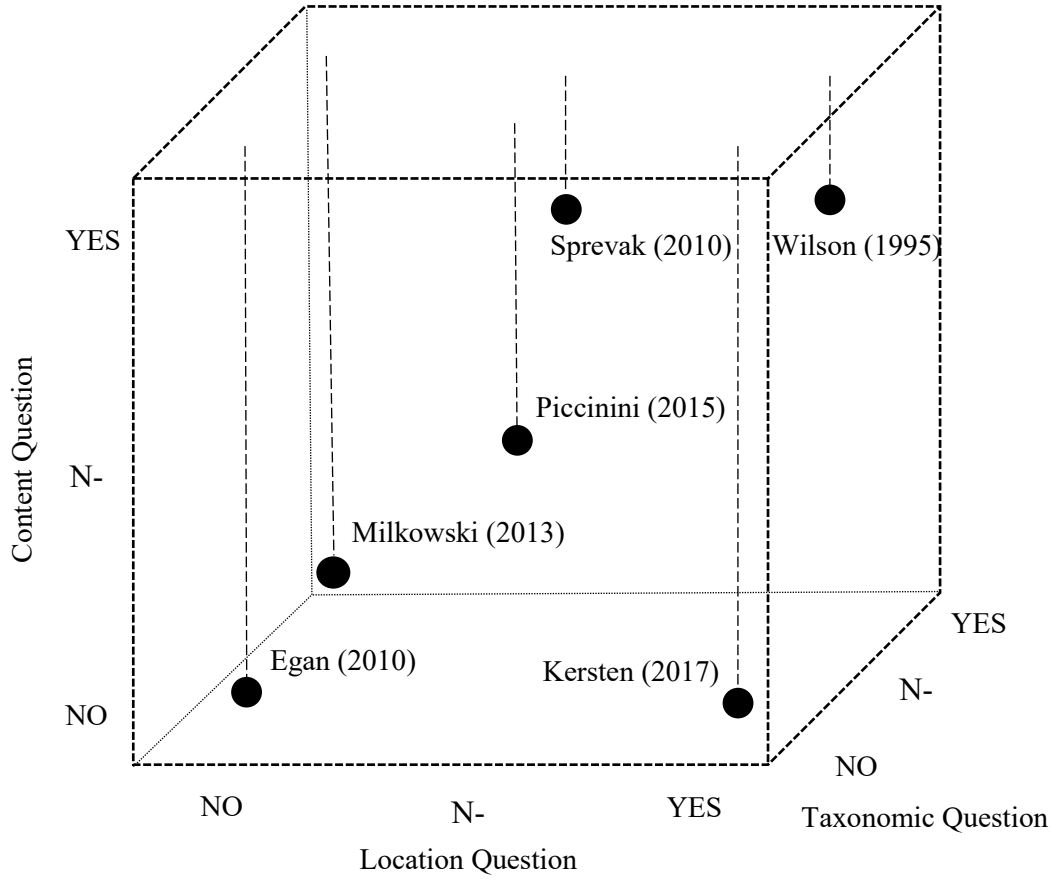


Figure 1. The Three-Dimensional Framework. 'N-' denotes neutral.

To get a better sense of the TDF, consider how three views are located within the space. Take Wilson's (1995) view, for instance. First, as we saw, Wilson maintains that computational systems are extended beyond the boundaries of the individual. Wide systems are those including elements of the surrounding environment. The account answers 'yes' to the location question.

Second, while retaining a commitment to the causal mapping account of computation, Wilson also makes room for semantic content or representations in his account. In a later work, for example, he writes: "Wide computationalism constitutes one way of thinking about the way cognition, even considered computationally, is 'embedded' or 'situated' in its nature, and it provides a framework within which an exploitative conception of representation can be

pursued" (p.165). Exploitative representations, for Wilson, are representations that take advantage of mind-world constancies; he offers the example of an odometer that directly represents distance via wheel rotations. Wide systems often trade-in exploitative representations. The account also answers 'yes' to the content question.

Finally, Wilson maintains that computational systems, in virtue of being locationally wide systems, are also taxonomically wide. He writes: "Since they [wide systems] literally extend beyond the boundary of the individual, not all of the states they contain can be taxonomised individualistically" (p.354). At least some wide computational states are individuated by extrinsic features of the environment. The account therefore also answers 'yes' to the taxonomic question. Taken together, then, Wilson's answers situate his account in the top right hand corner of the space.

Consider a second example: Egan (2010). First, as mentioned, Egan answers 'no' to the taxonomic question. According to her view, computations are individuated solely in virtue of their mathematical function, such as Marr's (1982) Laplacian of the Gaussian function for vision (a curve-smoothing function). The mathematical functions used by computing devices fully suffice to distinguish one computational system from another.

Second, representations only play a secondary, explanatory role in computational investigations. She writes, for example: "representational content, on this view, is a crucial component of what I am calling the 'explanatory gloss' on the theory. A representational gloss is necessary for the theory to yield explanations of the cognitive phenomena that form its explanatory domain." (p.258). While not type-individuating computational mechanisms, semantic content can provide a general motivation for computational accounts; a useful explanatory gloss. Thus, Egan's view also answers 'no' to the content question.

Finally, while slightly more subtle in its statement, Egan's view also offers an answer to the location question. She writes: "Whether a computationally characterized device succeeds in computing, say, the depth of objects and surfaces in the scene from information about the disparity of points in the retinal image depends on whether its internal states co-vary with

changes of depth in the environment. This requires a certain fit between the mechanism and the world" (p.257). While not exactly explicit, there is a suggestion here that computational mechanisms are firmly fixed within the individual. Talk of the 'fit' between mechanisms and the environment is suggestive of an internally bounded conception of computational systems. The account seems to answer 'no' to the location question. Taken together, such answers position Egan's account at the bottom left hand corner of the space. The TDF works equally well in classifying non-wide views of computation.

Finally, my own position finds its place in the bottom right hand corner of the space. The main reason for this, as we will see in more detail in Chapters 1 and 2, is that while the wide mechanistic view of computation answers 'yes' to the location question, it answers 'no' to the taxonomic and content questions. While it sides with wide views on the location of computational systems, in integrating mechanistic views of computation, it also adopts a functionalist approach to individuation. This means that semantic properties play, at best, only a heuristic role within the account. The view lands somewhere between Egan (2010), Piccinini (2015) and Wilson's (1994) views.

I should make a few caveats about the TDF at this point. First, while I have plotted several views in the space, I by no means take this list to be exhaustive. It simply reflects those views which have either already featured in the discussion so far or ones that will prove important later in the thesis. That said, of the sample provided, what the views do help show, I think, is the size of the space. The TDF captures a wide range of views on computation, everything from representationalist (Sprevak, 2010) to mechanistic views (Milkowski, 2013). Second, I do not assume that the positions I have plotted are in any way definitive statements on the views' theoretical commitments. Their placement reflects more a rough sense of how the views are usually interpreted. My goal is provide a general framework for thinking about computational accounts, and only secondarily an exegesis of specific author's commitments. That said, I think there is a good case to be made for positions of all the authors.

More positively, there are several advantages to adopting the TDF.

First, it helps to identify various differences *within* views of computation. For example, while many classifications draw a distinction between semantic and non-semantic accounts of computation, the TDF allows one to tease out differences within semantic views specifically. Shagrir (forthcoming), for instance, points out that there is difference between questions about computational *implementation* involving semantic properties and those about *individuation*. In the formers' case, one is interested in whether semantic properties are 'essential' to computation; while the latter's case, one is interested in whether semantic properties distinguish one computation from another.

The TDF is able to capture this difference. Because of its inclusion of a content dimension *and* a taxonomic dimension, the TDF is able to pull apart the differing roles for semantic properties. An account can adopt an individuative role for semantic properties (the taxonomic question) without also necessarily being classified as committed to an implementational role (the content question), and vice-versa.

As mentioned, this is important because not all classifications have been able to do this. In his opening description of three main views of computation, for example, Fresco (2014) writes: "The semantic view of computation, according to which is individuated by its semantic properties. On this view, computational individuation makes an essential reference to the representation and content" (p.15). Note the two different roles for semantic properties here. There is an individuative role, one concerning how to explain computations, and there is an implementational role, one concerning the status of semantic properties in computation; whether semantic properties are essential to ascriptions of computations. Such broad classifications fail to appreciate that it is possible for semantic views to be committed to the former type of claim without being committed to the latter.

Second, the TDF helps to draw out elements that may otherwise go overlooked within computational accounts. For example, Piccinini's (2015) account remains neutral on at least two of the three questions. With respect to the location question, for example, Piccinini writes: "I am officially neutral on whether the components of psychological computing mechanisms

extend beyond the spatial boundaries of organisms" (2015, ch.7). Mechanistic computing systems, at least in Piccinini's eye, are neutral on the question of location. This is suggestive because it points to a potential compatibility between wide and mechanistic views of computation. If computational mechanism are location neutral, then there is no, in principle, bar to computational mechanisms extending beyond the individual. I explore this potential compatibility in greater detail in Chapters 1 and 2.

With respect to the taxonomic question, Piccinini (2015, ch.7) writes: "Of course, many (though not all) computational vehicles do have semantic properties, and such semantic properties *can* be used to individuate computing systems and the functions they compute. The functions computed by physical systems that operate over representations can be individuated either semantically or non-semantically". Again, Piccinini is non-committal about the role of semantic properties. He flirts with the idea that semantic properties make some contribution to individuation, but to what extent remains unclear.

Because Piccinini remains non-committal on the location and taxonomic questions, his account falls in the middle of the space. More interesting, though, is that in posing different questions about computation the TDF draws our attention to otherwise underappreciated aspects of Piccinini's view, such as his neutrality on the location question. Rather than running roughshod over the view's subtleties, by focusing on key dimensions of analysis the TDF can capture some of the nuance within Piccinini's account.

Finally, the TDF respects a number of existing distinctions within classifications of computational accounts. We have already seen one such example with the individuation/implementation distinction. Another is the distinction often drawn between *semantic* and *causal/functional* accounts of computation (Fresco, 2014). As mentioned, semantic views are those implicating semantic or representational properties; while causal/functional views are those implicating either causal or functional/organisational properties. The TDF preserves this distinction via its content and taxonomic dimensions. Simply put, semantic views are those answering yes to one or more of the content and

taxonomic questions; while causal and functional views are those answering no to both. One admitted limitation of the TDF is that it doesn't neatly draw a distinction between causal and functional accounts. However, this is not so bad, as most classifications do not draw a distinction between these views either (see, e.g., Fresco 2014, ch.7).

The TDF also respects a more recent distinction between *externalism about computation* and *computational externalism*. As Shagrir (forthcoming) describes it: "Computational (or wide) externalism is a claim about the location of the vehicles of computation (e.g., Wilson, 1994), whereas externalism about computation is a claim about what individuates computational states, regardless of where they are located." The TDF captures this distinction via its taxonomic/location dimensions. While wide accounts may claim that some computational systems extend beyond the individual, this does not entail that they are also committed to viewing extrinsic features as playing an individuative role in those systems; the computational identity of a process might still be determined solely by the semantic properties intrinsic to an individual.

So, to summarise, the TDF can not only identify subtle differences within views of computation and pull out some otherwise overlooked features, but it can also preserve a number of distinctions already in use within existing discussions of computation. All grist for the TDF mill. However, there is an outstanding question that still needs to be addressed: namely, how can one be sure that the three dimensions of analysis are really distinct? Perhaps one or more of the dimensions of analysis collapse into each other. If so, then maybe there is nothing distinct about the TDF over and above what is already captured by existing classifications.

**Three (Distinct) Dimensions of Analysis**

One way to be sure would be to go through each view individually and show that their answers to one question don't constrain their answers to another. While thorough, this approach is overly demanding. A simpler solution, I think, is to provide a general scheme, and describe

the structure of the space. This should achieve the same result but without the detailed exposition; the reader is also then free to move the space at her own leisure. Consider Figure 2.

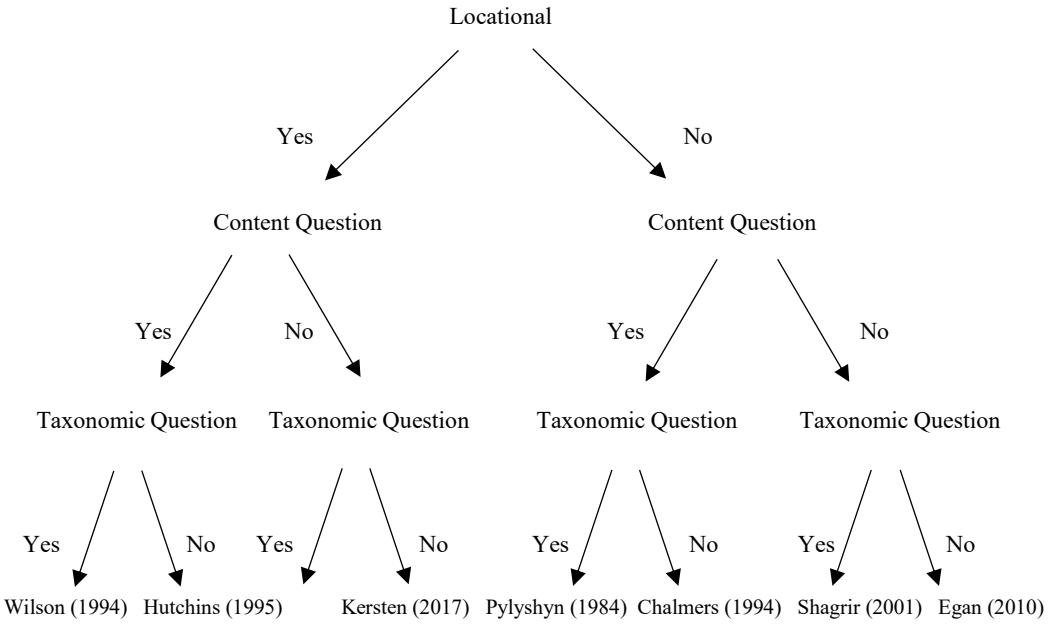


Figure 2. A flowchart of several possible answers to the three questions.

Figure 2 provides a map of a series of pathways through the logical space.[9] It highlights a number of different positions one might adopt in response to each of the three questions.

If we descend down the right hand side, we find more classic or standard views of computation, such as Chalmers (1994) or Egan (2010). Once it has been decided, implicitly or explicitly, that computational systems are centred on the individual, whether for methodological or empirical reasons, many of the classic and contemporary debates about computation, such as those about representation or external content, begin to take shape.

If we descend down the left hand side, having answered yes to the location question, we find different versions of wide computation. While the content and taxonomic questions are still answered, they are slightly overshadowed by the rather divisive location question. For

[9] In principle, the tree could be expanded to include terminal branches, representing those views which do not take a stance on one or more questions, and connecting lines between branches, representing answers to different questions. However, for simplicity, I've plotted only those views which take a clear stance on the three questions, i.e. yes or no.

instance, Wilson and Hutchins' views part ways on the taxonomic question, for while Hutchins thinks that group level systems trade in representations, he does not think that representations are what individuate computations (see Hutchins 1995, p.51); whereas, as we saw, Wilson thinks that one follows from the other.

More importantly, though, the flowchart illustrates that one is free to move through the space in any number of ways. An answer to the content question, for example, no more determines one's answer to the taxonomic question as it does the location question. While both Shagrir (2001) and Egan (2010) answer no to the content question, they provide differing answers to the taxonomic question; and while Sprevak (2010) and Chalmers (1994) both answer yes to the content question, their opinion is split on the taxonomic question. Of course, there might be all sorts of specific reasons why one chooses to answer the questions one way rather than another, such as worries about providing an adequate computational taxonomy (Chalmers, 1994) or accounting for voltage gates cases (Shagrir, 2001), but there is no, in principle, reason why an answer to one question has to commit one to a specific answer on any of the other questions.[10]

The reason, I think, that things might appear otherwise is that a number of positions have, for historical reasons, tended to group together. For example, as we saw, it was quite natural until recently to group semantic and externalist views of computation together (see, e.g., Shagrir, 2001). Semantic content, after all, is often widely individuated. These groupings offer the impression that certain positions naturally entail one another. Yet, as authors have increasingly appreciated, there are number of other extrinsic features that might also individuate computations but which are not semantic, such as evolutionary features (Dewhurst 2018; Shagrir forthcoming). The separation of the content and taxonomic questions reflects a

[10] One might worry that I've somehow forced this conclusion by structuring the space starting with the locational question; that begun with a different question the space would substantively change. This isn't the case. One could just as easily construct the space starting from either the taxonomic or content questions and the only difference would be a more disorganised flowchart. The same views could still be placed, it is just that the differences between the views would not be as clear.

growing appreciation of the number of conceptually distinct positions one might adopt about computation.

It is also no surprise that not all of the space is filled. Until rather recently, most classifications would have ended with the content question, or swapped the content question for the taxonomic question. Wide and individualistic views of computation, for example, equally painted over the difference between individuation and implementational questions. Both views assumed that an answer to the content question also brought with it an answer to the taxonomic question. One way to see the current thesis, then, is as an attempt to fill in one lesser appreciated position within the space.

So, while some answers to the three questions have been a bit friendlier to one another than others, and so have tended to group together more naturally, e.g., semantic and externalist views of computation, in principle, this does not mean that any of the three questions constrain each other in any substantive way. The three dimensions of analysis reflect three logically distinct ways to carve up the space of computational views.

## The Scope

I want to close out the introduction by laying out the scope of the current thesis.

There are three assumptions that I want to flag at the outset. The first is that computation, in one form or other, constitutes a viable approach to studying the mind and cognition. While I make some overtures to anti-computationalists in Chapter 3, in fact a good deal of Chapter 3 is dedicated to buildings bridges with such views, I nonetheless assume in what follows that computational approaches have something important to say about how we think about and study the mind. That said, if the thesis' arguments are successful, then the work as a whole can be read as an extended defence of a particular computational theory of mind, albeit a non-standard one.

The second assumption is a general reliance on mechanistic views of computation. While I extoll the virtues of mechanistic computation in several places, I do not offer a sustained

defence of the general approach in the thesis. I take it for granted that mechanistic views offer a plausible approach to computation. The main reason for this is that, for the most part, I think this has been done successfully elsewhere. Piccinini (2015), Milkowski (2013), and Fresco (2014), for example, have all offered comprehensive analyses and defences of the approach. Moreover, I think that pursuing such a defence would also involve too much of a detour from the thesis' main aim. Entering into a protracted defence of mechanistic thinking would only take away from articulating the case for wide mechanistic computation.

The third assumption is a weak form of pluralism. I assume in what follows that there a number of different approaches to computation in cognitive science, and that plausibility of one does not detract from the plausibility of others. For example, regardless of the truth of wide computation, I take it that individualistic or internalist views of computation, such as those of Egan (2010), still constitute viable approaches to studying the mind. There is nothing about wide computation, mechanistic or otherwise, that precludes others from pursuing alternative, narrow approaches. My claim is only that we should not consider ourselves *restricted* to follow a narrow approach.

In terms of its general approach, the thesis attempts to wear two hats.[11] It is an exercise in what Andrew Brook (2009) calls 'philosophy *in* and philosophy *of* cognitive science'. Brook describes the difference as follows:

> The former embraces work done on topics such as mind and language that are also studied using other approaches such as behaviour experiments and theoretical linguistics, so philosophy of mind and language. The latter is a branch of philosophy of science and is a meta-study. It studies what others do—rather than doing cognitive science, it studies cognitive science (p.218).

On the one hand, the thesis attempts to have its hands in the day-to-day practice of cognitive science. It attempts to articulate an account of wide mechanistic computation that appeals to and is responsive to work in cognitive science. For example, in Chapter 1, I appeal to empirical research in animal and human psychology to motivate the view's plausibility; while in Chapter

[11] A hat on a hat if you will.

2, I attempt to show that the view has empirically testable consequences; that it is not simply an abstract philosophical thesis about computation.

On the other hand, the thesis also attempts to take a step back from the details of computational cognitive science and clarify several key concepts. For example, in Chapter 4, I examine the notion of 'medium independence'. I do so in order to resolve a specific problem that arises for mechanistic views of computation, what is called 'the abstraction problem'. The thesis also aims to provide some normative guidance on how concepts should be used within computational cognitive science.

Finally, I should say what the thesis is not.

First, while I have said that wide computation bears on questions relevant to cognition, the thesis itself does not take a stance on what cognitive states or processes are, beyond assuming that they are at least partially computational. I try to remain neutral on the nature of mind and cognition. The one foray the thesis does make into such waters in Chapter 3 is done only to resolve particular tensions within the so-called 4Es of cognition.

Second, the thesis is not a defence of extended cognition. While it is true that wide computation can form one argument for extended cognition, and that wide computation has traditionally been an important forerunner to extended views of cognition (see, e.g., Wilson 2004; Wilson & Clark, 2009), my focus here is on developing a wide mechanistic account of computing for cognitive science as a whole. Wide mechanistic computation stands or falls independent of the arguments for extended cognition; any ancillary benefits it has for such views are simply a bonus. That said, chapter 5 does take up recent proposals for how to connect predictive processing and extended views of cognition. However, I do not take a stance on the truth of extended cognition in this chapter but, rather, simply raise concerns about the value of these recent proposals.

Third, the thesis should not be read as an extended inference to best explanation. The argumentative strategy is not one of trying to read off wide mechanistic computation from the best available evidence. It is not an attempt to reason from a set of empirical data to the best

hypothesis. Rather, I take the thesis to weave together several types of evidence, including empirical, theoretical, and methodological considerations. Part of the point of Chapter 3, for example, is to show that wide mechanistic computation is theoretically useful, in that it can help resolve specific problems within 4E cognition. The thesis is better seen as an attempt to integrate empirical and conceptual work into an independently well-motivated approach to computation.

So, we have gotten a sense of what wide computation amounts to, why it might prove interesting, and how it relates to other views of computation. The view is straightforward in its statement and yet far-reaching in its consequence. Wilson (1994) sums up the situation facing wide computation nicely when he says:

> The central idea behind wide computationalism is extremely simple. However, fleshing out the idea and being explicit about the implications it has for issues in philosophical psychology allow one to see the respects in which it represents a radical departure from the conception of the mind underlying much contemporary research in computational psychology (p.370).

For an account of wide computation to be successful, it needs to not only flesh out the idea of wide computation but also spell out what implications the view has for cognitive science. It is to this task I want to now turn.

## Chapter Summaries

The thesis is organised into six substantive chapters, each with its own introductory section. These introductory sections serve as bridges between the chapters, identifying key ideas, areas of argument, and the broader context. Their necessity stems from the publication-based format of the thesis.

I should also insert a small disclaimer. Again, largely owing to the publication-based format in which the thesis was written, there is some overlap between various parts of the thesis, particularly with respect to the explanations of wide computation and mechanistic computation. In some sense, this was unavoidable due to the self-contained nature of

manuscript writing. One silver lining, however, is that the chapters and their arguments, can, to a certain degree, be read independently of one another without loss of coherence.

Chapter 1 lays the groundwork for the mechanistic account of wide computation. It begins by unpacking some of the core assumptions of previous formulations of wide computation, such as Wilson (1995) and Hutchins (1995), and then raises several issues for these accounts via Piccinini's (2015) six desiderata for computation. The suggestion is that while traditional accounts handle the first three desiderata well, they struggle to accommodate the remaining three. To overcome these problems, I appeal to Piccinini's (2007, 2015) account of mechanistic computation, laying out his three main requirements on computational implementation. After demonstrating the conceptual compatibility between wide computation and Piccinini's account, I outline several examples that satisfy the implementation requirements, but which also crucially include elements of the environment. In particular, I examine research on bat echolocation (MacIver, 2009) and spatial navigation among sightless individuals (Ricciardi et al., 2009). The result is an argument for not only the coherence of wide mechanistic computation but also the view's empirical plausibility. I further develop the view by defending it from several objections.

Chapter 2 continues to develop wide mechanistic computation by further integrating it with mechanistic computation. It does so by addressing two outstanding challenges, what I call the *parsimony* and *testability* challenges. To motivate the threat of the challenges, I measure each worry against Wilson (1995) and Losonksy's (1995) formulations of wide computation. Neither account is able to satisfactorily respond to the challenges. I then propose appealing to Craver's (2007) work on 'constitutive relevance' to overcome the challenges. It turns out that the notion of constitutive relevance supplies wide computation with the resources it needs. This is because it (i) undermines the parsimony challenge's underlying 'ceteris paribus' assumption and (ii) it draws on a connection between constitutive relevance and experimental investigations, and so makes wide investigations empirically testable. The uptake is that I show

not only how the wide mechanistic computationalism can defend its ontological commitments but also how it can be made empirically testable. I conclude by responding to objections.

Chapter 3 uses wide mechanistic computation to resolve two tensions troubling 4E cognition. The first, raised by Clark (2008), says that the body-centric claims of embodied cognition run counter to the distributed tendencies of extended cognition. The second, raised by Clark & Kiverstein (2009), says that the body/environment distinction emphasised by enactivism militates against the world-spanning claims of extended cognition. I attempt to resolve these tensions by showing that in virtue of placing a simultaneous emphasis on location neutrality, concrete physical implementation, and autonomous systems, wide mechanistic computation is able to accommodate many of the central insights and concepts relevant to embodied, enactive and extended theorists. I argue that this result is important, as it shows that computational views have an important role to play in discussions of 4E cognition. I conclude by responding to possible objections.

Chapter 4 takes a step back and examines the foundations of mechanistic computation. It takes up a recent challenge to mechanistic computation called *the abstraction problem*. After outlining two of the mechanistic computation's key assumptions, constitutive explanation and medium independence, I articulate how these concepts help the view to satisfy a number desiderata on a theory of implementation, what I call the 'adequacy conditions'. Following this, I outline the abstraction problem, along with a recent response from Kuokkanen and Rusanen (2018). I argue that Kuokkanen and Rusanen's proposal, while interesting, nevertheless comes up short. This is because it subtly reframes mechanistic computation as an anti-realist view of implementation, and in so doing makes problematic trade-offs among the adequacy conditions. This, I suggest, reveals a more general a dilemma facing mechanistic computation. To avoid the dilemma, I argue that computations and their vehicles should be thought of as *abstracta*. This re-framing not only provides a response to the abstraction but also offers a way of retaining the three adequacy conditions. The benefit of this chapter is that

shores-up the conceptual foundations of mechanistic computation, and also provides a meta-theoretical framing for the thesis.

Chapter 5 sounds a cautionary note about recent proposals connecting predictive processing, the view that the brain is a hierarchical prediction machine, and extended cognition, the view that cognitive systems and processes extend beyond the boundary of the individual and into the environment. It addresses two proposals, in particular: Clark (2016a, 2017a) and Ramstead et al. (2019a). After outlining the two proposals, and identifying their key features, I argue that both proposals come up short as accounts of extended cognition, as both remain committed to 'first' and 'second' wave approaches to cognitive extension. This, I suggest, leads to a dilemma. Either the two proposals (i) fail to advance discussion of extended cognition or (ii) they function as demonstrations of a broader compatibility between PP and extended cognition, and so fail to offer any novel insights about extended cognition. I conclude sketching a general argument for how PP could be used to move the needle forward in discussions of extended cognition.

Finally, Chapter 6 examines a particular conception of levels within in cognitive science, what I call the 'hierarchical correspondence view of levels' (or HCL). It outlines the main elements of the HCL, and then provides two examples of its view in action: Newell (1990) and Sun et al. (2005). Following this, I provide a novel argument against the HCL. Using a several motivating cases, I argue that the HCL offers an overly restrictive view to cognitive science. It does so because it fails to appreciate an important distinction between *shifts in grain* and *shifts in analysis*. I diagnosis this failure as a result of the HCL's inability to appreciate an important part of our explanatory practices about complex systems: namely, the role of 'functional contextualisations'. I close by anticipating potential objections and articulating a version of the HCL that respects the shift versus grain distinction. The wider lesson of this chapter is that while cognitive science can avail itself of the concept of levels it should not to do so in virtue of positing a hierarchy of levels arranged in one-to-one correspondence.

# Introduction to Chapter 1

## Key ideas

I defend two key ideas in Chapter 1.

The first is that wide computationalism follows, in part, from the concept of the 'medium independence' of physical computation. This is the idea that a computational state or variable can be implemented in a variety of physical media in virtue of possessing certain 'degrees of freedom', i.e. the ability to take on at least one dimension of variation. For example, the cells in a Turing machines can take on at least two values, '1' or '0', so they have one dimension of variation. Physical implementations of Turing machines process the cells in the tape in virtue of responding to different portions of the vehicle's structure, i.e., the shape or form of the '1' versus the '0'. So as long as the cells in the tape can take on two possible values, and the active components of the Turing machine flips those states, the vehicle is medium-independent. It can be implemented in any physical system, whether it is made of Lego blocks or silicon chips. Medium-independent vehicles are what are processed by physical computational mechanisms.

What I suggest is that certain external environmental structures can also be defined in a medium-independent way. For example, I suggest that there are certain invariant relationships in the acoustic array that are plausibly interpreted as possessing certain degrees of freedom. If certain environmental states can form medium-independent vehicles, and medium-independent vehicles are what are processed by computational mechanisms, then environmental states can also form parts of wide computational mechanisms. The medium-independence of certain environmental structures secures the possibility of wide computationalism. As we will see in Chapter 4, the notion of medium-independence is not

without its problems, but it's worth noting at the outset that insofar as the concept of medium independence remains plausible so too does wide computationalism.

The second idea is that the challenge of explaining what it means for a physical system to implement a computation, what is known as the problem of *computational implementation*, applies not only to traditional views of computation, such as semantic or causal mapping accounts, but it also affects accounts such as wide computation which rely on these implementational accounts. This is not something previous accounts of wide computation have traditionally appreciated. Wilson (1994), for example, writes: "given this notion of computation [the causal mapping account], the idea that there can be computational system that involve non-brain part of the world is trivial. Less trivial is the claim that the brain plus parts of the non-brain part of the world together can constitute a computational system" (p.168). For authors such as Wilson, the assumption has been that having an answer to the implementation question guarantees the coherence of wide computation. If the underlying implementational account of computation is correct, then wide computationalism follows.

However, as a recent collected issue of *Journal of Cognitive Science* suggests, the answer to computational implementation is anything but straightforward (Chalmers, 2011). Insofar as the problem of implementation forms a challenge to views of computation, it presents an obstacle to wide computationalism. I should be clear. The claim is not that wide computationalism is a rival to semantic or causal mapping accounts. Rather, the claim is that in relying on causal mapping accounts, wide computationalism stands to inherit its benefactor's problems. The view's failure to meet this challenge provides one reason to think we should try and update its foundations.

## The argument

In terms of its argumentation, the chapter roughly divides into two parts. The first part suggests that pervious formulations of wide computation, such as Wilson (1994) and Hutchins (1995), have fallen short as computational accounts due to their inability to satisfy several

desiderata on implementation, including: objectivity, explanation, the right things compute, the wrong things don't compute, miscomputation, and taxonomy. This failure, I argue, should lead to a re-alignment of wide computation with mechanistic approaches, such as Piccinini (2015). Since mechanistic views are better able to accommodate the key desiderata on implementation, they provide a firmer foundation for wide computationalism.

However, one might worry about this general approach to evaluating computational views. Why think, for example, that this specific set of desiderata is good method for judging the strength of an account of implementation? The main reason is that it incorporates a broad range of practices from the computational sciences, including computer science, engineering or cognitive science.

For example, if an account fails to establish whether some physical system, such as the brain, computes as a matter of fact, then it fails to serve as a foundation for cognitive science. If the account fails to explain how the behaviour of a computing system is the result of the procedures it executes, then it fails to make sense of basic facts in computer science and engineering. And if the account fails to deliver a computational taxonomy that includes paradigmatic cases, such as Turing machines, and excludes non-paradigmatic cases, such as solar systems, then it fails to do right by our intuitive taxonomic practices.

By using a broad range of desiderata different computational accounts can be grounded in real-world examples, while also doing justice to our more general conceptual and scientific demands. While it is true that there is some debate about the specific desiderata that are necessary for an account of implementation, the proposed six desiderata are generally thought of as plausible candidates. They are alert to a number of the practices relevant to the computational sciences (see, e.g., Fresco, 2014; Piccinini, 2015).

The second part of the chapter suggests that wide mechanistic computation also gains support from animal and human psychology – examples of bat echolocation and spatial navigation, in particular, are used to motivate the account. However, one might worry here

that two examples are not really sufficient to establish the empirical plausibility of the wide mechanistic computation.

While there is some merit to this concern, I think the two examples are nonetheless important because they point to the kind of evidence that could be used to further support the view. What I mean by this is that while the two cases do not alone establish the case for wide computation, they are suggestive of a more general class of phenomena that are plausibly interpreted as instances of wide mechanistic computation: *active sensory systems*. Bat echolocation and spatial navigation in sightless individuals both involve systems that actively probe and recruit information, whether haptic or perceptual, in order to carry out their functions. They are systems that use self-generated energy to control the intensity, direction and timing of various types of inputs. These, I think, are good markers for the presence of wide systems. The rich informational flow and tight causal integration of agent and environment in active sensory systems offers a good starting point for thinking about wide systems. The fact that such systems are also plentiful in nature is also promising.

## Broader context

I want to briefly locate wide mechanistic computation relative to some other important concepts in philosophy of mind and cognitive science.

First, with respect to the notion of representation, while wide mechanistic computation is not, strictly speaking, hostile to representations, they do not play a core role in the proposed account. This is because, in relying on the mechanistic approach, wide computationalism holds that computational specification only requires structural and functional analysis; it does not require an appeal to representational properties to flesh out how or why physical systems compute. This means that, at best, representations only have a secondary role within wide investigations. So while wide computation is compatible with the notion of representation, it is not necessarily friendly to it.

Second, with respect to functionalism, wide mechanistic computationalism can be seen, to a degree, as a limited form of functionalism, insofar as functional specification of various computational roles forms a core part of computational analysis more generally. That said, the two views are also distinct in an important sense. This is because, when raised to the level of theses about the mind, functionalism is a metaphysical thesis about the nature of the mind and computationalism is as an empirical hypothesis about the organisation of the brain (Piccinini, 2010). Wide computationalism is better interpreted as one possible way of cashing out the functionalist picture, but it is not necessarily the only way.

# Chapter 1 – A Mechanistic Account of Wide Computationalism

## 1.1 Introduction

Jerry Fodor once claimed that: "quite independent of one's assumptions about the details of psychological theories of cognition, their general structure presupposes underlying computational processes" (1975, p.28). Fast forward 30 years and views have changed little. Paul Thagard, for example, writes: "[t]he central hypothesis of cognitive science is that thinking can best be understood in terms of representational structures in the mind and computational procedures that operate on those structures" (2010, p.6).

Suffice it to say, the assumption that psychological states and processes are computational in character pervades much of cognitive science, what many call the 'computational theory of mind'. Yet in addition to occupying a central place in cognitive science, the computational theory of mind has also had a second life supporting 'individualism', the view that psychological states should be taxonomized so as to supervene only on the intrinsic, physical properties of individuals (Fodor, 1980, 1987; Stich, 1983; Egan, 1992). One route to individualistic psychology is what Robert Wilson calls the 'computational argument for individualism'. Wilson (1994, p.353) formulates the argument as follows:

1) Cognitive psychology taxonomically individuates mental states and processes only qua computational states and processes.
2) The computational states and processes that an individual instantiates supervene on the intrinsic, physical states of that individual.

3) Therefore, Cognitive psychology individuates only states and processes that supervene on the intrinsic, physical states of the individual who instantiates those states and processes.

One response to the computational argument has been to challenge premise (2). This is the route adopted by Wilson (1994, 1995). Wilson argues that since not all computational processes are instantiated in the head, not all psychology is individualistic. Wilson raises the prospect of 'wide computational systems', in which some computational units are instantiated outside the individual. By enlarging the concept of computation, Wilson attempts to sever the link between individualism and computational psychology.

The idea of 'wide computationalism' is more than a little interesting. Not only does it represent a substantial departure from orthodox thinking in cognitive science (Marr, 1982; Pylyshyn, 1984), but it also offers distinct grounds for thinking about extended cognition (Clark and Chalmers, 1998; Rowlands, 1999; Wilson, 2004; Wilson and Clark, 2009). However, for one reason or another, wide computationalism has received little attention in philosophy of mind and cognitive science. Though there has been some scattered discussion, no sustained analysis has been offered. This paper aims to revisit the prospect of wide computationalism.

The problem is that several issues plague initial formulations of wide computationalism. For this reason, focus here is given to buttressing the view via recent discussions of 'mechanistic' computation (Piccinini 2007, 2015; Milkowski, 2015). The argument is that by appropriating a mechanistic conception of computation the problems that emerge for earlier formulations of wide computationalism can be avoided (Sections 1.3 and 1.4). The goal is to show that cognitive science has overlooked an important and viable option in computational psychology. On route to this conclusion, the paper marshals empirical support for 'wide mechanistic computation' and responds to possible objections (Sections 1.5 and 1.6).

A quick clarification is in order before discussion gets going. Wide computationalism, as Wilson presents the view and how it is developed here, is not meant as a global thesis. It does not imply that all computational systems are instantiated, at least in part, outside the body. Rather, the view is better understood as a supplement to individualistic psychology. It is an extension of the logic of computational analysis in cognitive science, rather than a replacement.

## 1.2 Wide Computationalism

In the abstract, wide computationalism gains a foothold via the location neutrality of computational individuation. Because the method of computational analysis is noncommittal about the kinds of physical states that might be computationally characterized, it is at least possible that some of the relevant computational states might reside outside the individual. Wilson, for instance, writes: "There is nothing in the method of computational individuation itself…which implies that the class of physical features mapped by a realization function cannot include members that are part of the environment of the individual" (1994, p.355). Since formal systems are indifferent to physical medium and computation is a formal system, it is at least possible that some computational states and processes reside outside the individual.

The location neutrality of computational analysis also carries with it implications for psychological theorizing. This is because if psychological states and processes are computational in character and computational analysis is location neutral, then some psychological states or processes may extend beyond the boundary of the individual. As Hutchins points out in his discussion of Marr (1982): "Marr intended his [computational] framework to be applied to cognitive processes that take place inside an individual, but there is no reason, in principle, to confine it to such a narrow conception of cognition" (1995, p.50). The allegiance of psychology to computational theory carries with it the potential for extended cognition (see, e.g., Hutchins, 1995; Wilson, 2004; Kersten, 2015; Kersten and Wilson, 2016).

Whether or not computational cognitive systems are instantiated exclusively within the boundary of the individual is an a posteriori question.

Wilson furthers develops wide computationalism by outlining a method for identifying wide systems. He writes: "The account of actual implementation is a generalization of that in the case of narrow computational systems: a wide computational system implements the 'program' physically stored in the environment with which it causally interacts" (1994, p.360). Similar to identifying 'narrow' (in-the-head) computational systems, wide computational systems are present if there is a computational description that tracks causal transitions running from an organism's environment to its internal physical states.

Interestingly, in identifying wide systems in this way Wilson aligns wide computationalism with 'causal mapping accounts' of computation (Chrisley, 1995; Chalmers, 1994, 1996; Scheutz, 1999, 2001). Accounts of this stripe claim that for a physical system to perform a computation there must be a mapping from a subset of states ascribed to a physical system by a physical description to states defined by a computational description. For any computational state transition of the form S1 →S2, if a system is in a physical state that maps onto S1, then the physical state that maps to S1 must cause the system to go into a further physical state that maps onto S2 (see Piccinini 2015, ch.2). Causal mapping accounts articulate the conditions for ascription of computational implementation in terms of isomorphic mappings between computational descriptions and physical descriptions via transitions between physical states.[12]

Commitment to the causal mapping account is more fully revealed in Wilson's comment that: "[f]or a physical device to be capable of implementing a given program [computation] is for it to have its physical states configured in such a way that transitions between those states are isomorphic to transitions between states that the program specifies" (1994, p.360). The identification of actual computations, whether narrow or wide, requires mapping

[12] 1 Causal mapping accounts get cashed out in various ways. Some opt for a counterfactual approach, others a dispositional approach. The important point is that in all cases what actually ground computations is causal relations.

computational states to physical states in virtue of tracking causal relations between physical states.

Proponents of wide computationalism not only think that the notion is coherent, but that some organisms actually implement wide computational systems. Wilson, for instance, offers two examples of what he takes to be wide computational systems.

The first is Sekuler and Blake's (1990) multiple spatial channels theory of vision and form perception. The crucial feature of this account is that specific sets of neurons are sensitive to specific sets of stimuli. These stimuli are decomposable into sinusoidal gratings with four parameters: spatial frequency, contrast, orientation, and spatial phase. Any natural scene can be decomposed into these formal primitives. For Sekuler and Blake, perception is the result of organisms processing environmental inputs through spatial channels and turning them into complex internal representations.

What is crucial about this example for Wilson is that the account acknowledges the computational role states beyond the individual play within perceptual processes. The computational analysis involves, first, identifying and describing the formal primitives instantiated in the physical environment and, second, describing how such inputs function to produce internal representations. Instead of viewing computation as beginning at the retina and ending at the visual cortex, Wilson maintains that computational analysis begins further downstream in the environment of the perceiving organism. Framing things in terms of the causal mapping account, the claim is that there is a computational description amenable to the physical states internal and external to the organism that help to explain form perception.

It is also worth noting that although there is sometimes a tendency in discussions of extended cognition to devalue or even dismiss the need for internal representations, there is nothing strictly antithetical about the two notions (see, e.g., Wilson, 2004; Clark, 2008). Wide computational systems of the kind relevant to extended cognition can trade in internal information-bearing vehicles just as easily as they can external ones. What is important is the role internal or external information-bearing vehicles play in the larger computational analysis.

The appeal to internal representations in Wilson's examples is simply the logical extension of pushing computational analysis further out into the environment.[13]

The second example Wilson offers is Gallistel's (1989a, 1989b) account of animal spatial navigation. Wilson points out that according to Gallistel's theory, animals construct complex representations of their environments in order to guide behaviour by instantiating modules sensitive to formal geometric structures of the environment. One example of this process is dead reckoning in ants and bees. These organisms take as their inputs three features: the animal's solar heading, forward speed and a representation of the solar azimuth. What they produce is a representation of position relative to some landmark. There is a physical process characterizable as a computational process that begins in the environment and ends in the organisms, specifically as a representation of relative position. Framed again in terms of the causal mapping account, there is a causal transition between physical states isomorphically related to a computational description of spatial orientation. There is a subset of physical states that track transitions between computational state descriptions.

A third example of wide computation comes from Edward Hutchins' (1995) work on navigation. Hutchins contends that when members of a navigation team carry out coordinating actions in service of a larger task – for example, navigating a ship at sea – they form a larger computational system that transcends the individual team members. Navigation – that is, the task of figuring out where something is relative to other positions – is achieved not only through the local actions of individual team members, such as the Navigator or Fathometer Operator, but also through the coordinated activity of the team members as a whole. Hutchins, for instance, writes: "In their communication and in their joint actions, the members of the navigation team superimpose themselves on the network of material computational tools of the trade" (1995, p.219). The navigational team can be thought of as a wide computational system because the functional whole extends beyond the local actions of individual team

[13] See Chemero (2011) or Varela et al. (1991) for an alternative perspective.

members. The social organization becomes the computational architecture on which the larger functional task is carried out.

Putting aside minor differences among the examples for the moment, the central message is that wide computational systems are not only theoretically possible, but that they are physically implemented in a number of cases.[14]For authors such as Wilson and Hutchins, the fact that research in human and animal psychology provide putative examples of concrete computational systems beyond the boundary of the individual is further vindication of the idea that at least some portion of computational psychology is not individualistic.

## 1.3 Concrete Computation

One of the main problems for computational theories of mind is the problem of 'computational implementation' (Chalmers, 1994, 1996, 2011; Sprevak, 2012). The issue is one of how to specify the conditions under which computations can be said to take place in physical systems. If psychological states and processes are computational in structure, then any successful account of computation has to explain how those computational states and processes are instantiated in physical systems. Without a successful account of computational implementation, the chances of a robust computational cognitive science diminish.

Several proposals have been offered, perhaps most famously Putnam's (1967) mapping account. Of interest here is the solution offered by wide computationalism. Recall that wide computationalism subscribed to the causal mapping account of computation. For the causal mapping account, concrete computations occur just in those cases where there is a mapping of

[14] One point of difference between Wilson and Hutchins' views lies in the breadth of the computational system identified. Wilson's examples identify wide computational systems applying to largely brain/environment composites; while, in contrast, Hutchins broadens the reach to include individual agents. This results in a difference in emphasis when it comes to the location of representational states within each system. For Wilson, because the wide computational system runs from the formal primitives of the environment to the processing centers of the brain, the representational states are located within the brain; whereas for Hutchins, the representational states relevant to defining the navigational computation are spread out across the coordinated activity of the individual team members (see Hutchins 1995, ch.4). Though at this first glance this difference might seem notable, particularly in light of some hostility occasionally levelled by proponents of extended cognition toward internal representations, nothing crucially important hangs on it. Rather, the difference emerges largely as a result of the differing computational units emphasized within each analysis. The scope and elements of the computational system help determine the kind and location of representational states implicated.

computational to physical descriptions tracking causal transitions between physical states. By specifying the isomorphic relations between physical states and computational states via causal transitions, the causal mapping account provides a resolution to the problem of implementation, and thus so too does wide computationalism.

However, a solution to the problem of computational implementation is only one desideratum on an account of computation. This is because not only should an account of computation explain how physical systems compute it should also do justice to the sciences of computation. It should strive to do right by the practices of computer scientists, engineers, and cognitive scientists. Piccinini (2015, ch.1) highlights six such desiderata:

1) Objectivity. The account should make whether a system performs a particular computation a matter of fact. It should establish some form of objectivity on questions of computational implementation.
2) Explanation. The account should explain the behaviour of computing systems in terms of the procedures being executed. It should say how appeals to program execution, and more generally to computation, explain the behavior of computing systems.
3) The right things compute. The account should include the paradigmatic examples of computing mechanisms, e.g., finite state automata, Universal Turing machines, etc.
4) The wrong things don't compute. The account should not entail paradigmatic examples of non-computing mechanisms, e.g., galaxies, digestive systems, etc.
5) Miscomputation. The account should explain how computations can go wrong.
6) Taxonomy. The account should provide a taxonomy that is able to distinguish between different kinds of computing machines, e.g., finite state automata, Universal Turing machines, calculators, etc.

The question is whether wide computationalism, as a computational theory of mind, satisfies these six desiderata. If it does not, then it may not be a viable account of computational cognition.

First, does wide computationalism provide some form of objectivity about computation? By endorsing the causal mapping account, it would appear so. Wide computationalism restricts the class of physical systems that can be said to implement computations by admitting only those systems that map computational descriptions to physical state transitions.[15] Deciding whether or not a system computes is in some sense a matter of fact according to wide computationalism.

Second, does wide computationalism account for how program execution explains computing behaviour? Here the account stumbles. To qualify as a computational explanation, wide computationalism must explain how a program generates a system's behaviour via a rule or program.[16] The problem is that although wide computationalism describes where a program is represented by causal transitions (for example, the ephemeris function in the case of animal spatial navigation), it does not show how physical systems deploy or execute computational programs in the production of behaviour. Although the causal mapping account provides a computational model or description, it does not offer a computational explanation.

With respect to the third desideratum, wide computationalism performs satisfactorily. This is because it entails that there is some subset of physical states in most computing devices – for example, digital computers and calculators – that can be mapped to computational descriptions. Under a wide computational rubric, it is possible for each paradigmatic computing device to map from some of its physical states to the relevant computational states.

The fourth desideratum is not so easily met. On Wilson's formulation, physical states of paradigmatic non-computational processes, such as the weather or respiration, can also be mapped to computational descriptions. The problem is that non-paradigmatic cases also trade in the right kind of causal transitions such that they can be mapped to computational state

[15] This is in contrast to simple mapping accounts, such as Putnam (1967), that place no restrictions on the physical states that might form the equivalence class for computational description. This is why Putnam is skeptical of computational theories of mind.

[16] This does not necessarily imply that computing system need to represent a rule internally, for this would overly restrict computational explanation, excluding paradigmatic cases of computing devices such as the finite state automata.

descriptions. Although it manages to account for cognitive systems and computers, wide computationalism also problematically entails that physical systems that should otherwise not count as computing systems nonetheless qualify. In short, the causal mapping account underwriting wide computationalism is too liberal.

Fifth, does wide computationalism account for miscomputation? Here, again, the account stumbles. Miscomputation requires that a computational system deliver the wrong output. In the case of a wide computational system, this entails that, for example, with respect an animal's spatial navigation system in Wilson's second example, a wide computational process could be mapped to the wrong output. But this does not appear possible given the causal-mapping account. In each case of wide computation, the computational description will map to the correct physical states (e.g., the animal's solar heading, forward speed and a representation of the solar azimuth). The account cannot but deliver the correct relative position. Part of the problem is that causal mapping accounts can be generated regardless of whether a computation produces the correct result.

Finally, does wide computationalism provide a taxonomy of computing devices? In line with its handling of desideratum (1), wide computationalism does seem to be able, at least in principle, to furnish a categorization of computing mechanisms on the basis of casual powers. This is because only some systems will support physical transitions that can be mapped to a computational description. Wide computationalism provides enough matter of fact about which physical systems support computations to distinguish between the powers of different computing systems.

In sum, as an account of concrete computation, wide computationalism is strong with respect to desiderata (1), (3), and (6), but weak with respect to (2), (4), and (5). Not a bad performance, but things could be better. Part of the reason for this less than ideal showing is the view's allegiance to the causal mapping account. The weaknesses of the casual mapping account carry through to wide computationalism. More positively, the prognosis is that if the underlying account of computation can be updated, then this may supply wide

computationalism with the resources to address the outstanding desiderata, ultimately saving it from the dustbin of promising but unworkable ideas.

## 1.4 Mechanistic Computation

Gualtiero Piccinini (2007, 2015) has recently laid out an account of physical computation that he dubs 'the mechanist account' (see also Milkowski, 2015). In what follows, I argue that the mechanistic account can be used to update the conceptual foundations of wide computationalism. This update allows the view to meet the three desiderata that caused trouble for earlier formulations.

Piccinini outlines three conditions for concrete computation. The first is that physical computing systems must be kinds of functional mechanisms. The system has to possess properties that organize in such a way so as to produce or support some behaviour – the reverse of which is that if a system fails to perform its function it must be the result of a breakdown in the organization of the system's component parts. The second condition is that one of the capacities of a mechanism must be the ability to compute at least one mathematical function. The system must be able to map from an input I (and possibly internal states S) to an output O. The system's behaviour must satisfy at least one abstract description mapping inputs to outputs – this also suffices to show that the system is following a rule.

The final condition is that a physical computing system must compute its function via the manipulation of medium-independent vehicles. This means that informational vehicles – whether they are numbers, symbols, or retinal images–must be transformed over the course of a computation in virtue of a system's sensitivity to some part of the vehicle's structure. So, for example, in the case of numbers or symbols, this would involve processing vehicles in virtue of their syntactic structure; while in the case of neural representations, it would involve processing vehicles in virtue of their systematic relational structure. The point is that if the input–output mapping is sensitive to at least some portion of the medium-independent vehicle over which it is defined, then it counts as a computation.

So, putting these three components together, the mechanistic account claims that concrete computation occurs wherever there is a physical system that has an organization of spatiotemporal components such that it computes an abstract function in virtue of manipulating medium-independent vehicles. As Piccinini describes the view: "Concrete computing systems are physical systems that are functionally organized to manipulate medium-independent vehicles in accordance with a rule that applies to all vehicles and depends on the medium-independent properties of the vehicles (and possibly the system's internal states) for its application" (2015, p.5). The emphasis is on functionally integrated systems that compute at least one abstract function via vehicle manipulation.[17]

An example might help. Consider the neural network in the ocularmotor system responsible for horizontal eye movement (Robinson, 1989; Leigh and Zee, 2006). This system exhibits all the characteristics of a mechanistic computing system. First, the system is a causally integrated connection of spatiotemporal components (neurons) poised to produce some behaviour (eye movement). It is a functional mechanism. Second, the system computes at least one abstract function (an integration relation). It does so in virtue of preserving the relationship between eye-velocity and eye-position. Third, the system computes its function via the manipulation or transformation of medium-independent vehicles: information contained within the cortex. Morphic relations between eye-velocity and eye-position are manipulated to compute horizontal eye movement. The ocular-motor system satisfies each of the three conditions on mechanistic computation. The mechanistic account is also distinct from causal mapping accounts on at least two fronts.

First, it holds that computing systems are functional systems of a specific mechanistic type; second, it holds that computation is achieved through the use of medium-independent vehicles. Although the causal mapping account acknowledges the importance of causally integrated systems, the mechanistic account takes this condition further. This is because in addition to

[17] There is more to Piccinini's account then what is presented here. However, this description should suffice to outline the basic features of the view. For a fuller description see Piccinini (2015, ch.7).

specifying in what ways physical structures can be processed, the mechanistic account also requires substantive organizational integration. It places more stringent conditions on when physical states can be interpreted as performing computations than the causal mapping account.

As promising as all this sounds, before the mechanistic account can serve as the basis for an updated wide computationalism, it also has to be shown that it can satisfy the six desiderata previously outlined. Consider each desideratum in turn.

First, and perhaps unsurprisingly, the mechanistic account settles the objectivity question. The specification of several conditions for concrete computation makes it a matter of fact as to whether a given physical system is a computing system – correspondingly, this also means that the account provides a solution to the problem of computational implementation.

With respect to the second desideratum, the account is successfully able to navigate the computational explanation versus modeling distinction. To see why, consider the ocular-motor example again. There, it was in virtue of the integration function that the system was able to compute horizontal eye movement. The system implemented a program or rule in the service of particular behaviour. The output of the system, horizontal eye movement, was the direct result of the function computed. One of the system's functions is to compute eye position using an integration relation – the system instantiates a computational procedure rather than simply being described as having one.

How about desiderata (3) and (4)? Here, again, the mechanistic account is more resilient than its causal mapping counterpart. Recall that the problem for the causal mapping account was that it was unable to exclude non-paradigmatic cases, e.g., respiration, solar systems, etc. The reason was that it was too liberal in its mapping conditions; a large number of non-computing systems qualified as computational in virtue of having the right kind of causal transitions between physical states.[18] In contrast, the mechanistic account strikes a better

[18] This is why science is filled with computational models. The mathematics of computation is particularly effective at describing and predicating natural phenomena (see Frigg 2012).

balance. It places further restrictions on physical computation. Consider, for example, the digestive system. Though the digestive system is appropriately described as a functional mechanism, it fails to meet an abstract functional description in terms of computing medium-independent variables. It therefore fails to qualify as a computing system. The addition of the functional mechanism and medium-independent processing requirements serves to exclude the wrong sorts of physical systems.

Fifth, the mechanistic account explains how a computing system can miscompute. This is because it focuses attention on how mechanisms can break down. Piccinini (2015, ch.7) points out that a computing mechanism is evaluable according to five perspectives: (i) how its designers intended it to compute a function, (ii) whether it is actually designed to compute a particular function but was built in such a way that it actually computes a different function, (iii) whether what was built malfunctions, (iv) whether the system is misconfigured (e.g., programed wrongly), (v) and whether the system is used incorrectly. In each case, whether through fault of designer, builder, or user, the miscomputation rests with the functional organization of the system. Functional organization is indispensable to the construction, execution and interpretation of computing systems. By requiring physical systems to be functional mechanisms, the mechanistic account successfully offers an explanation of miscomputation, because it draws attention to how the integration of spatiotemporal components drives when, how, and why physical systems miscompute.

Finally, the mechanistic account can taxonomize computing systems. This follows in virtue of its ability to distinguish between various mechanistic properties. For example, 'being programmable' requires having certain transducers and storage capacities. Computing systems that fail to have these properties might still compute, but may not be programmable – most Turing-machines would meet this description. By appealing to properties of functional organization, the mechanistic account takes advantage of the fact that mechanistic properties have computational implications. It seems that the mechanistic account, at least as developed by Piccinini, provides a robust account of concrete computation. Not only does it provide a

solution to the problem of implementation, but it also does so in a way that does justice to the sciences of computation.

## 1.5 Wide Mechanistic Computation

The question to consider is whether it is possible for physical computing systems of the type described by the mechanistic account to include elements outside the individual – that is, whether the mechanistic account can be squared with wide computationalism.

The answer turns out to be rather straightforward. The reason is that, similar to the casual mapping account, the mechanistic account also remains neutral about what physical parts of the world can be integrated so as to form a physical computing system. Similar to the causal mapping account, the method of computational individuation is location neutral. Whether or not a functional mechanism, one that processes medium-independent vehicles, is constituted by spatiotemporal components squarely localized within the individual or crisscrossing into the world is entirely an *a posteriori* question. Some physical computing cognitive systems might be entirely ensconced within the body, but some might as easily spread out over brain, body and world.

Piccinini even acknowledges this possibility in a footnote to his chapter on the mechanistic account: "I am officially neutral on whether the components of psychological computing mechanisms extend beyond the spatial boundaries of organisms" (2015, ch.7). Piccinini recognises that whether psychological computing mechanisms are constituted in part by components outside the individual is an empirical question. Because the mechanistic conditions on concrete computations are medium and location neutral, the question of wide computational systems is open. The real task, then, is to show that some organisms in fact implement wide mechanistic computing systems.

There are two examples I want to draw on in establishing the empirical plausibility of wide mechanistic computation. In each case, the focus is on active sensory systems.

The first is sonar-emitting bats. Bats have become a staple example of active sensory systems since it was discovered that they hunt using self-generated acoustic signals (echolocation) (MacIver, 2009). Consider one aspect of echolocation: object detection. Bats use two different methods for object detection. On horizontal planes, bats use time and intensity differences in the returning acoustic signals to detect objects; while on vertical planes, the bats' inner ear, the pinna-tragus, forms a pathway through which the incoming signal is filtered. The formation of the skin and supporting tissue transforms the signal into a range of spectral cues, which then get further processed neuronally.

There are two points to note about this example. First, the acoustic signal is more than just a passive input to the bats' navigation system. The propagation of acoustic pulses actively drives obstacle detection. Second, it is the neural processing plus morphology and acoustic environment that facilitates object detection. The neuronal processing alone is insufficient. What this suggests is that the bats' ability to detect objects along vertical planes is realized by spatiotemporal components spread out across the brain, body and world. The bats' spatial navigation system is supported by a wide mechanism.[19] This is the first condition on mechanistic computation.

Why view these components as forming a wide mechanism rather than casually related but distinct elements? The answer lies in the high degree of organization and structure exhibited by the components. It is only through the coordinated activity of parts spread across the brain, body and world that vertical object detection is achieved. It is not just that the acoustic signal and morphology of the bat that play a role in the delivering inputs to the internal processing. It is that both actively construct and transform the information latter used for internal processing. They are part of the underlying causal mechanism responsible for the animals' perceptual capacity.

[19] For another example of a wide mechanism see Wilson's (2010, 2014) discussion of the giant water bug: Lethocerus.

What about condition two? What abstract function is being computed? The answer here is one already encountered: object detection. Because there is a clear mapping from the acoustic signals outside the bat to the internal outputs (representation of objects in the vertical plane) via the simplifying structure of the pinna-tragus, the wide mechanism can be said to compute at least one abstract function. The difference in this case is that the input resides neither on nor in the bats' sensory transducers. Rather, it is part of the environment. The acoustic signals on which the bats' inner ear operates already contain information about objects in the environment.

Finally, Piccinini also claimed that a mechanism must compute a function in virtue of manipulating some portion of medium-independent vehicles. Gibson's (1966, 1986) notion of 'invariants' offers a useful starting point here. For Gibson, interactions of mechanical forces often produce systematic, structural regularities in the environment. These invariants are picked up by organism's perceptual system. They are used to guide, sustain, and regulate behaviour. For example, because light is constantly diffused and reflected throughout the environment, there are optical textures that overlay surfaces. The amount of texture corresponds to the amount of terrain. As the density of optical texture increases, the scale of the space is revealed. Optical textures provide crucial information about the environment that an organism's perceptual system can use to gauge distance.

Acoustic signals operate in an analogous way. Because sonic energy refracts and diffuses through gases, liquids, and solids, the arrangement of the environment and the medium of transmission shape the vibratory field in which bats navigate. It structures acoustic signals, providing important cues about object location and distance. This means that the vehicles transformed over the course of processing are external information-bearing structures. They are medium-independent vehicles that are persistent and unchanged over time and which carry information about the environment. Manipulation of the informational vehicles runs from the environment through the pinna tragus to the bat's brain.

Notice that the external vehicles are not medium-independent in virtue of the fact that they carry information, but rather in virtue of the fact that the relevant computations are sensitive to only some portion of the informational vehicles, i.e. the invariant structure of the acoustic signal carrying information about location and direction. What matters for computation is not that the bat is responsive to the sound qua sound, but that the bat is responsive to the sound qua information-bearing structures within the sound. As Piccinini explains: "Since concrete computation and their vehicles can be defined independently of the physical media that implement them, we shall call them 'medium independent' (2011, p.8). The abstract character of computational descriptions means that medium-independence follows in virtue of the relevance of specific parts of physical media to the overall computation being carried out, whether that is syntactic structure in symbols or invariant relational structure in sound.

Given the above, there seems to be good reason to think that the bats' navigation system instantiates a wide computational system of the mechanistic variety. Not only does it involve a mechanism that spans the brain, body and world, but it also computes an abstraction function via the transformation of medium-independent vehicles. It meets all three requirements of the mechanistic account.

Animal cognition is one thing, but is there a human example? Continuing with the idea of active sensory systems, the next case to consider is spatial navigation by sightless or blind individuals.

One common assumption is that sightless individuals are at a greater disadvantage than sighted individuals during spatial navigation because of a lack of crucial visual information (Lynch, 1960). Several studies have recently begun to cast doubt on this assumption, as spatial competence has been found to be increasingly less dependent on visual experience than initially thought. Ricciardi et al. (2009), for example, have shown that the sound of an action engages the mirror neuron system for action schemas, even when not learned through the visual modality.

One important take away from this research is that it suggests the use of a supramodal sensory representation in spatial navigation. This result is important because it comports well with how spatial navigation is mainly achieved within sightless individuals – that is, through the collection of spatial information via haptic and audio channels. A critical component of this collection process is the systematic feedback of haptic information through prosthetic devices, such as canes or personal assistive devices, in addition to hands, palms and fingers – canes, for example, provide low-resolution information about the immediate environment through a semi-spherical exploratory sweeping motion. What I want to suggest is that, similar to the bats' sonar emitting echolocation system, the prosthetic-aided navigation system can be thought of as a wide mechanism, one that supports an abstract function through which medium-independent vehicles are manipulated. Notice how the example meets the three mechanistic conditions.

Consider the first condition. Under normal circumstances, it is undoubtedly correct to say that spatial navigation is supported by an internally constituted mechanism. However, in the case of sightless individuals, the breakdown of internal components requires recruitment of external substitutes. This is the role occupied by prosthetics such as canes or personal assistive devices, which can even include sonar-emitting devices (Lahav and Mioduser, 2008). Under these conditions, the integrated spatiotemporal components form a wide mechanism. They form an integrated mechanism that spreads out beyond the neuronal.

Next, consider what function is computed by this wide mechanism. One possible answer is that the haptic information delivered by prosthetic devise aids in the construction of a 'survey representation', a disposition or layout of spatial features – direction and distance, for example (Loomis et al., 1993). The purpose of this representation is to facilitate finding trajectories or routes through the environment. Interpreted in this way, the abstract function is one that runs from the environmental signals generated from the repeated tapping of the prosthetic device to the internal spatial representation of the environment. There is a mapping of inputs I to output O via the prosthetic device.

Lastly, the spatial navigation system of sightless individuals involves manipulating medium-independent vehicles. The active exploratory strategies used by sightless individuals transform and manipulate environmental structures, particularly sonic and tactile information, in order to simplify and reduce internal processing. Much like the pinna-tragus of the bat, the prosthetic devices form morphological tools through which information is externally directed. There is a transformation of sensory information through external then internal structures. Once again, these considerations point toward the presence of a wide mechanistic computational system.

Actual examples of wide computational systems, such as the above, are important. As Segal points out in his review of Wilson: "the question of the truth of wide computationalism is a question about the proper domain of psychological theories (or at least cognitively scientific theories), it is a question about the extent of the natural phenomenon of cognition" (1997, p. 154). Without a demonstration of wide physical computing systems, the idea of wide mechanistic computation remains plausible but unsubstantiated. What I have tried to do, like Wilson and Hutchins beforehand, is show that in some cases wide computational systems are, in fact, implemented in cognizing agents.

## 1.6 Objections

Time to consider some possible concerns. First, one might worry that wide computationalism, even of the mechanistic variety, fails to offer principled grounds on which to distinguish internal (brain bound) cognitive systems from wide cognitive systems. Given that a fair amount of cognitive science attempts to distinguish the underlying systems supporting psychological behaviour, wide computationalism should be able to make such an important differentiation.

Two points are worth considering here. First, recall that the computational theory of mind leaves open the possibility that some computational systems are realized by elements outside the individual. That is at least part of the moral of classic multiple realizability. Second, recall

that any successful account of concrete computation is going to have to provide a taxonomy of different computing systems (desideratum six). Together, these points offer one potential answer to the 'being principled' concern. Since cognition is a form of computation and computation is location neutral, then insofar as one is able to identify how computational systems are implemented via the account of concrete computation offered, one has principled grounds on which to distinguish internal and wide cognitive systems. Individuating internal and wide systems simply requires identifying the right kind of conditions on physical computation.

Consider, again, the two previous examples. There, the two systems qualified as physical computing systems in virtue of meeting all three conditions of mechanistic computation. The only difference was that in the one case the system was partly constituted by parts in the world, while in the other it was completely contained within the individual. Insofar as wide computationalism relies on an account of concrete computation that is sensitive to differences in mechanistic properties, it can distinguish between computing systems instantiated both within and beyond the individual.

Consider a second concern. One debate that is near and dear to the heart of many philosophers of mind is whether cognitive states represent intrinsically (in virtue of themselves) or whether they represent in virtue of having meanings assigned to them. The question is whether the mind has 'original' or 'derived' intentionality (Searle, 1980, 1983; Dretske1981). Segal (1997) raises concerns about original intentionality in the context of wide computationalism:

> Someone who claims that original intentionality is restricted to brains (or things enough like brains) and certainly not something present in pieces of paper, or even pocket calculators, is likely to be unimpressed by wide computationalism. She would likely draw a distinction between cognition proper and mere computation. Cognition proper would be restricted to systems the symbols of which are originally intentional. (p.153).

The worry is that whereas cognition trades in original content, wide computational systems trade, at least partly, in derived content. Since wide systems do not deal in the right kind of

content, they should not be thought of as properly cognitive. There is a long-standing debate over original intentionality, and this is not the place to enter into the discussion. Suffice it to say, several authors have questioned the claim that original intentionality forms any sort of 'mark of the cognitive' (Dennett, 1987; Wilson and Clark, 2009).

For present purposes, what is important is that although semantic accounts are compatible with computation, they in fact still require non-semantic individuation. Functional and structural (non-semantic) properties, such as programming language or architecture, are also going to contribute to the individuation of physical computations. Individuating computational systems by function alone – by what they represent – will often fail to individuate physical computations finely enough. As long as semantic information is built out of syntactic structure in some way, which is not only plausible but the going view of most cognitive scientists, wide computationalism is going to remain plausible. Even if it turns out that semantic accounts of computation are correct, the syntactic underpinnings of semantic information still deliver wide computationalism. Segal's worry, then, is too quick. It fails to appreciate that computational cognition still is going to imply syntactic considerations of the kind pertinent to wide computationalism, whether or not cognition turns out to trade in original or derived content.

Finally, one might worry more generally about the value of wide computationalism. Does wide computationalism have anything to add as a research strategy to cognitive science? Wilson (1994) has something like this concern in mind when he says: "the most interesting issue concerns not the coherence of wide computationalism but the extent to which a wide computational research strategy is and could be employed within cognitive psychology" (p.371).

Here is one way wide computationalism might be of wider use to cognitive science: the ongoing debate over extended cognition. Some opponents of extended cognition have suggested that environmental and bodily processes and states are too unwieldy to be brought under a framework that also contains neural processes (Adams and Aizawa 2008; Rupert 2004, 2009). The 'motley crew' of extended cognition undermines its chances of developing a

scientifically tractable approach to cognition. Taking wide computationalism seriously affords one answer to this challenge. This is because wide computationalism provides a framework for investigating concrete physical computing systems that cross into the world. Parts of the body and world can be integrated into and thought of in terms of performing computations, insofar as they are part of wide computational systems. Wide computationalism offers a potential rubric from within which to conduct extended cognition research. One of the potential uses of wide computationalism is its ability to link individualist forms of computational psychology with more externalist-friendly extended approaches. It offers another prospective ship on which to navigate the conceptual waters of cognitive science.

## 1.7 Conclusion

Time to take stock. I set out to show that wide computationalism offered an important but overlooked option for cognitive science. I attempted to show that it offered not only a coherent and plausible account of concrete computation, but that it also found empirical support from examples in animal and human cognition. I further developed the view by defending it from several potential objections and charting its potential use for cognitive science more generally. This is by no means the final word on wide computationalism. Much more work is required – for example, it still remains unclear how wide computationalism fits with other core concepts in cognitive science and philosophy of mind such as functionalism, intentionality or representation. Nonetheless, what the preceding discussion has done, I hope, is convey the sense that wide computationalism offers a substantive and viable supplement to existing individualistic research strategies; that it provides a plausible and theoretically fruitful avenue for cognitive science to further explore. Summatively speaking, it would not be unfair to say that the prospects of wide computationalism are looking up.

# Introduction to Chapter 2

## Key ideas

Chapter 2 picks up where Chapter 1 left off, continuing to develop the case for wide mechanistic computation. It is guided by two central thoughts.

The first is that wide computationalism can be further enhanced by incorporating additional concepts from the mechanistic framework. I focus on two concepts in particular: *constitutive explanation* and *constitutive relevance*. Constitutive explanation is the idea that mechanistic explanations not only have to accurately describe the task or competence being performed by a mechanism, but they must also provide insight into the mechanism's component parts and activities. Applied to computation, the idea is that after the abstract function-theoretic description has been provided for some phenomenon, a structural description of the computational mechanism's component parts and activities needs to take place. Constitutive relevance, on the other hand, is the idea that there is a method for determining whether a component's activity is constitutively relevant to another component's activity. The thought, from Craver (2007), and the one adopted here, is that two components must be part-whole related and mutually manipulable in order to be constitutively relevant. The general picture of wide mechanistic computation begun in Chapter 1 is further fleshed out here using the notions of constitutive explanations and constitutive relevance.

The second thought is that wide computationalism, contrary to some suggestions, is, in fact, explanatorily and experimentally desirable. What I mean by this is that wide computationalism, particularly of the mechanistic variety, is not simply an abstract philosophical gloss on existing computational research and practice, it is also a substantive

explanatory and methodological position researchers might adopt. This is not to say that it is a fully worked out research programme; far from it. But it is to say that wide mechanistic computation has something important to add to cognitive science. It is not only a conceptually coherent and empirically plausible position, but it is explanatorily and methodologically attractive as well.

## The argument

To establish this last point, the chapter focuses on responding to two outstanding worries facing wide computationalism. The first is what I label the *parsimony* challenge. The parsimony challenges says that, other things being equal, we should prefer an internalist account of computational mechanisms, i.e. computational mechanisms ensconced entirely within the individual. Not only is the internalist position explanatorily simpler, but it is also ontologically less costly. It doesn't require positing unwieldy brain-world hybrids. The second worry is what I label the *testability* challenge. The testability challenges says that there is an outstanding question about how useful wide computationalism is as an empirical theory; that it is unclear whether wide computation as a whole offers anything empirically interesting to cognitive science. This is because, regardless of whether the view turns out to be conceptually coherent or not, it doesn't seem to be empirically testable.

In response to the two challenges, I argue that the notions of constitutive explanation and constitutive relevance can be pressed into service. First, the notion of constitutive explanation helps to undermine the 'other things being equal' assumption of the parsimony challenge. The parsimony challenge argued that we have reason to prefer internalist explanations, on balance, because they are ontologically simpler than their wide counterparts. However, if accurately describing the computational task is only one part of providing a good computational explanation, and computational explanations also require structural detail, then wide explanations may be better or worse than internalist ones depending on the amount of insight they allow into a mechanism's underlying component parts and activities. Incorporating a

notion of constitutive relevance makes it so that the computational explanation that best elaborates the underlying mechanism, which might in some cases be wide, is explanatorily superior. Second, the notion of constitutive relevance helps to connect wide computationalism to several experimental methods. This is because it opens up an array of possible 'interlevel' experiments – various types of interventions we might use to investigate mechanisms – to wide investigations. Adopting a notion of constitutive relevance offers a means to identifying the distributed computational mechanisms relevant to wide computation. It provides a clear hook into experimental investigation.

However, there is a preliminary worry I want to address. For one might not be entirely convinced by the challenges I've outlined. One might think, for instance, that the two challenges follow more from misguided ideas about parsimony and empirical testing than from genuine philosophical dilemmas.

One of the main reasons we should take these challenges seriously, though, is that we already accept analogues in debates about extended cognition. For example, as I point out in Section 2.2, Rupert (2004) suggests that we can accommodate all of the relevant facts about cognitive processes, such as how they are dynamically coupled with environmental structures, without incurring the additional ontological cost of moving to extended cognition. The facts can be handled by a simpler view, what Rupert calls the 'embedded' view of cognition. Similarly, Wilson (2002) raises questions about the empirical tractability of extended cognition, writing: "If we recall that the goal of science is to find underlying principles and regularities, rather than to explain specific events, then the facultative [temporary] nature of distributed [extended] cognition becomes a problem" (Wilson, 2002, p.631). Given their temporary nature, extended systems don't look like they are going to be empirically very fruitful.

What this shows, I think, is that one either has to outright dismiss the extended mind debate or one has to admit that there is something genuine about the two challenges. If the two challenges are genuine philosophical problems in the context of the extended mind, then the

same should be true of their wide computation counterparts. Of course, one could always accept the first option and reject their use in extended mind debates, but I think this would be a needlessly hefty price just to avoid admitting the significance of the challenges.

## Broader context

Finally, it will be worth saying a bit about more the relation between wide computationalism and other views one might hold about computation, such as individuation or implementation views. For it might not be entirely clear whether such views are competitors or distinct from wide computationalism (see also the Introduction).

As I formulate the view here, wide mechanistic computation is neither, strictly speaking, a theory of implementation nor a theory of individuation. Instead, it is a general claim about the scope of computational analysis. It is claim about the intrinsic logic of computational analysis, and what type of systems it might apply to: namely, world-spanning ones. In the same way that extended functionalism (the basis for some forms of extended cognition Clark [2008]) is not a rival to causal theories of content but, rather, is an offshoot of functionalism, so too is wide computationalism an off shoot of generic computationalism. For this reason, it is not a competitor to semantic or causal mapping accounts of implementation or individuation. This fact also helps elucidate the relation between wide computationalism and mechanistic computation. For while wide computationalism is a subset of computationalism and mechanistic computation is a subset of computationalism, wide computationalism is not a subset of mechanistic computation, nor the reverse. The two views form mutually independent position. The relation is one of non-overlapping Venn diagrams. This is why an integration is both novel and interesting. It sharpens up the conceptual lines between two otherwise distinct but interesting positions within the logical space of ideas.

# Chapter 2 – Two Challenges to Wide Computationalism

## 2.1 Introduction

The 'mechanistic account of wide computation' has revealed a general compatibility between two approaches to computation: mechanistic views, such as Piccinini (2007, 2015), which hold that computation is a function of concrete computing mechanisms, and so-called 'wide' approaches to computation, such as Wilson (1994, 2004) and Hutchins (1995), which hold that computational processes spread out beyond the boundaries of the individual.[20] This account claims that computational processes are *wide* in virtue of being part of *mechanisms* that extend beyond the boundary of the individual (Kersten, 2017a; Nowakowski, 2017).

Milkowski et al. (2018), for instance, have recently proposed that 'mindshaping' – the process of making interactions with conspecifics easier to interpret – could be driven by a wide computational mechanism: "Rather than tackle the seemingly computationally intractable task of predicting our con-specifics by building ever more complex, intracranial computational capacity, natural selection seems to have developed systematic means of structuring the social environment in ways that make it easier to predict using relatively simple intracranial resources". On their view, the computational mechanism that underlies mindshaping is not to be found exclusively inside the brain.[21]

One of the proposed benefits of combining the mechanistic and wide view of computation is that it helps to resolve a number of problems that troubled previous formulations. For

[20] This view has gone by a few different names in its time. Wilson (1994, 2004), calls it "wide computationalism", Losonksy (1995) labels it "environmentalism", Wells (1998) dubs it "interactive computation", Shagrir (2018) labels it "computational externalism". For consistency and simplicity, I stick with Wilson's original moniker.
[21] Similar claims have also been made about emotions and music cognition, see, e.g., Krueger (2014), Kersten (2017b), and Kersten & Wilson (2016).

example, according to Wilson (1994) and Hutchins (1995), wide computational processes are those that involve isomorphic mappings between abstract computational states and physical states involving the individual's brain, body, and environment. Wide systems are those that mirror transitions among abstract states in a formal computation.

One issue for traditional wide accounts is that their commitment to 'causal mapping' accounts make it difficult for them to provide an adequate *computational taxonomy*.[22] Casual mapping accounts maintain that physical systems implement computations only when they mirror the causal transitions among the states of a computational model (Chalmers, 1994; Chrisley, 1995; Scheutz, 1995). The trouble, as several authors have pointed out, is that such accounts are too liberal. They underspecify the possible mappings between computational and physical states. This means that they allow a number non-paradigmatic cases, such as solar systems or digestive systems, to qualify as computing systems (Milkowski, 2013; Piccinini, 2015).

Mechanistic accounts of computation, on the other hand, do a better job accommodating such cases. Since concrete computing systems are only those that implicate functional mechanisms processing medium-independent vehicles, non-paradigmatic cases, such as solar systems or digestive systems, fail to qualify as computing systems, while paradigm cases, such as Turing machines and digital computers, continue to qualify, as the latter but not the former involve the right kind of functional mechanisms (Kersten 2017a).

But in spite of such encouraging results, a number of worries remain for a mechanistic account of wide computation. One is whether the view can overcome rival 'internalist' accounts. That is, even if one admits the existence of wide computational mechanisms, it seems that one should still remain an internalist about computational mechanisms, as it seems

[22] This alliance is part historical, part conceptual. Historically, causal mapping accounts constituted the major approach to implementation. As such, early formulations of wide computation largely took their lead from these accounts. Conceptually, causal mapping accounts were, and still are, committed to medium-independence, the idea that computational analysis is indifferent to medium. Wide computation exploits medium-independence in order to get a conceptual foothold. It is the explanatory indifference of computational analysis to material composition that opens up the possible of applying computational analysis to environmental features.

to be the simpler, and less ontologically costly, of the two accounts. I label this the *parsimony challenge*. Another worry is that even if there are reasons to prefer wide computational explanations, there still seems to be a residual concern about whether wide computation as a whole offers anything empirically interesting. That is, there seems to be a question about the view's bearing on possible experiments, what I label here the *testability challenge*.

The goal of this paper is to develop a mechanistic account of wide computation by responding to these two challenges (*parsimony* and *testability*). My aim is to show that not only that positing wide computations is not ontologically costly, but that it is also empirically testable. My argument is that wide mechanistic computation can gain the necessary conceptual resources to address the outstanding challenges by embracing two aspects of the mechanistic approach to computation: 1) the structural aspect of mechanistic explanation and ii) the notion of constitutive relevance.

I begin by laying out the two challenges, showing how they apply to previous formulations of wide computation (Section 2.2). Next, to further flesh out the view, I outline a method for identifying distributed functional mechanisms based on Craver's (2007) work on 'constitutive relevance' (Section 2.3). Then, with an updated version of the mechanistic account in hand, I return to the parsimony and testability challenges (Section 2.4). I argue that the first challenge can be addressed by undermining its underlying 'ceteris paribus' assumption via appeal to the structural aspect of mechanistic explanation, while the second challenge can be addressed by drawing on the connection offered between 'constitutive relevance' and experimental investigation within mechanistic investigations. Finally, I conclude by responding to possible objections (Section 2.5).

## 2.2 Two Challenges

The parsimony challenge asks why someone should be moved to posit wide mechanisms if there always is a simpler, ontologically less radical position available: namely, internalism, the

claim that the whole subject minus the environment is the largest integrated system available for a computational study of the mind.

An analogue of the parsimony challenge can be found in debates about whether cognitive processes extend beyond the individual. Consider two views: one which says that cognition depends heavily on external props and devices in the environment (embedded cognition), the other which says that cognition, in virtue of its dependence on external props and devices, extends into the environment (extended cognition).

Rupert has the following worry about such views:

> We can grant that cognition often involves intimate interaction with its environment…This way of putting matters, however, is best accommodated by HEMC [hypothesis of embedded cognition]; and given the costs to intuition – and to the general principle of conservatism in theory acceptance – of spreading the mind out into the world beyond the organism, there seems no reason to reinterpret the situation in keeping with HEC [hypothesis of extended cognition]. (2004, p.405).[23]

HEC and HEMC posit the same causal processes and interactions between the brain and external environment. However, HEC claims that some of those causal processes are also extended cognitive processes. The problem is that this ontological inflation appears to reap no additional explanatory reward. The same explanatory work can be done by HEMC without the ontological inflation. If the appeal to hybrid cognitive/environmental systems do no extra explanatory work, then we should prefer the more parsimonious HEMC.

Substitute the word 'computation' for 'cognition' and a similar worry applies to positing wide computations. Consider two views: one which says that cognitive computation heavily depends on external structures in the environment (the hypothesis of internal computation)

[23] As Sprevak (2009) points out, there are two virtues to keep distinct here. One, which relies on a principle of conservatism, says that a theory is preferable if it proposes minimal disruption or revision to existing thinking. It is backward looking: it speaks to a better fit of a new theory with the existing picture. The other, which deals with considerations of simplicity, says that a theory is preferable if it proposes a more minimal ontology. It is sideways looking, in that it speaks to the merits of a new theory relative to its current alternatives. Sprevak (2009) argues that conservatism as a virtue should count for little in deciding between internalist and wide approaches to explanation. It mistakenly suggests that internalism is the output of mature scientific theory, rather than an import into cognitive science from outside. Conservativism is generally only a virtue when there are good reasons for accepting the old theory. However, since internalism about computation has never between explicitly considered by cognitive science, but largely assumed, it is unclear that it should weigh heavily in the assessment. For this reason, I focus on simplicity considerations alone in what follows.

(HIC); and another which says that cognitive computation, in virtue of its causal dependence on external structures, extends into the environment (the hypothesis of wide computation) (HWC).

If HWC and HIC are competitors, and each describes the same computational activity via the same set of causal events and interactions, then we should prefer HIC insofar as it offers the ontologically more parsimonious view, e.g., it does not appeal to hybrid agent/environmental entities. Even if the mechanistic account of wide computation is coherent, there is still a theoretically simpler way available to accommodate its insights.[24]

To illustrate, consider the case of cricket phonotaxis. To identify and localise potential mating partners, a tracheal tube, which connects the inner ears of the female cricket, inhibits certain informational relationships in the environment, such as those outside the 4-5 kHz range, which, in turn, facilitates the amount of information poised for motor control by cricket's brain. There is a systematic relation between the structure of the cricket's acoustic environment, its morphology, and its neural and auditory systems (Webb & Harrison, 2000; Webb, 2008). Such examples have provided grist for the mill of wide computationalists. Not only do they highlight the intimate role of the body and environment in complex, informationally-laden tasks, but they also highlight how the resultant functional whole might be treated as a distinct unit of analysis (Kersten, 2017a).

The problem is that there does not seem to be a clear explanatory advantage to this move. The same processes and behaviours could be described under both internal and wide accounts. Why not, then, simply treat the case as an instance internal computing that is heavily dependent on the environment? Other things being equal, simplicity suggests we should limit the expansion of our ontology, accept internalism, and reject the wide view.

[24] I am using the word 'internalism' instead of 'embedded' not because the views aren't roughly the same for present purposes, but simply because it brings out the contrast better. I will also intermittently use the term 'narrow' when referring to internalist explanations, but nothing is reflected in this change of verbiage.

The testability challenge asks, even granted there is some reason to prefer a wide computational approach, should this view be of interest to empirical cognitive science? As Wilson (1994) frames the concern: "the most interesting issue concerns not the coherence of wide computationalism but the extent to which a wide computational research strategy is and could be employed within cognitive psychology" (p.371). Since we want our theories to not only be descriptively adequate but also experimentally informative, we should be wary of any account that does not supply a means for empirical testing.

Consider Losonsky's (1995) account, for example. Losonksy argues that there are a number of empirical cases that demonstrate that "subjects plus structures in the environment and relations to them form integrated computational systems" (p.364). For example, in discussing Rutkowska's (1993) research on infant development, Losonksy argues that feedback loops in grasping and fixating behaviour point to 'action systems' that are constituted by internal and external components. What such research shows is that "we can no longer treat behavior such as grasping or fixating on an object as a molar entity without an internal structure. Instead, we will have to analyze it into sequences of steps that involve internal as well as environmental structures" (p.364). In short, we need to look for wide computational systems and processes.

But notice that Losonksy provides little direction about how to do this. The account is largely occupied with providing a re-description of existing empirical work. It offers an informal gloss on phenomena that heavily depend on environmental structures. This is understandable, but the question still remains as to how to test for wide systems. Are all cases of systematic feedback loops between subject and environment instances of wide computing systems? Only some? As we will see, similar points hold of other accounts of wide computation. As it stands, wide computationalism does not offer much in the way of empirically testable consequences.

These are our two challenges, then. The first is based on worries about wide computation's ontological cost; the second on worries about the view's empirical testability. Notice that the two challenges are distinct. Even if it turns out that there are reasons to prefer wide

computational ontology, it is still legitimate to wonder whether the view is empirically testable.

### 2.2.1 Two Challenges Applied

To drive home the force of the challenges, consider how a specific account of wide computation fares against each challenge.

Recall that for Wilson's (1994) account, wide computational processes and systems follow from the substrate neutrality or medium-independence of computational analysis – that is, wide computation gains a conceptual foothold via the indifference of computational analysis to physical media.[25] Finding wide processes requires identifying isomorphic mappings between physical states inside and outside the individual and transitions among states within a computational model. Environmental structures are apt for wide treatment when they possess certain formal properties. For example, Wilson points to several cases, such as spatial navigation in animals, which involve re-descriptions of formal, geometric properties of the environment, such as the solar-azimuth. [26]

Does Wilson's account offer grounds for preferring wide explanations over internalist ones? It doesn't appear so. Even granting that geometric features are apt for wide treatment, the question remains as to whether such properties constitute proper parts of a computational process or simply parts of an external causal process that triggers an internal computational process. If the same set of processes and activities, formal or otherwise, are described by narrow and wide computational explanations, then it seems that, other things being equal, we should prefer the account that results in the least ontological inflation.

[25] This expression is typical of other accounts of wide computationalism which also rely on substrate neutrality. Hutchins (1995), for example, writes: "Marr never intended his [computational] framework to be applied to cognitive processes that take place inside an individual, but there is no reason, in principle, to confine it to such a narrow conception of cognition" (p.50).

[26] As mentioned, most accounts of computation rely on mapping accounts of computation, such as Hutchins (1995) and Losonksy (1995), so the worries I lay out here should also spell trouble for these views as well.

One line of response that Wilson gestures at but does not pursue is to argue that wide computational explanations are, in fact, favoured by other, more general explanatory virtues, such as 'theoretical appropriateness' and 'causal depth'(see Wilson 1994, p.368).[27]

First, an explanation is said to be theoretically more appropriate if it characterises a phenomenon at the right level of description. For example, to take a case of theoretical *in*appropriateness, it seems to make more sense to describe a job's candidate's failure of an interview via the resources of economics and sociology than it does neurology. An explanation at the level of neurology would seem to fail to adequately describe the phenomenon in question. Theoretical appropriateness can be thought of as a horizontal dimension of explanatory power. It concerns the relationship between the explanans and explanandum.

Second, an explanation is said to have more causal depth if it describes a counterfactually more robust set of properties in a system. For example, an explanation that describes the locomotion of the passive dynamic walkers (agents that move without drawing on an energy supply) without referencing the worldly structures that generate and sustain their walking behaviour (i.e. the terrain) is counterfactually less robust than an explanation that does. The explanation will be at a loss to account for how the walker fails to walk on surfaces not conducive to the natural dynamics of the agent. Causal depth can be thought of as a vertical dimension of explanatory power.

Wilson might argue that when weighed against parsimony theoretical appropriateness and causal depth tip the scales back in favour of wide computational explanations. The reason why is that both theoretical appropriateness and causal depth are typical of wide explanations. Wide explanations are focused on capturing a certain degree of coarseness in their explanations. Unlike narrow explanations, which focus on local, intrinsic properties, wide explanations appeal to the contextual, non-intrinsic properties of a system, e.g., the terrain used by

[27] Wilson raises this prospect but that moves on to argue from the specific cases under consideration. Also, while Wilson's (1993) initial discussion is directed at psychological explanations, I am adapting it slightly here to apply specifically to wide computational explanations.

dynamical walking systems. They are more likely to exhibit theoretical appropriateness and causal depth. Thus, even if internalist explanations are simpler, there are still independent reasons to prefer wide computational explanations on balance.

The problem is that while theoretical appropriateness and causal depth may give us reason to favour wide explanations in general, they seem ill-suited to help computational explanations specifically. What would it mean, for instance, to say that a computational explanation is more or less theoretically appropriate? One thing it would seem to require is an ability to designate some phenomenon as more properly 'computational' than others. This is presumably what made sociological accounts more theoretically appropriate than neurological ones in explaining a failed interview, for example; only the latter supposedly described the phenomenon at the right level of description.

However, what phenomena are apt for computational treatment is exactly what the dispute between the internalist and wide computationalist is about. The internalist's suggestion is that only individual bound systems and processes are proper subjects of computational treatment, while wide computationalists argue that hybrid agent-world spanning systems and processes should also qualify. We cannot decide ahead of time which is more appropriate, or else this would beg the question in either direction. Whatever else its merit, theoretical appropriateness does not seem to get the wide computationalist very far in responding to the parsimony challenge.

Second, notice that computational explanations already single out those properties that are causally robust. Simply what it means for a physical system to compute is that a system can be described in terms of supporting causal state transitions that mirror transitions among computational states of a model. There is no sense in which computational explanations could be causally deeper, because what it is for property to be causally robust is that it can be mapped to a computational property in a system. So, while Wilson's virtues may apply to wide explanations at higher levels of abstraction, such as those in psychology and sociology, but

they do not seem to apply to the computational level. It seems they cannot be used, therefore, to tip the scales back in favour of the wide computationalist.

How about the second challenge? Does Wilson's account fare any better in terms of providing empirically testable consequences? Here again the view struggles. The trouble this time is that while one can grant that medium-independence is what accounts for the possibility of wide computational systems, this does little to tell us about how to go about finding such systems. Simply identifying the isomorphic mappings between a computational model and a physical system does little to help figure out which phenomena are apt for wide analysis, nor how such analysis should be conducted. As it stands, wide computationalism seems to offer little in the way of empirical tractability. It simply points to the general possibility of isomorphic relations obtaining between wide physical and computational systems, but not how such investigations should be conducted, and this seems at least part of what we want from a plausible account of computational cognition. So, again, the view seems in trouble.

In sum, more seems required to respond to the two challenges than what is on offer from existing accounts of wide computation. Appealing to additional explanatory values alone fails to supply the resources to respond to the two challenges. For this reason, I suggest supplementing wide computation with additional resources from mechanistic computation.

## 2.3 The Mechanistic Turn

Recent years have seen a surge of interest in applying mechanistic thinking to computational accounts of implementation and individuation. For an increasing number of authors, physical computing systems are special types of functional mechanisms, ones that perform concrete computations (Milkowski 2013, 2015; Fresco 2014; Piccinini 2007, 2015; Dewhurst 2018a, b).

### 2.3.1 Wide Mechanistic Computation

In keeping with this tread, in Kersten (2017a), I explicitly adapted Piccinini's (2015) account, articulating three conditions for wide mechanistic computation. These conditions, I suggested, provided grounds for thinking about how wide computing systems are implemented.

First, a system's behaviour must satisfy at least one abstract functional description. A system much be describable in terms of computing at least one input-output function. Second, the system must involve a set of integrated components spread out over brain, body, and world. The computational activity of interest must be supported by a distributed mechanism. Third, the components or variables that form the vehicles of the distributed mechanism must be manipulable in virtue of some portion of their structure. The mechanism must process its inputs in virtue of transforming medium-independent vehicles.

These three conditions distill a good deal of the wisdom of mechanistic views of computation. The first condition is the legacy of causal mapping accounts (Chalmers 1994; Chrisley, 1995; Scheutz, 1995). It says that one must be able to find structural isomorphisms between computational and physical states that criss-cross the individual-environment boundary, ones that ideally support counterfactuals. The second condition takes on the lessons of mechanistic explanation (Milkowski, 2013; Piccinini 2007, 2015). It says that computing systems must be realised by a robust set of components, which in virtue of their coordinated activities, give rise to the phenomenon in question. The third condition, which addresses issues of tractability, says that if computational processes are implemented in functionally integrated systems, then such systems need to be responsive to differences in the form or structure of its vehicles (Piccinini & Scarantino, 2011).

To be clear, by 'medium-independent' it is meant that a state or variable possesses certain degrees of freedom and functional organisation within some medium (Garson 2003; Coelho Mollo 2018; Piccinini, 2018). For example, 'digits' are medium-independent insofar as they are distinguishable by processors in virtue of some portion of their structure, such as where

they lie along a string. Computational vehicles, which are medium-independent, must be transformed over the course of processing in virtue of the system's sensitivity to some portion of their structure.

Another way to explicate medium independence is to contrast it with multiple realisability. As Ritchie and Piccinini (2018) describe the difference, one way to think about multiple realisability is as a relation that holds between a property of a whole system and properties and relations of that system's component parts. A higher-level property is said to be multiply realisable if it reflects the causal powers of its lower-level properties and relations. So, for example, replacing a lithium battery with a magnesium-copper lemon cell reliably produces the required voltage to power a LED light because both types of batteries, despite varying in material composition, contribute the same causal power to the circuit.

In contrast, medium independence can be thought of as a more restrictive claim about how the vehicles of computation are implemented. Implementing a computation only requires that, whatever the physical medium, the vehicle possesses the right degrees of freedom and functional organisation. While multiple realisability is defined in terms of specific physical effects (e.g., producing voltages), medium-independence is defined in terms of the relation between variables (e.g., a cell in a Turing machine taking on either a '1' or '0'). So while everything that is medium-independent is multiply realisable, not everything that is multiply realisable is medium-independent (cf. Polger and Shapiro [2016]).

Phrased more generally, the view is that physical systems are apt for wide computational treatment if they involve a world-spanning or distributed mechanism that manipulates medium-independent vehicles in accord with at least one abstract function or rule. When a system implicates a functionally integrated set of components that spans the agent and world, one which supports at least one abstract function and which achieves its function in virtue of manipulating medium-independent vehicles, it can be said to form a distributed computational mechanism, which forms the basis for a wide computational explanation.

As an example, consider the case of cricket phonotaxis. First, in terms of an abstract function, the cricket is computing a function for localisation and identity. The inputs, in the form of the acoustic signal, are mapped to internal outputs, a representation of song location and identity. The only difference between this case and that of, say, a human is that the inputs reside neither on nor in the cricket's auditory receptors but, rather, as part of the external environment.

Second, the cricket's ability to detect and locate male crickets is achieved via spatiotemporal parts spread out across the brain, body, and world. The propagation of acoustic pulses is what actively drives and maintains mate detection. It is the neural processing plus the morphology and acoustic environment that supports phonotaxis. There is a distributed functional mechanism supporting the cricket's ability to detect male crickets mating song.

Finally, as the cricket moves through the acoustic space it samples and adjusts its head to maximize regularities in the incoming sound. What matters is not the sound qua sound, but the sound qua external information-bearing structure. The acoustic array is a medium-independent vehicle apt for manipulation; the only difference between this and normal cases is that the manipulation takes place in the interaction between the agent and the world, rather than within some internal channel.

So, not only does phonotaxis involve a mechanism that spans the brain, body and world, but it also involves a mechanism which computes an abstraction function via the transformation of medium-independent vehicles. It meets all three requirements of the mechanistic account. To be clear, I am not claiming that the cricket example establishes the empirical plausibility of the view exclusively, I take it that this has been done to a large extent

elsewhere.[28] Instead, I am attempting to provide a sense of how the view is supposed to work relative to its implementation conditions.

As a point of comparison, consider extended accounts of cognition, such as Menary (2007) and Wilson (2010). According to these views, what allows for cognitive extension, or integration as Menary (2007) calls it, is the presence of functionally integrated causal systems, ones which implicate component parts both inside and outside the boundary of the individual. Similarly, the mechanistic account says that distributed computing mechanisms also involve casually integrated sets of component parts, ones that conspire to support some wide functional process. A system is said to compute only if its component parts and activities conspire to support a given computational function.[29,30]

The problem is that, as things stand, the mechanistic account lacks an explanation of what makes something part of a distributed functional mechanism. It says little about what mechanisms are or how their component parts should be identified. This problem is particularly pressing given that the view claims that some mechanisms constitutively include parts of the embedding environment. Without an account of what parts are constitutively relevant to a phenomenon, it remains unclear how to distinguish a functional mechanism from its wider embedding environment. We cannot say much about how to either identify wide mechanisms (the viability challenge) or how the mechanisms that feature into wide explanations could prove importantly different from the ones implicated in internalist explanations (the parsimony challenge).

[28] For mechanistic examples of wide computation, see Kersten (2017a, 2017b), Nowakowski (2017), and Smart (2018), while for more general examples see Wilson (1994, 2004), Hutchins (1995), Losonksy (1995), and Kersten & Wilson (2016).
[29] In Piccinini's (2015) view, these are teological functions, but that detail isn't too relevant at present.
[30] The key difference is that while traditional, functionalist-inspired accounts of extended cognition require causally integrated systems to support a cognitive capacity (Wilson, 2010), the mechanistic view also requires that the description satisfy a characterisation in terms of an abstract function (input-output mapping) in virtue of manipulating medium-independent vehicles (Piccinini, 2015).

### 2.3.2 Distributed Mechanisms

Mechanistic explanations are generally said to have two parts: (i) they involve isolating some phenomenon to be explained and (ii) they involve positing a mechanism; mechanisms are always mechanisms *of* something. One has to identify the relevant subcomponents of a mechanism and show how these conspire to produce the phenomenon of interest (Craver 2006).

In terms of the mechanisms themselves, these are usually said to have four characteristics: (i) they are phenomenal, in the sense that they perform tasks; (ii) they have at least two components relevant to the explanadum; (iii) the components are causally interrelated; and (iv) the spatial organisation of the mechanism's components, e.g., their locations, shapes, orientations, etc., and the temporal organisation, e.g., their order, rates, and durations of their activities, are relevant to generating the phenomenon (Craver 2006, pp.469-70).

One way mechanists have attempted to explicate componential relations, and thus identify mechanisms, is via the notion of 'constitutive relevance' (Craver 2007a, 2007b, 2009; Couch 2011; Kaplan 2011, 2012; Baumgartner & Casini 2017; van Eck 2018). Craver (2007b, p.159) explicates the notion in terms of what he calls 'mutual manipulability', which he says has two criteria:

(CR1) when $\phi$ is set to value $\phi_1$ in an ideal intervention, then $\psi$ takes on the value $f(\phi_1)$;

(CR2) if $\psi$ is set to the value $\psi_1$ in an ideal intervention, then $\phi$ takes on the value $f(\psi_1)$.

(CR1) and (CR2) suggest that a component's activity is constitutively relevant to another component's activity if the two are related part to whole, and they are mutually manipulable. So, X's $\psi$-ing is constitutively relevant to S's $\phi$-ing if one is able to manipulate S's $\phi$-ing by intervening on X's $\psi$-ing (by stimulating or inhibiting) and one is able to manipulate X's $\psi$-ing by intervening on S's $\phi$-ing. For example, to say that synaptic depolarisation is constitutively relevant to action potentials, one must be able to show how changes in the

distribution of the neurotransmitters, such as Sodium $N^+$ and Potassium $K^+$, affect the occurrence of action potentials, and one must be able to show that intervening on action potentials, such as through ET, in turn affects the distribution of neurotransmitters.

The benefit of appealing to mutual manipulability is that it provides a method for identifying the component parts and activities of distributed mechanisms. To say that X's ψ-ing is constituted not only by component S's $\phi_1$-ing internal to the individual but also component Y's $\phi_2$-ing external to the individual there must exist some ideal intervention $I_1$, such that intervening on X's ψ-ing via $I_1$ is associated with changes in S's $\phi_1$-ing and changes in Y's $\phi_2$-ing.

For example, to return to the phonotaxis case, if the acoustic properties of the acoustic array (S's $\phi_1$-ing) and the tracheal tube (Y's $\phi_2$-ing) are constitutively relevant to phonotaxis (X's ψ-ing), then under some ideal intervention ($I_1$), intervening on either the acoustic array (S's $\phi_1$-ing) or the tracheal tube (Y's $\phi_2$-ing) should affect the occurrence of phonotaxis, and vice-versa. Given the systematic relation between the acoustic array and the cricket's morphology, any intervention on the acoustic array or the cricket's morphology will affect the cricket's ability to detect and localise the male mating signal, while any interference on phonotaxis will affect the activity of either the acoustic array or the tracheal tube. I will say more about this in Section 2.4.

What we have, then, is a way to flesh out the notion of distributed mechanisms. A functional mechanism is distributed only if it includes a set of components spread out over individual and world which, in virtue of their spatio-temporal organisation, are jointly responsible for a given phenomenon. Not only are wide computational mechanisms ones that compute an abstract function in virtue of manipulating medium-independent vehicles, but they are also those which involve constitutive relations between internal and external components, which are identifiable using the mutual manipulability criteria.

A few caveats. First, while it might be tempting to think of the relationship between S's $\phi_1$-ing and Y's $\phi_2$-ing as symmetric – that is, as mutually affecting – this is not, strictly speaking, required. All that is required for S's $\phi_1$-ing and Y's $\phi_2$-ing to be constitutive of X's $\psi$-ing, is that they have both have a mutual effect on X's $\psi$-ing given some ideal intervention $I_1$. For example, the activities of V1 and V2 in the visual cortex are both constitutive of visual processing, in that inhibiting or exciting either affects visual processing. But, this does not mean that in course of visual processing V1 and V2 are causally linked in both directions. Rather, the casual flow is simply one that goes from V1 to V2. V1 might excite V2 but the reverse isn't necessarily the case. The causal relations between S's $\phi_1$-ing and Y's $\phi_2$-ing can be purely asymmetric and yet still constitutive.

Second, it is worth emphasising that this is not an account of computational identity or individuation. Identifying which properties are required to taxonomise computations is a different task to that of outlining which properties are required to implement computations. There is a substantive and interesting question about the role of the environment in individuating computations, but this is importantly different from the question of how computing systems are implemented (Shagrir, 2019). That said, there are some close connections.

Dewhurst (2018a), for example, maintains that computational individuation can follow in virtue of the physical properties of computing systems. The key idea is that when two processors, such as two logic gates, with mirror input/outputs, are placed in the same physical system they will exhibit differences at the level of physical description. These differences, Dewhurst thinks, are enough to distinguish the two processors, even if an abstract interpretation is ambiguous, such as whether the processors implement an OR or AND gate. The physical mechanism alone is sufficient for computational individuation (c.f. Harbecke & Shagrir 2019).

Dewhurst's central insight sits nicely with the view on offer here. If computational systems are implemented in distributed mechanisms, then individuation of the computations being performed can follow from attention to the environmental elements that are constitutive of the distributed mechanisms. I raise Dewhurst's account not as the final word on individuation, but simply to emphasise that the wide mechanistic account has options when it comes to fleshing out the individuation side of the story.

Finally, mutual manipulability is not the only way to cash out constitutive relevance. Baumgartner & Casini (2017), for instance, offer a "no-decoupling account" (NDC), while Kastner (2017) offers a "temporal EIO-part account" (EIO). The NDC and EIO share important similarities to Craver's account, in that they share a common emphasis on the notion of 'difference-making', but they are nonetheless importantly different, as they move away from Craver's focus on causal relations. So while I explicate distributed mechanisms in terms of mutual manipulability here, I acknowledge that there are other useful ways to think about constitutive relevance. I am not saying that the wide mechanistic account solely rests on mutual manipulability, but the view does offer a constructive proposal for how to further flesh out the notion of distributed functional mechanism. I think that, in principle, the wide mechanistic account could be made to fit with other accounts of constitutive relevance, but I stick with Craver's here because it is well-developed and offers the clearest connection to experimental practises.[31]

### 2.4 Two Challenges Revisited

Time to revisit the two challenges. First, recall that the parsimony challenge argued that we have reason to prefer internalist explanations, on balance, because they are ontologically simpler than their wide counterparts.

[31] See Kastner (2017) for further discussion.

The trick to answering this challenge is to undercut the 'ceteris paribus' clause, the idea that wide and narrow explanations provide equally good descriptions of the relevant events and interactions *all things being equal*. We can do this by focusing on the relationship between mechanistic explanation and computational explanation.

As mentioned, there are two key features to mechanistic explanations (Piccinini & Craver 2011). First, one has to provide an abstract specification of the computational function under investigation. One must provide a functional analysis of the task being performed – in the phonotaxis case, for example, the task was localisation and identification. This is what is typically called a mechanistic *sketch* or *schemata*. Second, one has to provide a blueprint of the mechanism underlying the task or capacity. One must provide an explanation of the structural details of the mechanism in question. Once the task or competence has been fixed, analysis has to shift to looking for the supporting computational mechanism, i.e. the set of components and activity realising the task or competence.

A mechanistic explanation is only *complete* when it outlines not only the task being performed but also what actually performs the task. As Milkowski (2013) points out: "The norms of mechanistic explanation require that computational mechanisms be completely described. This means that they cannot contain black boxes or questions" (p.123). Not only must a computational explanation accurately describe the task or competence being performed, but it must also provide insight into the underlying mechanism's component parts and activities.

This is particularly relevant in the present context because it means that a wide computational explanation can be explanatorily richer when it not only functionally describes the computational task being faced, but also when it accurately describes the supporting mechanism. This is what Thagard (2007) calls having greater 'explanatory depth'.[32]

[32] An explanation is explanatorily deeper, according to Thagard (2007), when it provides a more specific or illuminating account of the underlying mechanism. Explanatory depth is distinct from causal depth, in that it refers to the epistemic insight offered by an explanation rather than whether it identifies the most causally robust set of properties or not.

Accurately describing the computational task is only one part of providing a good computational explanation.

What the search for the underlying component parts and activities does is take the 'ceteris paribus' assumption of the parsimony challenge off the table. It makes it so that the computational explanation that best elaborates the underlying mechanism is explanatorily superior. While narrow and wide accounts might describe the same processes and events at the functional level of description, the same is not true for the structural level of description. The abstract, functional characterisation is only half of the computational story.

To illustrate, return again to the phonotaxis case. Whether or not cricket phonotaxis is best explained via a wide or internalist account depends not only on how the higher level computational task (phonotaxis) is characterised, but also on which underlying mechanism is identified (distributed or internal). If elements outside the cricket's brain, such as the tracheal tube or acoustic array, prove constitutively relevant to the supporting mechanism, then a wide computational explanation is required. But if, on the other hand, only elements wholly internal to the cricket prove constitutively relevant, such as those in the auditory and neural systems, then an internalist explanation is preferable. I will say more about how to determine the mechanism shortly, but the important point is that there will be an empirically testable difference.

Notice how this response differs from the one proposed by Wilson's account. The problem for Wilson's account was that it had to appeal to additional explanatory virtues, such as theoretical appropriateness or causal depth, to counter simplicity considerations. It accepted the *ceteris paribus* assumption, but maintained that there were independent reasons for thinking that wide computational explanations were preferable. The mechanistic account, on the other hand, directly challenges the *ceteris paribus* assumption. While the functional aspects of computational explanation may be equivalent, the structural aspects are not. Simplicity considerations do not get a chance to take hold, because two computational explanations will implicate different underlying mechanisms.

Of course, one is free to disagree about whether, in a given case, a distributed mechanism is in fact implicated. However, in taking the *ceteris paribus* considerations off the table, the bite of the worry is largely removed. Wide computational explanations are now potentially superior so long as they implicate the right mechanism. Unlike previous formulations, the mechanistic account ties computational explanations explicitly to structural descriptions, which means that higher-level processes and events are by themselves insufficient to determine when two explanations are equal.

One might further object that there are still additional reasons to prefer internalist explanations; that other explanatory virtues such as 'explanatory breadth' – the idea that theories which cohere with a wider set of theories are preferable – might tip the scale back in favour of the internalist. And while this might be an open question, it nonetheless importantly shifts the debate. This is because the debate would no longer be about the value of simplicity in deciding between narrow and wide explanations, but about how to order the explanatory virtues in general. Of course, this is an important discussion. But once we have admitted that there is a matter of fact as to which computational mechanisms are implicated in a given computational explanation, then simplicity alone cannot tip the balance in favour of the internalist.

Recall next that the testability challenge worried that the mechanistic account failed to furnish testable, empirical consequences.

A response to this challenge can be mounted using the connection offered between the mutual manipulability criteria and 'interlevel' experiments via constitutive relevance. As Craver (2002, 2007) describes them, interlevel experiments address the different types of interventions we can perform on mechanisms.[33] They vary along two dimensions. First, there are bottom-up/top-down experiments, which either disable or stimulate components or higher-level phenomenon in order to observe the effects on the corresponding lower-level

[33] See also Darden (2002).

phenomenon or components. Second, there are inhibitory/excitatory experiments, which intervene on a mechanism's component parts or higher-level phenomenon in order to observe the effects on the corresponding lower-level phenomenon or components.

Three specific types of inter-level experiments can be constructed using these two dimensions. The first, interference experiments, which are bottom-up inhibitory experiments, intervene to diminish or disable a mechanism's component in order to detect effects on some higher-level phenomenon. For example, lesion experiments remove a portion of the brain to detect effects on task performance. The second, activation experiments, which are top-down excitatory experiments, intervene to active or trigger a target phenomenon in order to detect the properties or activities of one or more of the mechanism's component parts. For example, in PET or fMRI studies, cognitive systems are engaged on some task to investigate the different brain areas associated with the system's activation. Finally, stimulation experiments, which are bottom-up excitatory experiments, intervene on one component of a mechanism to detect effects on some higher-level phenomenon. For example, applying low-grade electrical stimuli to different motor cortical areas in dogs produces movements in specific muscles, e.g., the legs, tail, and facial muscles.

The advantage of talking about constitutive relevance is that it opens up this array of possible interlevel experiments to test the mechanistic account of wide computation. There is now a corollary to each kind of interlevel experiment in wide investigations.

To see this, return one last time to the cricket phonotaxis case. First, consider that to employ an interference experiment, one might disrupt or modulate the carrier frequency in the acoustic array using an acoustic pulse. This would reveal the range of frequencies in the carrier wave that are relevant to triggering phonotaxis (e.g., 4-5Hz). Second, to use a stimulation experiment, one might intervene on the cricket's morphology, either lengthening or shortening the tracheal tube, and observe what effects this has on the cricket's ability to direct and localise the mating signal. This would establish which morphological features are relevant to sustaining phonotaxis. Finally, in the case of activation experiments, one might engage the

cricket on a phonotaxis task, or set of tasks, and then use a wire-tapping method to reveal which parts of its brain were relevant to the control mechanism (Dawson, 1998).

Of course, this discussion oversimplifies matters. For example, under normal circumstances, any given interlevel experiment tests several interventions simultaneously, intervening on several pathways to control for as many variables as possible (see Baumgartner & Casini, 2017; Kastner, 2017). Nonetheless, in spite of individual methods varying, the general apparatus offered by the interlevel experiments is now available to wide investigations. It is no longer the case that wide computation is restricted to merely re-describing existing phenomenon at a functional level.[34] Instead, once the initial functional description has been provided, substantive structural investigations can now be undertaken to uncover the supporting computational mechanism. Further integration with mechanistic thinking provides a nice hook into experimental investigation for the mechanistic account. To be clear, the interlevel experiments just noted have not actually been carried out. What I am offering here are descriptions of experiments that *could* be used to empirically test wide computationalism. The worry, recall, was that wide computationalism was not empirically testable. It now is.

So, to review, the mechanistic account of wide computation can address each of the two challenges in virtue of further embracing mechanistic thinking. First, in requiring that computational explanations describe the underlying computational mechanism, it can undercut the underdetermination concerns motivating the parsimony challenge; second, by adopting a notion of constitutive relevance such as Craver's, it can connect wide investigations to interlevel experiments, which undermines the viability challenge in virtue of providing a clear hook into experimental investigations.

[34] See Wilson (1994), Hutchins (1995), and Losonksy (1995) and Kersten and Wilson (2016) for instances.

## 2.5 Objections

Some might be sceptical of my appeal to Craver's version of constitutive relevance. Some worry, for instance, that on Craver's account it is possible to have experimentally equivalent models that do not involve constitutive relations (Baumgartner & Wilutzky, 2016, 2017).

For instance, if Y's $\phi_1$-ing and S's $\phi_2$-ing change as a result of an ideal intervention $I_1$ on X's $\psi$-ing, then the experimental story stays the same regardless of whether $I_1$ effects the behaviour X's $\psi$-ing via Y's $\phi_1$-ing and S's $\phi_2$-ing. It's still up for grabs whether S's $\phi_1$-ing and $\phi_2$-ing are constitutively or merely causally relevant to X's $\psi$-ing. If one wants to maintain that constitutive relevance is not a causal relation, then it seems difficult to also appeal to the type of interventionism assumed by mutual manipulability. Note that such a worry would apply equally to internal and wide views of computation.

One way to side-step this concern is to supplement mutual manipulability with additional constraints. Recall the three implementation conditions: (1) said that a system's behaviour must satisfy at least one abstract functional description, (2) claimed that a system must involve a distributed mechanism, while (3) said that the components or variables that form the vehicles of the distributed mechanism must be manipulated in virtue of some portion of their structure.

Notice that if (3) is taken seriously, then we have a further constraint on which features are constitutively relevant to distributed computing mechanisms. It is only those vehicles that are medium-independent that can be included as part of a distributed computational mechanism. Since the issue is one of restricting the set of elements that might be properly included as parts of the computational mechanism, so long as external structures, such as the acoustic array, have the requisite degrees of freedom they can qualify as the proper vehicles of computational mechanisms.

In other words, if only (2) holds but not (3), then a component may be part of a distributed mechanism, but not necessarily one that is implicated in a distributed computing mechanism;

there may be cases of extended digestion, for example, that involve such distributed functional mechanisms, but which do not involve computing processes (Wilson, 2010). If only (3) holds but not (2), then the component may form the input to an internally bound mechanism, but it may not be constitutively relevant to an external mechanism. But if both (2) and (3) hold, alongside (1), then a given external structure may be incorporated into the distributed computing mechanism. The vehicle manipulation requirement helps to restrict the set of elements that can be said to be relevant to identifying a given computing mechanism.

To be clear, the claim is not that the mechanistic account affords a novel response to Baumgartner and Wilutzky's worry, there is no suggestion of using constitutive relevance to identify the boundaries of cognition or anything more metaphysically ambitious (c.f. Kaplan, 2012). Rather, the claim is that constitutive relevance simply helps spell out how to carve up the boundaries of computational mechanisms specifically, and that this, in conjunction with other constraints, is enough to say when a given physical system is computing. On this view, the boundary of the organism might simply be a more conventional boundary, such as the central nervous system.[35]

A second objection is that one might worry that there is a confusion going on in terms of how mechanisms are being analysed. Many mechanists hold, for instance, that mechanisms are analysable at three distinct levels: the contextual level, which outlines the capacity to be explained (the explanandum); the isolated level, which contains parts of the whole, but not the system itself; and the constitutive level, which spells out the composition of the parts (Craver, 2007; Milkowski, 2013).[36] At the contextual level, a mechanism is embedded in some environment, while at the isolated and constitutive levels, the mechanism is described only in terms of the entities that are part of the mechanism. The worry, then, might be that all this talk

[35] Romero (2015), Kastner (2017), and Krickel (2018) offer other interesting replies to Baumgartner and Wilutzky's worry.
[36] These levels, it should be added, are levels of the mechanistic explanation, not levels of nature.

of wide or distributed mechanisms conflates analysis at the contextual level with analysis at the isolated and constitutive levels.

One reason to think there is no such conflation is that mechanistic analysis, to a large extent, is context sensitive. That is, which parts feature into either contextual or isolated levels depends, in part, on how the explanadum phenomenon is initially characterised.

Take vision, for example. If characterised in terms of facilitating object recognition, a whole suite of properties and activities are salient to investigation. These include fine-grained properties, such as those in the primary visual cortex involved in edge detection, and coarse-grained properties, such as those involved in categorisation. However, if characterised in terms of facilitating action-guidance, than an entirely new set of properties and activities are implicated, such as those involved in maintaining body images or reaching behaviour. The functional decomposability of vision depends not only on structural organisation of the phenomenon but also on how the activity is initially characterised.

Once it is appreciated that functional characterisations fix part-whole relations within a given mechanistic analysis it follows that decomposability into different levels is not only tied to organisational structure, but also to how the mechanism is initially functionally contextualised. Part of what matters for mechanistic analysis is how the contextual function of the phenomenon is initially described. It is only after this description that questions about part-whole relations at the isolated and constitutive levels begin to take shape – we can analyse any constitutive level of one mechanism as a contextual level of another mechanism. As Milkowski (2013) frames the point: "If contextual effects prove important, they will be analysed as parts of larger mechanisms and thereby included in the isolated level; if not, they won't. To see what an explanation of a phenomenon should include, one has to identify the boundaries of the mechanism and its causally relevant parts" (p.124). One mechanist's isolated level is another's contextual level.

But does not this appeal to the context sensitivity of functional descriptions re-introduce trivialisation concerns? Does it not show that mechanistic analysis is problematically observer

relative?[37] The simple answer is no. While it is true to say that there are several viable ways to carve up the same system, this does not mean that any function can be ascribed to a given physical system (Craver 2013). This is particularly true in the case of computational functions.

Consider the digestive system, for example. While the digestive system is appropriately described as a functional mechanism, it fails to meet an abstract functional description in terms of computing medium-independent variables. It does not have the right kind of structure to be described as processing medium-independent vehicles. As Dewhurst (2018b) expresses the point: "For a mechanism to perform the function of computing is to possess the correct kind of physical structure to be interpreted as performing this form an explanatory perspective". So while mechanistic analysis does make function attribution partially dependent on the observer, this does not mean that the underlying physical structure of the system is still not enough to circumscribe the range of functions properly attributed to the system. The medium-independent processing requirement still excludes a sizable set of computational functions.[38] This is enough to block the majority of spurious mapping cases.

## 2.6 Conclusion

My aim in this paper has been to update and extend the account of wide computation by integrating it with mechanistic views of computation. I have sought to do this by addressing two outstanding challenges (*parsimony* and *testability*). I have shown not only how the wide computationalist should defend her ontological commitments but I have also highlighted important connections between interlevel experiments and wide investigations that make such a proposal empirically testable. Wide mechanistic computation may be both explanatorily desirable and experimentally tractable.

[37] For discussion of triviality concerns more generally see Sprevak (2018a).
[38] Isaac's (2013) 'homomophoric' account or Millhouse's (2018) 'simplicity' criterion offer other ways of delimiting the physical structures that might prove relevant to implementation.

# Introduction to Chapter 3

## Key ideas

The goal of Chapter 3 is to demonstrate the theoretical 'legs' of wide mechanistic computation. In particular, I attempt to illustrate the general value of the view by using it to resolve two tensions in 4E cognition. The first, raised by Clark (2008a), says that the body-centric claims of embodied cognition run counter to the distributed tendencies of extended cognition. The second, raised by Clark & Kiverstein (2009), says that the body/environment distinction emphasised by enactivism militates against the world-spanning claims of extended cognition. The tensions point to some rather deep incompatibilities between the various Es.

My argument is that wide mechanistic computation can reconcile these two tensions in virtue of placing a simultaneous emphasis on location neutrality, concrete implementation, and functionally-closed autonomous systems. In particular, it addresses the first tension by emphasising functional mechanisms and medium independence, thereby satisfying the embodied theorist's demand for a privileged role for the body and the extended functionalist's demand for abstract, functional analysis. It addresses the second tension by highlighting a shared set of conceptual resources between wide computationalism and enactivism; in particular, the notion of autonomy.

## The argument

The trouble is that this argument rests on two assumptions that need further defence. The first is that wide mechanistic computation offers a viable approach to computation. While I

gesture at its plausibility in the chapter, I do not explicitly argue for the view. If the argument is to be made convincing, then the general case for the view needs to be explicit.

To that end, I want to consolidate some of the evidence in favour of the view that has been presented so far. First, recall the previous examples of bat echolocation and spatial navigation from Chapter 1. These, I suggested, pointed to a more general class of phenomena plausibly interpreted as instances of wide mechanistic computation: *active sensory systems*. The rich informational flow and tight causal integration of agent and environment in active sensory systems offered a plausible starting point for investigating wide systems. There seems to be broad empirical support for the view.

Second, by most lights, coherence with existing theories or background knowledge is an important explanatory virtue (van Fraassen 1980; Psillos 1999; Glass 2007). The more a theory coheres with background beliefs the better it is. As I previously argued, wide computationalism is not compatible only with causal mapping and mechanistic accounts of implementation, but also some forms of non-semantic individuation. It coheres with implementational and individuation theories of computation. Such broad coherence again speaks in favour of the view. To be clear, the claim is not that wide computationalism coheres *more* than its rivals (e.g., individualism) with wider views about computation. Rather, the claim is that, despite appearances to the contrary, there is broad coherence between such views where one might have otherwise thought none existed; the individualist, recall, often frames their position as uniquely following from computational considerations (see, e.g., Fodor, 1981, 1983). Wide computationalism is *at least as good as* rival/alternative internalist or individualist views.

Finally, as we saw in Chapter 2, wide mechanistic computation is both explanatorily and methodologically attractive. It can be explanatorily superior to rival internalist accounts and it can be empirically tested. These features again point to two important explanatory virtues: (i) explanatory depth and (ii) empirical fruitfulness. The view is not only able to provide insight into the causal mechanisms underlying computational phenomena but it can also generate

empirical discoveries via experimental interventions. Taken together, such considerations should, I think, make a plausible case for the viability of the wide mechanistic view of computation.

The second assumption is that there is, in fact, some value to be had in reconciling the two tensions facing 4E cognition. As I mention in the chapter, a sceptic might argue that there is little reason to expect that the 4Es should form a unified research programme in the first place, as the various Es merely constitute a loose set of nominally related approaches. While I point to some general considerations that I think chip away at this scepticism in the chapter, I want to say a bit more about the assumption here.

The most straightforward reason for trying to resolve the two tensions is that it may produce a simpler, more unified framework for 4E cognition. Consider an analogous point sometimes offered about the predictive processing framework, the view that the brain is best treated as a multi-layered hierarchical prediction machine. According to Clark (2016a), one reason for pursuing the predictive processing framework is that it helps to bring together a number of otherwise disparate theories and models about cognition, everything from bottom-up re-constructivist models of vision to top-down Bayesian theories of concepts acquisition. By the same token, resolving the two tensions may help to unify the 4Es. It could potentially serve to articulate a common theoretical core, one which could pull together insights about embodiment, environmental scaffolding, dynamic causal coupling, and functional analysis. In short, we have good epistemic grounds for trying to pull together the various strands of 4Es, as it provides a framework that allows us to see how various overlapping/related conceptual resources and insights hang together.

The big idea of the chapter is that wide mechanistic computation can serve just this unifying role. It can articulate a common theoretical core for 4E cognition. In the process of resolving the two tensions, the view offers a 'philosopher's stone' of sorts, one which can translate some of the core insights and concepts of one E into that of another, e.g., autonomous systems in autopoetic enactivism to extended autonomous systems. Rather than being orthogonal to 4E

cognition, as is sometimes wrongly suggested, computational thinking turns out to play a central role in its unification.

## Broader context

The chapter links up with a number of existing attempts to assuage some more general tensions between more 'classical' views of cognition, such as computationalism, and 'radical' views of cognition, such as enactivism or embodiment. For example, in a series of recent papers Dewhurst and Villalobos (2016, 2017, 2018) have attempted to show that there is nothing particular to the enactivist or autopoietic approaches that bias them against computationalism. They write, for instance: "AT [autopoietic theory], as we can see, has the potential to accommodate some form of computational explanation in its explanatory framework, and in that way, to retain the notion of computation as a useful theoretical element in the study of cognitive systems" (p.125). While their target is slightly more specific (autopoietic theory), Dewhurst and Villalobos' general strategy is the same: to show that the explanatory and conceptual apparatus used by a radical view of cognition is amenable, if not compatible, with computational thinking. In the same way, and drawing on several of their insights, I argue that enactive, embodied and extended cognition can be fruitfully linked using a wide mechanistic account of computation. The chapter adds to a growing chorus of voices that see computational thinking as having a substantive role to play in discussions of 'radical' views of cognition, such as the 4Es.

# Chapter 3 – Resolving Two Tensions in 4E Cognition

## 3.1 Introduction

Enactive, embodied, embedded and extended cognition, or simply '4E' cognition, has often been thought to form a collective challenge to classical cognitive science (Menary, 2010; Ward & Stapleton, 2012). Central to many of these views is the idea that cognition is often integrated with and heavily dependent on body and world (Varela, Thompson, & Rosch 1991; Clark & Chalmers, 1998; Noe, 2004; Wilson & Foglia, 2016).

Recently, some authors have worried about the potential unity of 4E cognition. Two tensions, in particular, have been raised. The first, raised by Clark (2008a), says that the body-centric claims of embodied cognition run counter to the distributed tendencies of extended cognition. The second, raised by Clark & Kiverstein (2009), says that the body/environment distinction emphasised by enactivism militates against the world-spanning claims of extended cognition. Together these tensions seem to present a challenge to 4E's status as a distinct framework for cognitive science. They suggest a troubling a lack of unity for a framework under which an increasing number of empirical and philosophical projects are conducted.

The aim of this paper is to provide a resolution to these two tensions. Building on existing proposals, I argue that a mechanistic form of 'wide computationalism' can be used to help reconcile the various strands generating the two tensions.[39] The aim is to show that a renewed focus on computation can help strengthen the 4E framework.

[39] The fourth 'E', embedded, is the least contentious and so is not discussed here.

I begin by sketching the three strands which generate the two tension, alongside examining Clark's (2008a) proposed solution (Section 3.2). Next, I introduce wide computationalism, unpacking its conceptual and empirical support (Section 3.3). I argue that a mechanistic version of wide computationalism is able to reconcile the two tensions in virtue of placing a simultaneous emphasis on abstract functional analysis, concrete physical systems, and autonomous systems (Section 3.4). Finally, I conclude by addressing some outstanding issues (Section 3.5).

## 3.2 Three Strands, Two Tensions, One Solution

Three strands generate the two tensions. The first, body-centrism, says that the body has a non-trivial role in determining mental states and functioning, that the details of a creature's embodiment have a profound effect on the functioning of the mind (O'Reagan & Noë, 2001; Noë, 2004, 2009; Gallagher, 2005). For body-centrists, without mention of the unique and ineliminable contribution of bodily structures and activities, cognitive explanations are crucially lacking. The view is also sometimes called the "constitutive-contribution claim" or "special contribution story" (Clark, 2008a; Shapiro, 2019).

The second strand, extended functionalism, says that cognitive systems are functional wholes distributed across diverse sets of components and processes; cognitive activities involve a complex balancing act between brain, body and world (Harman, 1998; Wilson, 2004; Clark, 2008b). The spirit of extended functionalism is captured in Clark and Chalmers' (1998) 'parity principle'. The parity principle says that we should not exclude *a priori* external structures that might otherwise form part of our cognitive process or system should they occupy the right functional role (Clark, 2010). We should be 'unbiased' in our cognitive investigations. The extended functionalist is keen to highlight the location neutrality of cognitive analysis.

The third strand is what has sometimes called 'autopoietic enactivism' (Villalobos & Ward, 2015; Villalobos & Dewhurst, 2017). Autopoietic enactivism, a particular interpretation of

autopoietic theory, says that cognitive systems are autonomous, self-determining systems, ones created by the reciprocal interaction of internal and external components (Thompson, 2005, 2007; Pi Paolo, 2009; Froese and Di Paolo, 2011; Di Paolo & Thompson, 2014). One of the key functions of autopoietic systems is to bring forth meaning or value, sometimes called 'sense-making'. [40] Sense-making is required for maintaining a system's boundary in the face of perturbations from the environment. As Kiverstein and Clark (2009) describe the view: "An autonomous [autopoetic] system that seeks out only those interactions that contribute to its continuation, and avoids those interaction that threaten its survival will, it is claimed, exist in a world in which things have a meaning or value" (p.2). One of principal concepts of autopoetic enactivism is what is known as 'autonomy' (Dewhurst, 2018). As we will see, the concept of autonomy has many sides, but the basic idea is that autonomous systems are self-determining and operationally closed systems – for example, a living cell is self-determining and operationally closed in the sense that it regulates its interactions with the environment in virtue of its intrinsic dynamics.

To clarify, by 'autopoetic theory' I mean any attempt to ground cognition in the biodynamics of cognitive systems. This notion should be kept separate from autopoetic enactivism, for, as we will see later, there are importance differences between certain historical strands of autopoetic theory and modern iterations. It is also worth mentioning that enactivist orthodoxy has changed somewhat, as 'sense-making' is now identified with 'adaptive' rather than merely 'autopoietic' organisation (Di Paolo, 2016). However, because this development makes little difference to the structure of the argument to follow, I ignore it here.

The three strands sit roughly as follows within 4E cognition. First, embodied cognition endorses body-centrism in virtue of emphasising what it takes to be the unique contributions of bodily structures and activities, such as sensorimotor contingencies or knowledge. Second, extended cognition endorses extended functionalism in virtue of assigning a non-trivial

[40] For a classic discussion of autopoietic theory, see Maturana (1981).

functional role to environmental elements in sustaining cognitive activities, such as representational systems in problem solving. Third, enactivism adopts a version of autopoietic theory in virtue of demarcating an organism/environment boundary via autonomous, sense-making systems.

Of course, not every version of enactivism is committed to autopoietic theory, and not every version of embodied cognition is committed to body-centrism. Some versions of embodied cognition, for example, align with extended functionalist views of the body (Clark & Chalmers, 1998), while some versions of enactivism align with body-centric views of cognition (Varela, Thompson, & Rosch, 1991). There is, at least in principle, a degree of compatibility between the various views. The issue is that the whole looks weaker than the sum of its parts. While each of the three strands looks plausible on its own right, when taken collectively, they look increasingly incompatible. The challenge facing 4E cognition seems to be how to account for the truth of all three strands simultaneously; or, failing this, figuring out which one to drop.

Consider, then, how the two tensions are generated. First, if the body has a non-trivial role in determining mental states, then this seems to imply that cognition cannot also be location neutral. The body cannot have a unique role in cognition if it is simply one physical resource among many. Here is Clark (2008a) diagnosing the situation:

> There is a potential tension, it seems to me, between the kinds of account that typically stress features of human embodiment and the kinds of account that typically stress environmental embedding and intervention…[Embodied cognition] depicts bodily form and sensorimotor patterning as elements that might make a special contribution to human thought and reason. But [extended cognition] seem[s] to depict bodily action and environmental structuring as merely additional elements in a wider computational, dynamical, and representational nexus. (p.49).

Body-centrism's push for a privileged role for the body in cognition seems to run up against the location neutrality of extended analysis, particularly as embodied in something like the parity principle.[41]

[41] The tension has also been described as an incompatibility between a 'special contribution story' and 'larger mechanism story', but this is basically the same distinction (Shapiro, 2019).

Second, if extended functionalism is correct, and cognitive systems stretch out into the world, then it seems that living systems, being autopoietic, autonomous systems, cannot also be co-extensive with cognitive systems. As Clark and Kiverstein (2009) make the point:

> If cognitive extension is possible then an extended cognitive system must also be an autopoietic system. However the boundaries of an autopoietic system are the boundaries of the organism. The boundaries of an extended cognitive system are not however the boundaries of the organism. Thus the autopoietic system cannot be an extended cognitive system. (pp.2-3).

Extended cognition seems to require the possibility of recruiting resources outside the individual, but autopoietic enactivism denies this possibility. It undercuts the identification of cognitive systems with extended systems in virtue of maintaining a sharp distinction between the physical boundaries of the organism and the environment.[42] As Wheeler (2010) notes: "If the living system is identical with the cognitive system, then the boundary of the living system is the boundary of the cognitive system. And it's this that (finally) generates the inconsistency with EM [extended cognition]" (p.10).[43]

Taken together the two tensions seem to undermine the collective thrust of 4E cognition. They seem to disrupt the framework's otherwise harmonious picture by pointing to some rather deep incompatibilities. As Kiverstein and Clark phrase the worry: "[D]oes that large and sheltering slogan (mind as 'embodied, embedded, enacted') hide points of (perhaps even irreconcilable) disagreement about the nature of the mind and the shape of a mature cognitive science?" (2009, p.2).

Of course, one might dispute that such tensions really need resolving. One might think there is little reason to expect that the 4Es should form a unified research programme in the

[42] This also assumes that extended cognition and autopoietic theory are global theories about cognition.

[43] It is worth mentioning that is no tension between enactivism and embedded cognition, because while the autonomous/adaptive dynamics are essentially world-involving (embedded cognition) they do not go so far as to claim that such dynamics are constitutive of cognitive systems (extended cognition). Unlike extended cognition, cognitive systems for the enactivist are living systems, which means they are bound to the body. The enactivist's commitment to the strong continuity of life and mind only ensures a compatibility with embedded views, not necessarily extended views.

first place; perhaps the various Es simply constitute a loose set of alphabetically related approaches, for example (see Chemero, 2009; Myin & Hutto, 2013, 2017).

But consider the deep similarities one would have to explain if this were true. Consider, for example, that one would have to explain why embodied and extended cognition both appeal to the tight causal coupling of action and perception cycles (Clark, 1997; Wilson, 2004); why the enactivist and ecological psychologist invoke organism-environment dynamics to explain perceptual capacities (Gibson, 1979; Varela, Thompson, Rosch, 1991); why the sensorimotor enactivist and phenomenologist reflect on agent mobility and temporal spread when explaining visual experience (Merleau-Ponty, 1945; Noe, 2004); or why the embedded and extended theorists both deploy the concepts such as scaffolding when explaining cognition (Kirsh, 1995; Clark, 2003). Such overlaps seem quite the coincidence should there be no common ground; an odd instance of philosophical parallel thinking. The more likely scenario seems to be that the shared ground is owed to the various Es tracking the same phenomena using a suite of similar concepts.

But such an admission does not preclude others from trying to develop the various Es separately, nor does it mean that a resolution to the two tensions necessarily has to come at the cost of one E being annexed or reduced to another; a unification might come from a synthesis, rather than a replacement (Shapiro, 2011). The two tensions simply offer an invitation to reflect on what might be shared by the various views.

Aware of the looming fractures, Clark (2008a) offers the following proposal:

> The proper resolution of this tension…is to display the body as (for all cognitive purposes) nothing but the item, or items, that play a certain complex functional role in an information-processing economy. Within such an economy, mental sameness is determined by the overall balance achieved using neural, bodily and environmental resources. The body plays a special role in determining and stabilizing this balance and as such it is a key player on the cognitive stage. (p.57).

Clark's solution is to view the body as playing an 'enabling' role within cognition. The strategy is to conceptualise the body in terms of its ability to enable different kinds of information processing. Thus the significance of the body emerges as a function of its role within

'intelligent' organisation. The Es are unified by a shared emphasis on complex functional systems. Some systems are individual bound (as per enactive and embodied cognition), while others are spread out across brain, body and world (as per extended cognition). Note that Clark's proposal only addresses the first tension, it does not address the second. I return to the second tension in section 3.4.

More important, notice that Clark's solution essentially sides with the extended functionalist; his use of the phrase 'nothing but the item' is telling. Though there is mention of the 'special' role of the body, it does not denote a privileged status in virtue of understanding the body's fine-grained contributions to cognition. Rather, it simply refers to the unique mediating role the body can take on between brain and world. What is important for Clark is that the body stabilizes the relationship between brain and world, but this role is still very much functional in nature.

One issue with Clark's proposal, despite its advantages, is that it fails to specify the relationship between physical systems and abstract analysis in sufficient detail. Sprevak's (2009) discussion of extended functionalism helps bring this point out. Sprevak unpacks what he sees as an unsustainable tension between the demands of functionalism and the claims of extended cognition:

> All varieties of functionalism contain a parameter that controls how finely or coarsely functional roles should be specified (how much should be abstracted and ignored). If this parameter is set too fine, then one is committed to Martians who differ from us in minor ways not having mental states. If the parameter is set too coarse, then functional role specifications are too easy to satisfy, and systems that are intuitively non-mental wrongly count as mental. (2009, p.7).

Sprevak's point is that insofar as functionalism can be pitched at finer and coarser grains of analysis, counterintuitive consequences emerge for proponents of extended cognition. Either functionalism entails a rampant expansion of the mind into the world or it includes counterintuitive cases about what counts as cognitive. Neither outcome is desirable for the proponent of extended cognition.

Sprevak's conclusion is a bit more negative than is required here. The point for present purposes is simply that the issue of granularity reappears in Clark's proposal. If the body is merely one instrument among others within functional analysis, then any 'unique' position in structuring intelligent organization remains unexplained. The indiscriminate setting of functional grain leaves matters unclear as to what allows the body to play its 'enabling role' in cognitive activities. Extended functionalism, absent further qualification and precisification, operates at too coarse a level to usefully specify the organisational structures relevant to understanding the information-processing role of the body. It is not that Clark is wrong in proposing that the body has an enabling functional role within information-processing, but that this suggestion alone does not specify what the role amounts to or why it should prove important.

### 3.3 Wide Computationalism

What seems to be missing is a story about how abstract analysis is grounded in concrete physical structures. To flesh out this side of the story, I propose appealing to a form of 'wide computationalism'.

In its simplest form, wide computationalism says that some of the units of computational systems and processes can, on occasion, reside outside the individual, that the agent or individual does not form a privileged boundary around which to draw the limits of computational analysis (Wilson 1994, 1995, 2004; Hutchins, 1995; Losonksy, 1995; Wells, 1998; Kersten, 2017a).[44] To use a simple example, in explaining how cognisers solve certain types of arithmetic problems it seems to make as much sense to appeal to the computational work being on the page as it does to the work being done in the head (Wilson, 2004). When the interaction of information-processing structures inside the agent and information-bearing

[44] Though the view has taken on a few different forms over the years, this is its core claim.

states outside the agent become central to analysis, a wide computational story begins to take shape.

Traditionally, wide computationalism has been committed to "causal mapping accounts" of computational implementation (Wilson 1994, 1995, 2004; Hutchins, 1995).[45] These accounts articulate the conditions of implementation in terms of isomorphic mappings between computational and physical descriptions via causal transitions (Chrisley, 1995; Chalmers, 1994, 1996; Scheutz, 1999, 2001). Similarly, the wide computational version of the story requires finding causal transitions between physical states across agent and world that mirror the transitions among states in a computational model (Wilson, 1994; Hutchins, 1995).

However, more recently, wide computationalism has been re-aligned with 'mechanistic' accounts of computation (Kersten, 2017a, 2017b; Nowakowski, 2017). Unlike their causal mapping counterparts, mechanistic accounts frame the conditions of concrete computation in terms of functional mechanisms. For the computational mechanist, concrete computation requires a physical system to have component parts whose activities conspire to perform some function (Piccinini, 2007, 2015; Milkowski, 2013, 2015; Fresco, 2014; Dewhurst, 2018). Again, the wide computational story involves finding components and activities distributed across brain, body, and world, ones which conspire to support some computational function.

What is interesting about wide computationalism, in either of its forms, is that it follows from the underlying assumptions of computationalism. In particular, it gains a conceptual foothold via the location neutrality of computational analysis, what is also sometimes called 'substrate neutrality' or 'medium-independence' (Piccinini, 2007; Milkowski, 2013). The thought is that because computational analysis is itself location neutral it is at least possible that some of the features relevant to physical computational systems may, on occasion, reside outside the individual – in the case of causal mapping accounts, for instance, this means finding external states that isomorphically map to a given computational model; while in the case of

[45] Computational implementation refers to the conditions under which a physical system can be said to compute.

mechanistic accounts, it means finding environmental elements that conspire to support some computational function.

But the plausibility of wide computationalism does not solely rest with its foundational assumptions. It also gains support from empirical work. For example, research on spatial navigation in insects reveals formal, geometric properties of the environment, such as the solar azi muth, that are computationally relevant to explaining the construction of complex, internal representations (see, e.g., Sekuler & Blake, 1990). Wilson (1994, 1995) interprets such work as revealing the presence of a wide computational process, as it acknowledges the unique computational role of states beyond the individual within perceptual processing. Hutchins (1995), moreover, has long argued that ship navigation is best viewed as a cognitive process fundamentally involving a distributed computational architecture. That is, because the task of ship navigating is not reducible to the individual actions of the team members on a ship it should be seen as a process involving a distributed computational system. Finally, in Kersten (2017a), I argued that the navigational system in bats is best interpreted as instantiating a wide computational system in virtue of the systematic relation it exploits between an organism's morphology, acoustic environment, and neural systems.

For each of these authors, there are a natural set of phenomena usefully understood in terms of implementing wide computing systems. Of course, this is only a small sampling of the work that has been appealed to, others include: Losonsky's (1995) discussion of development walking systems and child development, Wilson's (2004) examination of Ballad's (1990) theory of animate vision, Kersten & Wilson's (2016) study of music perception and performance, Kersten's (2017a) analysis of spatial navigation in sightless individuals, and Nowakowski's (2017) discussion of insect morphology and mantis shrimp. Taken in conjunction with the view's conceptual foundations, and such considerations point in the general direction of the view's plausibility. Not only is wide computationalism conceptually coherent, but it also receives a good deal of support from animal and human psychology. For these reasons, I propose adopting the mechanistic formulation of wide computationalism in

what follows, though, for ease of exposition, I often simply refer to the view as wide computationalism.[46]

## 3.4 Taking a 'Wide' View on the Two Tensions

Let's return, then, to the two tensions.

### 3.4.1 Tension #1

Consider the first tension. The body-centrist was at pains to highlight the unique role of bodily actions and structures in cognition. The worry was that the extended functionalist failed to acknowledge the special contribution of the body in virtue of relegating it to being another functional node within a larger web of tools. The wide computationalist can accommodate such worries by focusing on 'functional mechanisms'.

Functional mechanisms are systems whose component parts and activities conspire to support a given behaviour. The heart, for example, is a functional mechanism in virtue of the fact that its pumps and valves conspire to circulate blood around the body. Functional mechanisms come in all different shapes and sizes. Some are individual-bound, appealing to only the intrinsic properties and activities of a capacity; while others are more world-traversing, drawing on diverse sets of components and activities spread out over brain, body, and world.

For example, if muscle contraction is explained via describing the coordinated firing of a group of neurons in the motor cortex, then the explanation is narrow: the supporting mechanism is formed by elements and activities contained wholly within the organism. However, if the withdrawing behaviour of an organism is explained via, say, muscle contraction and the presence of predators, then such an explanation is wide: the supporting

[46] The reason for adopting the mechanistic versus causal formulation is that, in addition to helping resolve the two tensions, the mechanistic view is able to better accommodate concerns about 'implementational' and 'triviality' in virtue of providing additional constraints on computational implementation (see Kersten, 2017a).

mechanism spreads out across elements and activities internal and external to the agent. The width of a mechanistic explanation depends, in many cases, on the properties and activities (either wide or narrow) it identifies (Menary, 2007).

Framed in terms of functional mechanisms, the claims of the body-centrist become fine-grained statements about the set of bodily and neural structures responsible for delivering cognition and perception. What makes sensorimotor contingencies crucial to vision, for example, is not that they pick out some special set of properties possessed only by the body, but rather that they track certain causal relations in a tightly integrated functional mechanism. They identify the wide properties of the supporting perceptual mechanism. The constitutively embodied systems relevant to the body-centrist are really specific types of functional mechanisms localised to the brain-body complex.

Consider the extended functionalist side of the equation. In applying the method of computational analysis to world-individual spanning systems, wide computationalism also maintains a commitment to the location neutrality of cognition. What matters are the abstract computational properties responsible for carrying out the function in question, not the physical properties responsible for realising the system as a whole. This focus means that the extended functionalist's emphasis on location neutrality is also preserved within the wide computationalist picture. Focusing on the relation between computational models and physical systems allows wide computationalism to retain a commitment to the type of abstract analysis relevant to extended functionalism.

By re-envisaging body-centrism in terms of the implementation of wide computational systems, a space is opened up for the tight integration of bodily and neuronal processes in support of cognition (i.e., functional mechanisms) and the locational neutrality of computational implementation (i.e., wide computational analysis). The special status of body structures becomes a more general class of wide computational systems. The difference is that whereas some functional mechanisms are instantiated within individuals (as highlighted by body-centrism), others are instantiated by the brain, body and world (as highlighted by

extended functionalism). The dual focus allows of computationalism to systematically connect abstract analysis with the concrete organisation and structure in physical mechanisms.

But what exactly is the functional organisation that proves relevant to wide computationalism? What connects the abstract analysis to concrete functional mechanisms? In Kersten (2017a), I adapted Piccinini's (2015) implementation conditions to offer three conditions for wide systems.

First, a wide system must be a kind of functional mechanism. The system has to possess properties internal and external to the agent that organise so as to produce a given behaviour. Second, one of the capacities of the system must be the ability to compute at least one function. The system must be able to map from an input I (and possibly internal states S) to an output O. The system's behaviour must satisfy at least one abstract description mapping. Third, the system must compute its function via the manipulation of medium-independent vehicles. The vehicles must be transformed over the course of a computation in virtue of the system's sensitivity to some part of the vehicle's structure – in the case of digits, for example, this would involve processing the vehicles in virtue of their syntactic structure, while in the case of neural representations it would involve processing the vehicles' systematic relational structure.

Consider Turing machines as an example. Turing machines usually have three basic components: a finite state controller (or automaton), a read-write head, and a tape. Together these define the 'structural architecture' of the device. The tape, divided into equal parts, contains characters from a finite set (e.g., '1' and '0's). The head reads characters on the tape and performs various actions, such as erasing or writing, changing the internal states, or moving along the tape.

The standard line is to interpret the structural architecture as being implemented in the agent – for example, the tape is construed as an internal memory store, the automaton as a CPU, and the read-write head as a central executor. However, as Wells (1998) points out, an equally viable interpretation, one that aligns closer to Turing's original vision, is to view part

of the structural architecture as itself implemented in the environment.[47] Wells calls this the 'interactive' reinterpretation. On this interpretation, part of the Turing machine's computational machinery is distributed into the world.

Interpreted along interactivist lines, Turing machines offer nice examples of wide computing, meeting each of the three conditions. First, as an example par excellence of functional mechanisms, Turing machines support complex computing behaviours in virtue of the interaction and activity of their component parts (e.g., the interaction of finite state controller, read-write head, and tape). Second, one of the capacities of any Turing machine is the ability to compute some function. Any Turing machine will map at least one set of inputs to a set of outputs, such as computing an arithmetical sum, for example. Third, the characters processed by Turing machines (i.e. digits) are classic examples of medium-independent vehicles. They are distinguishable in virtue of where they fall along a string, while also being transformed in virtue of the system's sensitivity to some part of their syntactic structure.

So, not only do Turing machines involve functional mechanisms that span agent and world, but they also compute their functions via manipulation of medium-independent vehicles. What this shows, I think, is that wide computationalism articulates specific conditions under which physical systems can be said to compute. Unlike extended functionalism, wide computationalism offers a detailed picture of the profile of functional structures and organisation relevant to implementing computational systems.

I should qualify how the term 'input' is being used. It might seem like the term is being used as both part of the system (condition one) and something outside the system (condition two). The confusion stems from the fact that there are two ways to understand what an input is, corresponding to the two interpretations one might have of Turing machines. On the 'internalist' interpretation, an input is just a causal trigger for some internal computation. It is

[47] Wells' (1998) thinks there are a number of benefits of moving to this interactivist interpretation. Not only does it align closer to Turing's origin vision of pencil-plus-agent computer, but also it solves what he calls the 'evolutionary' and 'transduction' problems. I won't go into these here, but they jointly provide a plausible motivation I think for moving to the interactivist picture. For further discussion, see Wells (1998, p.275-80).

an instruction telling the system how to behave. Since the structural architecture is implemented solely in the agent, it makes sense to see inputs as being 'external' to the system. Inputs are simply the causal catalysts outside the system that trigger its internal processing; they are functional 'entries' or 'exits' to the system. However, on the 'interactivist' interpretation, inputs 'come back' to the system. For example, because the tape is located in the external environment, it acts a functional node, linking the finite state controller to the dynamics of read/write head in an organisationally closed way. The system becomes functionally closed. Understood in this light, the 'inputs' are no longer coming from outside the system, but are themselves part of the system (see Villalobos & Dewhurst (2019) for further discussion).

However, the current proposal might seem to side too much with the extended functionalist, in effect devaluing the role of the body by focusing too heavily on abstract analysis. But recall that what proves crucially relevant to the mechanistic formulation of wide computationalism, as opposed to the causal mapping account, is physical implementation – this is what many see as the advantage of adopting mechanistic approaches to computation (Milkowski, 2013; Fresco, 2014; Piccinini, 2015; Dewhurst, 2018). The uniqueness of the body is delivered within the mechanistic account via the requirement to specify the concrete functional organisation relevant to abstract computational systems (i.e. functional mechanisms). The body is simply the mostly evolutionarily integrated unit of selection. The type of functional mechanisms that will often prove most relevant to sustaining cognitive activities are ones more often than not localised to the bodily envelope (Carruthers, 2006). So while it is true that wide computationalism emphasises abstract analysis, it does not do so at the expense of neglecting what it is that makes the body special.

Should the body-centrist insist that the body has some further unique role to play within cognition then the worry becomes they are beginning to impose a conceptual limit on theorising not unlike the individualist of old. Recall that for the individualist, such as Fodor (1980), only internal states, by conceptual necessity, could prove relevant to individuating

mental states. What was explanatorily and taxonomically important to individualists was that states play certain roles within individuals' internal mental economy. In a similar vein, if the body-centrist insists, over and above accounting for the unique role of the body, that the body has some special status, it seems fair to say that they are imposing an *a priori* claim on what cognition has to be, analogous to that of the individualist, which is, for a number of independent reasons, perhaps not a desirable path forward in theorising about the mind (see, e.g., Wilson, 1995).

### 3.4.2 Tension #2

Consider the second tension. Recall that autopoietic enactivism was unimpressed with the blurring of the organism/environment boundary offered by extended functionalism. The worry was that autopoetic systems, in virtue of being of autonomous systems, demarcate a boundary between organism and environment that ultimately undercuts the ability of cognitive systems to be extended. Note the important role played by the concept of autonomy. Because autopoetic systems are autonomous systems, they cannot be extended systems; some have even argued that autonomy is what fundamentally differentiates autopoietic enactivism from other approaches in cognitive science (Thompson, 2007; Di Paolo, 2009; Thompson & Di Paolo, 2014). One strategy for resolving the second tension is therefore to show that the perceived gulf between autonomous systems and extended systems is not as wide as some suggest. If extended systems, such as the ones offered by wide computationalism, are also autonomous systems, then this may go some way to assuaging the tension.

Consider, then, the several faces of autonomy.[48] First, in metaphysical terms, autonomy addresses the way in which certain systems are self-determining, i.e. how certain systems specify their own domain of interaction. For example, in order to extract energy from its environment (glucose in the bloodstream), a cell will 'trap' glucose in its membrane wall. This

[48] I am largely following here the dialectic laid out by Villabolos and Dewhurst (2019).

allows it necessary time to extract the requisite energy, what is called phosphorylation. The cell construct or specifies the ways it interacts with its environment. Second, in formal terms, autonomy deals with the closed (or circular) organisation of certain systems, i.e. how certain systems fail to receive inputs in the form of external instructions. To return to the cell example, during phosphorylation, a cell will also dispose or eject electrons from the transformed glucose. The external stimuli (the glucose molecules) do not function purely as 'instructions' that precipitate a certain output (electron disposal). Rather, the external stimuli act as 'triggers' for that precipitate a response given the cell's own dynamics (phosphorylation). Finally, in material terms, autonomy describes the constitutive precariousness of systems, i.e. how living system need to sustain themselves in the face of thermodynamic decay.[49] Di Paolo and Thompson (2014), for example, write: "When a process is enabled by the operationally closed network and by external processes as well, if the network is removed and the process remains – in new circumstances – thanks only to the external support, than that process is not precarious (p.72). Again, phosphorylation is a response to precariousness faced by the cell. It is the cell's attempt to sustain itself in the face of fluctuating power supplies.

Wide computationalism can accommodate each of these different sides of autonomy. To see how, return to the Turing machine example. First, on the interactivist interpretation, a Turing machine can be self-determining insofar as the system's functional structure specifies how it should behaviour. When the reader detects a particular inscription on the tape, for instance, the resulting change in state of the machine is determined by the configuration of the machine, not the nature of the inscription. If the same digit were used in a different Turing machine, it would trigger a different change in the machine's states and actions (e.g., it simply might bypass the inscription). How a Turing machine specifies its syntactic constraints, what Wells (1998) calls the 'task architecture', determines how its interactions unfold in any given case. As Villabolos and Dewhurst (2019) note: "A physically instantiated Turing machine,

[49] See Villabolos and Dewhurst (2019) for further discussion of each of dimension.

through its structure and configuration specifies its own domain of interactions (perturbations) and meets the environment on its own terms."

Second, Turing machines can also be organisationally closed when they are no longer open to the environment. Recall, for instance, that the tape (which is located in the environment) acts a functional node between a system's effector and sensor dynamics (the read-write head), and the control system (the finite state controller). The reader, via the finite state controller, influences the writer part of the head, which in turn influences the reader part via the medium of the tape. Such a scheme shows how a Turing machine can lack inputs and outputs in the functional 'entries' and 'exits' sense. The system is functionally closed. This does not mean that there is no distinction between what is inside and outside the system. Rather, the functional dynamics of the system remain the same regardless of which environment the Turing machine finds itself. What is functionally included depends on the particular coupling established by the system.

Finally, as a matter of implementation, Turing machines can be as robust or precarious as the engineering requires. A Turing machine could, in principle, be designed to address a problem relevant only to its own survival. Such a system could therefore be involved in sustaining itself in the face of precarious environmental engagements. So long as a system is operationally closed, it can be more or less precarious. [50]

So, it seems that paradigmatic examples of wide computing, such as Turing machines, can, in principle, accommodate a central tenet of autopoetic enactivism. Wide computing systems, which are kinds of extended systems, are also plausibly interpreted as autonomous systems. This is significant because it shows how some autonomous systems can constitutively include elements in the environment. Of course, this is not to say that autopoietic enactivism only cares about autonomy, there are other important concepts such as 'adaptivity' (Di Paolo, 2005;

[50] Not all autopoetic enactivists may be happy with this discussion. But I think the interpretative waters here are muddy and contentious on this concept, and so the ball is ultimately in their court to achieve some consensus in spelling out the senses of closure, precarity and self-determination that are important for them. I want to thank Dave Ward for pointing this out to me (personal communication).

Thompson, 2007). Nor is it to say that autopoetic enactivism has to buy the idea that all cognitive systems are computing systems; enactivism may well have other reasons for rejecting computationalism, such as the 'deep continuity' between life and mind (Weber and Varela, 2002; Di Paolo, 2005; Thompson, 2007). But it does show that there is no deep tension between the underlying features of autopoetic enactivism and those of wide computationalism, and by extension extended functionalism.

The question is what causes the perceived tension if there is no deep incompatibility? One potential cause is that many of the systems analysed by autopoietic enactivists happen to be ensconced within the organism. Autopoietic enactivists often appeal to examples from biology to motivate their case. Such examples help to illustrate the way in which organisms differentiate themselves from their physical environment in the service of preserving homeostatic properties – cells, for instance, employ semipermeable membranes to regulate metabolic functions. Most autonomous systems therefore are, as a matter of empirical fact, restricted to the physical boundary of the organism. However, as we have seen, this does not mean that all systems need to be analysed as such. Nothing in the basic toolkit of autopoetic enactivism prevents its analysis from applying to world-involving systems. As Thompson and Stapleton (2009) point out: "The grounding of cognition in sense-making and sense-making in adaptive autonomy does not imply either internalism or externalism about the processes of cognition. The internalist/externalist debate rests on assumptions that are foreign to the enactive approach" (p.25).

But perhaps there is still resistance to the idea of marrying wide computationalism and autopoetic enactivism. Can the autopoietic enactivist not agree that wide computing systems can be autonomous systems, but nonetheless deny that this shows that autonomous systems are also extended cognitive systems?

The problem with this reply is that misplaces the burden of proof. The autopoietic enactivist has to show why, in principle, these two further theses cannot be conjoined. The wide computationalist can admit that not all cognitive systems are computing systems and yet still

insist that some of the autonomous systems that are wide computing systems might also be cognitive systems. The wide computationalist does not have to show that every instance of autonomous systems are extended systems, only that some might be.[51] Autopoietic enactivists, such as Di Paolo (2009), deny that such cases are even possible, that it follows as a matter of definition that autopoietic systems are localised to the physical boundaries of the organism.[52] But once it is admitted that wide computationalism is compatible with the basic resources of autopoietic enactivism, a space is revealed in which autonomous, autopoetic systems might also be extended computational cognitive systems.

Moreover, as Villalobos and Dewhurst (2017) point out, computationalism is also sometimes seen as being committed to representational and/or information processing theories of cognition (Di Paolo, 2005, 2009). As a result, anti-computationalism seems to follow from a general rejection of representational and information processing theories (Hutto and Myin, 2013, 2017). However, as we have seen, not only are there viable accounts of computation that eschew appeal to representation when specifying implementation, but such accounts are also able to accommodate central concepts of autopoietic enactivism (c.f. Issac, 2018).

So, to review, wide computationalism addresses the first tension in virtue of placing a dual emphasis on functional mechanisms and abstract analysis, while it addresses the second tension in virtue of highlighting a shared set of conceptual resources in terms of the notion of autonomy. Have the two tensions been conclusively resolved, then? Probably not. I probably have not satisfied those who outright reject some of assumptions on which the current discussion has been built, such as that computation is in some sense relevant to studying cognition (Hutto & Myin, 2013, 2017). However, what I have tried to do is to chip away at the general scepticism about the possibility of unifying the various strands of embodied cognitive science I have considered by showing how wide computationalism provides a non-circular, non-representational, empirically grounded account that can flexibly incorporate many of the

[51] Though for some suggestive examples, see Villabolos & Dewhurst (2017).
[52] For a review of the debate between autopoietic and extended theorist, see Di Paolo (2009) and Wheeler (2010)

demands of the enactive, embodied and extended theorist. A full-throated defence might have to wait for another time, but for now it should suffice to note that there are a number of other avenues one might take in responding to the sceptic (see, e.g., Kuokkanen & Rusanen, 2018; Raleigh, 2018).

## 3.5 Outstanding Issues

It will be worth concluding by addressing some outstanding issues. First, one might worry that simply showing that the wide computationalism and autopoietic enactivism are conceptually compatible does not really show that they are deeply congenial; they may just be superficially related, for example.

Recall though that the claim was not that autopoietic enactivism and wide computationalism were compatible in every instance. Indeed, there are probably going to be a number of cases where it is more productive to pursue one approach rather than the other. Rather, the point I simply tried to make, following others, was that a shared emphasis on a set of conceptual tools, such as the concept of autonomy, showed that some phenomena might be productively understood using either lens. As Dewhurst and Villabolos (2019) put the point in the context of Turing machines:

> It would be equally valid, though probably of little interest to the observer, to take the viewpoint of the wiring between the sensor device and effector device, from where the sensor provides an outputs and the effector consumes inputs. Since the Turing machine is a deterministic system, the observer would find a different but equally perfect mapping or function between these 'alternative' inputs and outputs.

It is, to a certain extent, a descriptive convention whether we choose to analyse a given system in the ways relevant to autonomous systems, i.e. as either a functionally open or closed system. Some of the systems studied by autopoetic enactivism may be the types of autonomous systems ensconced completely within the organism, but some may also stretch out into the world. It is these systems that may be usefully addressed by both approaches.

Second, one might be concerned that since being a wide computationalist is simply a more specific way of being an extended functionalist (as computation is still functionally specified

at some level of granularity), then perhaps there is a way to translate Clark's proposal into the current one.

One reason to think this is not the case is that functionalism and computationalism come apart more generally. Piccinini (2010), for instance, argues that whereas functionalism is a metaphysical thesis about the functional organisation of the mind, computationalism is an empirical mechanistic hypothesis about the brain. Even if the brain is a computing mechanism, the mind may or may not be the brain's computational organisation. If Piccinini is right, then functionalism does not entail computationalism, nor vice-versa. The relevance of this, given that extended functionalism and wide computationalism rely on largely the resources of functionalism and computationalism, is that wide computationalism, although importantly related to extended functionalism, offers a distinct approach to thinking about the brain/mind. It offers distinct resources over and above those of extended functionalism, as was shown particularly with the implementational conditions.

However, one might think there is a tension in adopting this response given what was previously said about the compatibility of extended functionalism and wide computationalism. But it is worth noting that wide computationalism and extended functionalism, though distinct, are nonetheless importantly related. Functional specification of various computational roles is a critical part of a given investigation. Yet it is when the views are raised to the level of theses about the mind and brain that important differences begin to emerge (Piccinini, 2010). There is no tension in insisting that extended functionalism and wide computationalism are compatible in broad strokes so long as one is clear about when and why this is the case.

Finally, one might be worried about unifying the 4Es using wide computationalism rather than dynamical systems theory.[53] One might think that something such as Chemero's (2009) view might be a more obvious candidate for reconciling the 4Es than wide computationalism.

[53] Dynamical approaches view cognition as the product of a complex interaction of internal and external factors, such as internal neural dynamics and environmental structures such as affordances. Differential equations are often used to the model these complex interactions within dynamical approaches (van Gelder, 1995; Chemero, 2009; Hutto & Myin, 2013).

There is a much that can be said about the relation between computationalism and dynamical systems theory, but I will restrict myself to a few brief points here.

There are a few reasons why one might think wide computationalism is *at least as good* a framework for integration as dynamical system theory. First, wide computationalism has a good chance of incorporating representations and content into 4E approaches. This is a desirable feature insofar as one wants at least some level of integration between the 4E views sympathetic to representational talk (e.g., Wilson, 2004; Clark, 2008b, Wilson & Clark, 2009; Raleigh, 2018). For instance, there is a longstanding debate surrounding whether dynamical systems are the types of systems that can support representational descriptions (Shapiro, 2011).

The classic example is van Gelder's (1995) Watt Governor Machine; in what sense the watt-governor can be said to involve representations has been a matter of considerable dispute. Now, regardless of what one makes of these debates, it seems plausible to say that dynamical systems theory does not support representational talk straightforwardly (c.f. Chemero, 2009). Computational analysis, on the other hand, seems to stand a reasonable chance of integrating representational talk, as there have been a number of attempts to connect mechanistic approaches to computation with representational views of cognition (see, e.g., Milkowski, 2013, ch.5). One might think alongside Egan (2010), for example, that representational descriptions sit on top of but are not reducible to computational descriptions. As long as computationalism forms at least a partial base for cognitive investigations, and representational views are related to cognitive investigations, then there is a good reason to think the two views are related in some way.

Second, wide computationalism, in being entailed by computationalism, avoids the some of the hardships associated with being a global sufficiency thesis. For instance, it is entirely compatible with wide computationalism that narrow computational systems are also productively studied by cognitive science. Unlike something such as Chemero's (2009) view, where all cognition is said to be explainable by appeal to dynamical explanations, a version of what Shapiro (2011) calls the 'replacement hypothesis', wide computationalism is pluralistic

about computational cognition. Wide and narrow investigations can exist side-by-side. Wide computationalism is the natural extension of the conceptual resources of standard computationalism. Moreover, given the historical difficulties associated with making global claims about cognition – for example, individualism in psychology a la Fodor (1980) – a more accommodating, pluralistic approach might be preferable when it comes to integrating the various 4E views. Bold conceptual claims about cognition traditionally have a way of not panning out; better to let many flowers bloom (Wilson & Clark, 2009).

Third, the shift to mechanistic computation also carries with it a number of specific advantages. For instance, mechanistic approaches, in being constitutive in nature (Craver, 2007), avoid debates around whether or not explanations are merely descriptive or explanatory, something that traditionally hampered discussions of dynamical systems theory (Shapiro, 2011). Making the mechanistic turn also gains notable methodological traction for wide computationalism. This is because adopting the mechanistic approach brings with it a host of useful interlevel experiments for identifying wide functional mechanisms, such as interference experiments, activation experiments and simulation experiments.[54] Such experiments supply substantive proposals for identifying the constitutively relevant parts of wide functional mechanisms. They reveal a potential path to a methodologically tractable version of the view; a desirable feature of any philosophical account hoping to be relevant to cognitive science (Irvine, 2014).

Of course, all this is not to say that there might not be interesting points of connection. For example, as form of generic computation, wide computationalism might entertain treatment in terms of analog computation (see Issac, 2018). Analog computation has interesting connections to dynamical systems theory insofar as dynamical explanations often invoke continuous variables. However, given that further discussion would take us too far afield, it should suffice for now to point out that there are a number of lines of connection one might

[54] For recent criticism, see Baumgartner et al. (2016, 2017).

pursue. So, while the above considerations are by no means decisive, they do suggest that there are some good reasons to think that wide computationalism is *at least as good* a candidate for unifying framework as dynamical systems theory.

## 3.6 Conclusion

What I have tried to do in the preceding discussion is show that a renewed focus on computation, particularly of the wide mechanistic variety, may help to resolve the two tensions troubling 4E cognition. I have tried to show that in virtue of placing a simultaneous emphasis on location neutrality, concrete physically implementation, and autonomous systems wide computationalism is able to accommodate many of the central insights and concepts relevant to embodied, enactive and extended theorists. This may not please everybody, but it is nonetheless an important result; as we have seen, discussions of 4E cognition have often eschewed mention of computation for one reason or another.

Of course, more remains to be done. I have said little, for example, about how wide computationalism integrates with other areas of cognitive science, such as information-theory or predicative-processing; nor have I probably satisfied those staunch anti-computationalists who reject the move to reconciliation in the first place. However, in spite of this, I think the current analysis offers an important step in pulling the 4Es a bit closer together. If wide computationalism does have a role to play in resolving the two tensions, then it means that despite some calls to the contrary, 4E cognition may further benefit from integrating computational approaches into its basic conceptual apparatus. If nothing else, the current discussion shows that the wide computationalist has something important to add to the 4E conversation.

# Introduction to Chapter 4

## Key ideas

Chapter 4 switches gears and turns from a focus on wide computation to mechanistic computation. In particular, it focuses on the conceptual foundations of mechanistic computation. There are two central ideas of the chapter.

The first is that the concept of 'medium independence'– the idea that a state or variable can be implemented in a variety of media in virtue of its degrees of freedom – is central to the success of the mechanistic approach to computation (MAC), the view that computations are implemented in physical systems in virtue of being concrete mechanisms that process medium-independent vehicles. The reason, I suggest, is that medium independence is, in part, what allows MAC to satisfy several desiderata on a theory of implementation.

To provide a brief example, one desideratum on a theory of implementation is that it should articulate how a system's capacities follow from what it computes. If computational explanations are those that trade in medium-independent properties, then the capacities of a system, such as its cognitive capacities, can be explained by reference to such properties. However, if computations and their vehicles are not medium-independent, then distinguishing a system's computational properties from its causal/physical properties becomes increasingly difficult. Both types of properties will have functional effects, but it will be unclear which is relevant to defining a particular computation. In other words, without a robust notion of medium independence, the explanatory adequacy of MAC as theory of implementation is cast in doubt. The concept of medium independence is central to MAC's success as a theory of implementation.

The second idea is that the twin concepts of *abstracta* and *illata* offer a constructive way to think about computation. As Dennett (2000) deploys the distinction, *illata* refer to a theory's theoretical posits, such as mass or weight in physics, while *abstracta* refer a theory's calculation-bound entities, such as centres of gravity. The suggestion here is that computations should be thought of as abstracta, as they possess both of its hallmark features. First, computational states and activities are robustly predictive and explanatory. Since computational explanations generate a number of precise and accurate predictions about cognitive systems, and explanatory success is our best guide to what exists, computations should be admitted into our ontology. Second, computational descriptions involve idealisations. Because only a subset of the regularities produced by a system's activities are relevant to its description, computations can only be ascribed under certain idealising assumptions. Computations should be treated as abstracta, as they are predictively and explanatory useful entities defined over the theoretical posits of the theory (the illata) via idealisation.

## The argument

The importance of this result, I suggest, is that it charts a way through a looming challenge facing MAC, what is call the 'abstraction problem'.

The abstraction problem says that the concept of medium independence is deeply problematic when combined with another of MAC's key notions: namely, constitutive explanation. The trouble is that structural explanations, which form one half of constitutive explanations, require detailed descriptions of how a system's component parts are organised and operate. Computational explanations not only require an abstract, functional characterisation of a system, but also a detailed, structural description of a system's component parts. However, if computational explanations are simultaneously abstract and concrete, then MAC seems to make talk of physical processes manipulating abstract medium-independent properties mysterious. As Hutto et al. (2019) describe the concern: "The trouble is that if

medium-independent vehicles are defined by their abstract properties then it is unclear how such vehicles could be concretely manipulated by their abstract properties" (p.278). The abstraction problem forms the argumentative target of the paper.

My response to the abstraction problem is to redeploy the illata-abstracta distinction. Interpreted properly, I suggest, the illata-abstracta distinction provides a useful way of thinking about the two sides of the mechanistic approach. On the one hand, if MAC's computations are treated as abstracta rather than illata, then computational descriptions can be literally true or false. Computational explanations are truth apt so long as they remain predictively and explanatory successful. On the other hand, re-conceptualising the physical/causal states and activities of mechanisms as illata allows one to retain the structural aspect of constitutive explanation. Structural explanations function as physical descriptions of a system's physical/causal properties, ones that can be used to precisely define the computations to which they give rise. So while computations and their vehicles, the ones described by abstract, functional descriptions, form the abstracta, the physical/causal states and activities of mechanisms, the ones described by the concrete structural explanations, form the illata.

But perhaps one is still sceptical of the abstraction problem. Perhaps the severity of problem has been overstated. Consider, for example, Putnam's (1975) famous wood peg-hole case. To explain why a cubical wood peg 15/16ths of an inch on each side will fit through a square hole that is 1-inch on each side but not a circular one, Putnam points out that we usually do not appeal to the micro-properties of molecules or atoms. Rather, we are satisfied with simply appealing to the macro-properties of the wood peg and hole as described, i.e., in terms of its geometric properties. Appealing to the micro-properties only complicates the story by introducing a number of irrelevant details.

While Putnam uses the example to argue against a reductionist approach to explanation, one might also use it to formulate an analogous worry about the abstraction problem. One might point out that regardless of what stance one takes on the type of explanation that is required to explain the peg-hole case there is still no further mystery about the relation between

the micro- and macro-properties. The micro- and macro-properties simply stand in a realisation relation – we are not confused about the type of explanation that suffices to explain the case, nor the relation between to the two types of properties. Thus, by the same token, the relation between physical and computational properties might simply reflect a difference between two types of descriptions: one physical, one computational. There is nothing particularly mysterious about the relation between the physical and computational properties over and above the standard micro-macro realisation relation.

The trouble is that while such reasoning might account for classical accounts of implementation, such as Chalmers' (1994) causal mapping account, which conceives of implementation as a mapping relation between computational and physical states, it doesn't really work for MAC. This is because, unlike causal mapping accounts, MAC tries to fit two concepts of hierarchy together. Computational descriptions are often said to fit into a mechanistic hierarchy in virtue of standing in part-whole relations. Kuokkanen & Rusanen (2018), for instance, claim that: "[T]hese higher-level properties, such as the medium-independent properties, can be realized in different lower-level properties that constitute different mechanisms at the immediate lower mechanistic levels (p.289)".

To illustrate, consider a computational level $C_1$ of some mechanistic hierarchy, consisting of the component parts (e.g., registers and circuits), their function, and their organisation. One way to further analyse $C_1$ is to describe the computational components of an underlying computational-level $C_0$, such as logic gates. Another way $C_1$ would be to describe medium dependent physical properties, such as voltages. Call this level P. But notice that it is not clear how P and $C_1$ are supposed to relate. While P describes physical implementational details, $C_1$ describes medium-independent properties. Yet if the properties of $C_1$ are not parts of the properties of P and vice versa, then $C_1$ cannot be at a lower level than P. There are two different mechanistic hierarchies, one computational ($C_0$, $C_1$, $C_2$ …) and one implementational ($P_0$, $P_1$, $P_2$ ...).

The point for present purposes is that the implementation relation used by MAC is not as straightforward as the micro-macro view would have it appear. In trying to combine the functional and structural aspects of explanation, MAC complicates the story of how physical and abstract properties are supposed to relate. It claims that physical processes causally manipulate abstract, medium-independent vehicles. This might seem strange, but it has to be accounted for. The abstraction problem is responding to a very real tension in MAC's particular mixture of concepts. The problem is not so easily dismissed.

## Broader context

Chapter 4 serves two wider ends. The first is to shore up MAC's conceptual footing. As mentioned, the abstraction problem threatens to undermine MAC's status as a workable theory of implementation. But it does more than this as well. Implementational accounts also help to justify the central commitments of artificial intelligence and computational cognitive science. As Chalmers (2012) points out: "In order for the foundation to be stable, the notion of computation itself has to be clarified. The mathematical theory of computation in the abstract is well-understood, but cognitive science and artificial intelligence ultimately deal with physical systems. A bridge between these systems and the abstract theory of computation is required." Articulating an answer to the abstraction problem therefore not only helps to solve a particular challenge threatening MAC but it also helps to secure the foundations of computational cognitive science and artificial intelligence more generally.

The second is to provide a meta-theoretic framing for MAC. Echoing a growing chorus of voices on the importance of meta-theoretical commitments in computational debates (see, e.g., Dewhurst, 2018; Schweizer, 2019), the chapter pushes for what I label the 'abstracta realist' position. As the name suggests, abstracta realism straddles the line between a realist ontology and an epistemic perspectivism. It says that while MAC's implementational descriptions are literally true or false, their truth-aptness depends on the predictive and explanatory utility of computational explanations. While computational descriptions do not, strictly speaking, refer

to features of the world (simpliciter), they are nonetheless true or false in the ways relevant to MAC. One way to see the current proposal, then, is as an attempt to situate MAC with respect to a more general family of realist/anti-realist positions in the philosophy of science.

# Chapter 4 – How to Be Concrete: Mechanistic Computation and the Abstraction Problem

## 4.1 Introduction

What does it take for a physical system to implement a computation? One popular answer is to say that a system must have the right physical or causal structure, sometimes referred to as 'physical/causal structure' proposals (Chalmers, 1994, 2011; Chrisley, 1995; Scheutz, 2001). Another is to say that a physical system must manipulate representations, sometimes referred to as 'semantic' proposals (Fodor, 1981; Pylyshyn, 1984; Sprevak, 2010).[55]

However, more recently, a third option has appeared. A number of authors have turned to the concept of 'mechanism' to help articulate an answer to the implementation question, what is collectively labelled here as 'the mechanistic approach to computation' (or MAC for short).[56] (Piccinini, 2007, 2015, 2018; Fresco, 2014; Milkowski, 2013, 2015; Dewhurst, 2018). According to MAC, computational implementation is best explicated within a mechanistic framework. Computational explanation is a species of mechanistic explanation, and computational mechanisms are a special type of functional mechanism. A physical system implements a computation only if it processes medium-independent vehicles in virtue of being a functional mechanism.[57] Computing systems are said to be a type of concrete computing mechanism.

[55] See Sprevak and Colombo (2018) for review.
[56] I am concerned here with implementational questions – i.e. what it takes for a physical system to implement a computation – so I leave aside questions of how to individuate computations. For discussion, see Sprevak (2010).
[57] Proposals differ on what exactly this amounts to. Piccinini (2015), for instance, maintains that computational mechanisms have to support teological functions; while Milkowski (2013) maintains that they need to be responsive

The aim of this paper is to take up a recent challenge to MAC, what has been labelled 'the abstraction problem' (Kuokkanen and Rusanen, 2018). The abstraction problem says that one of MAC's central pillars – medium independence – is deeply confused when applied to the question of computational implementation. The concern is that while it makes sense to say that computational processes are abstract (i.e. medium-independent), it makes considerably less sense to say that they are also concrete processes *of* a mechanism. The worry is that explicitly incorporating talk of mechanisms into an account of computational implementation makes it difficult to articulate in what sense computational processes and vehicles are supposed to be abstract. While the abstraction problem has received some passing attention, no sustained analysis has yet been offered.

In Section 4.2, I describe two of the key features of MAC, showing how these help satisfy a number of desiderata on a theory of implementation. In Section 4.3, I outline the abstraction problem, and show how it challenges MAC's general use of medium independence. In Section 4.4, I examine a recent response to the abstraction problem from Kuokkanen and Rusanen (2018). I argue that Kuokkanen and Rusanen's response comes up short insofar as it makes problematic trade-offs among the various desiderata. In Section 4.5, I diagnose some of the underlying issues of the debate and outline a general dilemma facing MAC. The worry is that MAC either has to give up being an objective theory of implementation or it has to concede the abstraction problem, and so reintroduce triviality concerns. Finally, in Section 4.6, I propose a solution to the abstraction problem based on the 'illata-abstracta' distinction. I argue that conceiving of computations as *abstracta* rather than *illata* provides a way to avoid the proposed dilemma and articulate a notion of medium independence that addresses the abstraction problem.

to reliable causal differences in the input stream. These difference aside, most authors agree that it is something to do with a system's mechanistic structure, at core, that allows a physical system to implement a computation.

## 4.2 MAC

Broadly speaking, there are two main features that distinguish MAC from other proposals about implementation.

The first is its use of a 'constitutive' notion of explanation (Machamer, Darden, and Craver 2000; Craver 2007, 2013). The idea is that rather than being separate from computational explanations, as is sometimes maintained by traditional accounts of computation (e.g., Marr 1982), implementational details are actually necessary to turn computational descriptions into full-blown computational explanations. Because computational explanation is a form of mechanistic explanation, it not only requires providing an abstract, functional characterisation of a system, but also a detailed, structural description of how the system's component parts are organised and operate, what are respectively referred to as the 'functional' and 'structural' aspects of explanation (Milkowski, 2013; Piccinini, 2015).

For example, to explain horizontal eye movement, one is said to have to not only describe the function being computed by the ocular motor system (an integration relation), but also how the neurons in the ocular motor system carry out the function via preserving morphic-relations between eye-velocity and eye-position (Leigh and Zee, 2006). To qualify as implementing a computational system, the ocular motor system must not only be capable of sustaining a description in terms of an input-output relation, but also a structural description in terms of the activities and organisation of its component parts.

The second feature is that computations and their vehicles are often said to be 'medium-independent'. To take two typical instances, Piccinini and Scarantino (2011, p.8) maintain that "[s]ince concrete computations and their vehicles can be defined independently of the physical media that implement them, we shall call them 'medium independent'"; while Milkowski (2018, p.942), in a related vein, says that "because computational mechanisms operate on vehicles whose states are specified in terms of their degrees of freedom, they are, to some extent, substrate-neutral."

The trouble is that while many proponents of MAC rely on medium independence, the notion is rarely explicated in any rigorous way. The most precise formulation comes from Garson (2003): "Medium independence: The structure S–for example, the structure relation that obtains between the units of a sequence of action potentials–can be instantiated across a wide range of physical mechanisms" (p.927). For Garson, a state or variable is medium-independent if it (i) possesses certain structure (often interpreted as degrees of freedom) and (ii) can be implemented in different media in virtue of that structure. A physical state or activity only counts as a computational vehicle if it is able to be implemented in a variety of media in virtue of its structure. Medium-independence is a comparative property. It is a property shared by several types of processes, rather than a property of a token physical process.

For example, consider the vehicles of a standard digital computer. According to Garson's view, one explicitly endorsed by Piccinini (2015, 2018), digits are medium-independent insofar as they are (i) distinguishable by a computer's processors in virtue of where they lie along a string (i.e. they can be processed in virtue of their structure) and (ii) insofar as they can be implemented in a variety of media (i.e. they can be realised in silicon chips or vacuum tubes). The same computational vehicle (digit) can be implemented in completely different materials (silicon chips or vacuum tubes) in virtue of possessing the right structure (where they lie along a string).

Another way to explicate medium independence is to contrast it with multiple realisability. As Ritchie and Piccinini (2018) describe the difference, one way to think about multiple realisability is as a relation that holds between a property of a whole system and properties and relations of that system's component parts. A higher-level property is said to be multiply realisable if it reflects the causal powers of its lower-level properties and relations. So, for example, replacing a lithium battery with a magnesium-copper lemon cell reliably produces the required voltage to power a LED light because both types of batteries, despite varying in material composition, contribute the same causal power to the circuit.

In contrast, medium independence can be thought of as a more restrictive claim about how the vehicles of computation are implemented. Implementing a computation only requires that, whatever the physical medium, the vehicle possesses the right degrees of freedom and functional organisation. While multiple realisability is defined in terms of specific physical effects (e.g., producing voltages), medium independence is defined in terms of the relation between variables (e.g., a cell in a Turing machine taking on either a '1' or '0'). So while everything that is medium-independent is multiply realisable, not everything that is multiply realisable is medium-independent (cf. Polger and Shapiro [2016]).

The going method for evaluating theories of implementation is to weigh them up against different desiderata. The general idea is that if a theory of implementation does a better job accommodating a given set of desiderata than its rival, then it is a superior account of implementation. Richie and Piccinini (2018, p.193) provide three such desiderata:

1) *Metaphysical adequacy*: the account should entail there is a fact of matter about whether a physical system computes.
2) *Explanatory adequacy*: the account should say how a system's capacities can be explained by what it computes.
3) *Extensional adequacy*: the account should ensure that paradigmatic cases of computing systems do compute and paradigmatic cases of non-computing systems do not compute.[58]

Call these the 'adequacy conditions'.

Consider how MAC fares with respect to each condition. First, MAC seems to make computational implementation rely on specific features of the world. Only those systems which involve functional mechanisms processing medium-independent vehicles qualify as

[58] There are other desiderata that are sometimes offered, such as miscomputation, but these three broadly capture the majority generally offered by theorists (see, e.g., Fresco [2014], Milkowski [2013], Piccinini [2015] Egan [2018], Sprevak [2018]).

computing systems. It entails a matter of fact about whether a physical system computes. Second, since computational explanations are only those which implicate medium-independent properties, the capacities of the system, such as its cognitive capacities, can be explained by reference to a system's such properties. It provides a naturalistic basis for explaining cognition. Finally, because only those systems that process medium-independent vehicles count as computing systems, paradigmatic cases of computing, such as Turing machines or calculators, count as computing systems, while non-paradigmatic cases, such as digestive systems or solar systems, fail to qualify. This last condition is one often said to be possessed by MAC but not its rivals, such as causal mapping or semantic accounts (see, e.g., Ritchie and Piccinini [2018, p.198]). So, not only does MAC provide an objective, naturalistic basis for the study of cognition, but it also captures many of our taxonomic intuitions about computing systems.

For present purposes, the adequacy conditions are also interesting because they offer a useful benchmark for judging the strength of potential responses to the abstraction problem. Given that a successful theory of implementation should accommodate at least the above three conditions, any response which gives up one or more is a worse answer for it. Sacrificing one or more of the adequacy conditions should only come at a high price. In this way, the adequacy conditions provide a good yardstick for measuring the success of any response.

## 4.3 The Abstraction Problem

As mentioned, some have recently questioned what it would mean for a property to be 'medium-independent' in the sense required by MAC. Hutto et al. (2019), for example, write: "The trouble is that if medium independent vehicles are defined by their abstract properties then it is unclear how such vehicles could be concretely manipulated by their abstract properties" (p.278). There seems to be a tension in saying that abstract, medium-independent properties are also manipulated by concrete physical processes.

For example, consider Piccinini and Bahar's (2013) discussion of neural spike trains – the time-series electrical signals recorded from individual neurons. According to Piccinini and Bahar, neural spike trains are medium-independent because they (i) depend on functionally relevant aspects of the neural events, such as firing rates and timing, and (ii) because they can be implemented in different physical mediums, such as silicon chips. Brains are said to perform a generic form of computation in virtue of manipulating medium-independent vehicles.

The problem, as Hutto et al. see it, is that Piccinini and Bahar's account fails to explain why neural spike trains actually possess medium-independent properties. They write, for instance: "Understanding how neural processes can be sensitive to concrete, medium independent properties presents no conceptual difficulty. By contrast, we have no conception of how concrete neural process could causally manipulate, abstract medium independent vehicles" (2018, p.278). There is no explanation of how the causal manipulation of medium-independent vehicles could be even (in principle) achieved.

MAC's trouble stems from trying to combine two distinct notions: constitutive explanation and medium independence. The first says that good mechanistic explanations tend toward full structural detail; that good mechanistic explanations provide details about how a mechanism's component parts and activities are organised and operate, what Boone and Piccinini (2016) call the *requirement of maximal detail*. The second, however, says that good computational explanations are necessarily abstract; that good computational explanations only care about preserving certain degrees of freedom in their descriptions. When combined, these two notions seem to be at odds in computational explanations.

Haimovici (2013) was the first to point this out, albeit in a slightly different form. According to Haimovici, MAC faces a dilemma: either computational explanations are functional, in which case they cannot provide full structural detail, or they can provide full

structural detail, in which case they cannot be multiply realisable.[59] Either mechanistic computation is functional or it is structural, but it cannot be both. While Haimovici presents the dilemma as one between functional and structural properties, the tension is the same as the one pointed out by the abstraction problem. In trying to retain both the functional (abstract) and mechanistic (concrete) aspects of computational explanation, MAC is seemingly led into mysterious talk of concrete physical processes manipulating abstract medium-independent properties.[60]

To further drive home the problem, consider how MAC fares without a notion of medium independence. First, if MAC no longer identifies the abstract, medium-independent properties defining of computation, then it loses its claim to objectivity. It's now unable to say when it is, strictly speaking, true or false to say of a physical system that it implements a particular computation. It fails to provide a metaphysically adequate theory of implementation. Second, if computations and their vehicles are no longer medium-independent, then it is unclear which properties are relevant to explaining a system's capacities.[61] Without a notion of medium independence, distinguishing a system's computational and physical properties becomes increasingly difficult. Finally, disrupting MAC's use of medium independence appears to complicate the view's ability to distinguish paradigmatic and non-paradigmatic cases of computing. Non-paradigmatic cases, such as digestive systems or grain sieves, now qualify as computing mechanisms, insofar as they represent functional mechanisms but ones which do not process medium-independent vehicles. In short, undermining MAC's use of medium independence seriously complicates the view's status as a workable theory of implementation.

[59] Haimovici (2013, p.175) assumes that multiple realisability and medium independence roughly amount to the same thing.
[60] Coelho Mollo (2018) also picks up on this dilemma, although he does not formulate in so stark of terms as the abstraction problem.
[61] To be clear, it is the entities that the computational explanations refer to that are observer independent, not the explanations themselves.

## 4.4 A Response

Kuokkanen and Rusanen (2018) (henceforth K&R) have recently offered a reply to the abstraction problem. Their suggestion is that the abstraction problem equivocates between two important senses of 'abstractness'.[62]

First, there is an 'epistemological' sense of abstractness. According to this sense, concrete computations and their vehicles can be defined independently of the physical media that implement them. Calling a vehicle 'medium-independent' is simply a matter of omitting certain features from one's descriptions. Second, there is a 'metaphysical' sense of abstractness. According to this sense, concrete computation and their vehicles possess abstract properties. Calling a vehicle 'medium-independent' is to claim that its properties are themselves metaphysically abstract.

To motivate the distinction, K&R offer the case of modeling the famous plane the Spirit of St. Louis. In the process of modeling this plane, they suggest, one would normally have to omit a number of irrelevant details, such as those concerning its aerodynamics. But doing so would not affect the plane's original properties. The process of modeling merely abstracts away from irrelevant details, it does not change the properties of the thing being modelled.

There is an important difference, in other words, between 'abstractness' as a feature of our descriptions (the epistemological sense of abstraction) and 'abstractness' as a feature of the processes being described (the metaphysical sense of abstraction). As K&R make the point: "The medium independence of vehicles should be interpreted as a claim that the computational models describe the target phenomena in an abstract away rather than a claim that the properties or processes of target systems can exist in an abstract way" (2018, p.288-9). So, while it makes sense to say that the descriptions of concrete vehicles are abstract (medium-independent), it does not make sense to say that concrete vehicles are metaphysically abstract.

[62] Piccinini (2015) provides something close to this proposal as well, but given that his articulation precedes the abstraction problem I focus on K&R's response instead.

There is much to be said for K&R's proposal. It offers an elegant and tidy solution to the problem, seemingly preserving a sense of abstraction that is at once intuitive yet servicing of MAC's needs as a theory of implementation. The problem is that the proposal comes at too high a price. It requires sacrificing metaphysical and extensional adequacy in order to preserve explanatory adequacy. It trades-off conditions (1) and (3) for condition (2).

First, notice that if K&R are right, then medium-independence is a property of our descriptions, and not a property of the world, as they write: "one can have descriptions of concrete vehicles as medium-independent and abstract, but one cannot have (metaphysically) concrete vehicles that are (metaphysically) abstract at the same time" (2018, p.290). Whether or not something is medium-independent is a product of how we choose to describe it. However, this seems to entail that MAC no longer specifies under which conditions it is (literally) true or false to say that a physical system implements a computation. Instead, it makes the truth of computational descriptions dependent on our explanatory needs; abstraction is a property of our descriptions rather than a property of the process we are trying to describe. MAC no longer provides an objective theory of implementation.

Second, notice that if the abstract properties of a computational vehicle are not what make something medium-independent, then any system satisfying an abstract description would qualify as computing. If medium independence is no longer an objective feature of the world, then it is unclear why, for example, one set of abstract properties, such as firing rating and timing, should quality as computational and not another, such as resonance frequency or refraction rate. If K&R's proposal is right, then MAC seems to lose its ability to distinguish paradigmatic from non-paradigmatic cases. There might any number of instances where otherwise counterintuitive entities, such as digestive systems or grain sieves, are usefully interpreted in terms of possessing abstract, medium-independent properties.

But perhaps there is a more favourable way to interpret K&R's proposal, one that can avoid the previous issues. Boone and Piccinini (2016), for instance, point out that the concept of abstraction not only has an *epistemic* role in mechanistic analysis but also an *ontic* role. It helps

to explain a phenomenon at a particular degree of generality by describing the general types of components, properties, and organisational relations which constitute that phenomenon.

For example, to explain rat navigation researchers often abstract away from the details of a particular rat and event in order to identify the general types of components, activities, and organisation common to rat navigation. It is not only that the explanations are more or less general but the actual phenomenon itself. So perhaps this is the sort of abstractness K&R have in mind: medium-independence is a real property of (some) concrete physical vehicles, but one described by suitably abstract descriptions. If so, then this would save K&R's proposal from having to give up metaphysical adequacy, as medium-independence would still be a property *of* certain concrete vehicles. Call this the 'ontic' interpretation.

Crucially, for this interpretation to work, one needs to have a method for determining whether it is the explanation or the explanadum phenomenon that is more or less general. Without one, a phenomenon's generality could be determined or fixed by the generality of one's description. One plausible method for establishing a phenomenon's generality would be to appeal to its cross-situational stability (Boone and Piccinini, 2016). For example, in the case of rat navigation, we might say that because rats often make particular types of navigational errors in water mazes this reveals something important about how spatial maps are used in their memory. Going beyond specific token instances allows researchers to identify the robust, cross-situational properties that form general features of rat memory.

Notice that when couched in the mechanistic framework this method takes on a specific character. This is because mechanistic explanation is decompositional. Parts and activities at lower levels of a mechanistic hierarchy are part-whole related to phenomena at higher levels. As K&R remind us: "[T]hese higher-level properties, such as the medium independent properties, can be realized in different lower-level properties that constitute different mechanisms at the immediate lower mechanistic levels" (2018, 289). Rat navigation is a 'more' general phenomenon because of its position at the top of the mechanistic hierarchy. Explanatory shifts from higher-level phenomenon, such as rat navigation, to lower-level

component and activities, such as rat memory, involve a reduction in the generality of the phenomenon being explained.

The trouble is that this method only works if it is assumed that computational and mechanistic hierarchies systematically overlap. It assumes that computational descriptions fit into the mechanistic hierarchy in virtue of standing in part-whole relations, which is often not the case. This point is made forcefully by Elber-Dorozko and Shagrir (2018), and the discussion here largely follows their general argumentation, particularly (215—17).

Consider, then, a computational level $C_1$ of some mechanistic hierarchy, consisting of the component parts (e.g., registers and circuits), their function, and their organisation. On the one hand, $C_1$ can be analysed by describing the computational components of an underlying computational-level $C_0$, such as logic gates. But, on the other hand, $C_1$ can also be described as being implemented by medium dependent physical properties, such as voltages. Call this level P.

Notice that it is not clear how P and $C_1$ are supposed to relate if computations are part of a mechanistic hierarchy. While P describes physical medium dependent properties, $C_1$ describes abstract medium-independent properties. But if the properties of $C_1$ are not parts of the properties of P and vice versa, then $C_1$ cannot be at a lower level than P. There are two different mechanistic hierarchies, one computational ($C_0$, $C_1$, $C_2$ …) and one implementational ($P_0$, $P_1$, $P_2$ ...), and it is unclear how they relate on the ontic interpretation.[63] If the generality of a phenomenon is determined by its place in a mechanistic hierarchy, then the mechanist has to show how computational and implementational descriptions fit together, but they cannot do so solely in virtue of relying on part-whole relations.

I should also briefly mention that this ontic interpretation seems to come at a particularly high ontological price, as it seems to suggest that degrees of abstraction correspond to degrees

[63] The problem is made worse by the fact that computations are multiply realisable. There might be more than one computational hierarchy and mechanistic hierarchy.

of being. I do not want to make too much of this point here, but one might be wary of taking on such a substantive ontological commitment in order to solve such a specific problem.

So, the ontic interpretation faces a dilemma, either it fails to explain how mechanistic and computational hierarchies fit together, in which case it makes a mystery of the generality of computations, or it fails to explain how to identify the generality of computations, in which case it does little to help K&R's proposal. Neither option is palatable. Given this, in what follows, I interpret K&R's proposal along the original epistemological lines suggested, as at least this version of the proposal makes headway on the problem.

To clarify, the worry is not that K&R's response is wrong full stop. Their proposal captures something important about what has gone wrong with the abstraction problem: namely, that it blurs a distinction between epistemological and metaphysical senses of abstraction. Instead, the worry is that, the ontic interpretation notwithstanding, the proposal comes at too high a cost. It strips MAC of much of what makes it initially appealing as a theory of implementation (i.e. metaphysical and extensional adequacy). Thus, if there is way to preserve this insight, while nonetheless retaining the view's ability to accommodate the various adequacy conditions, then such a view might be doubly appealing.

## 4.5 A Dilemma

Notice that K&R's proposal relies heavily on a distinction between metaphysical/epistemological interpretations of abstraction. They write, for instance: "Hutto et al.'s claim about the neo-mechanists's difficulties with 'explaining how…abstract entities can be causally manipulated by 'concrete…processes' would not make any sense if it was not committed to the metaphysical interpretation of abstract vehicles" (2019, 289). As far as K&R are concerned, Hutto et al. overemphasise the metaphysical interpretation of abstraction. Properly understood, medium independence only deals with the epistemological interpretation, and so the abstraction problem dissolves.

Hutto et al. (2019), however, are well aware of the distinction, writing: "[O]ne can have medium-independent descriptions of processes – a description which abstracts from certain substrate-related properties and mentions properties which can be found in different substrates – but one cannot have concrete vehicles that are medium independent" (278). In spite of being aware of the two senses of abstraction, Hutto et al. think that only the metaphysical sense is relevant to MAC. The reasoning seems to be that if MAC is to function as an objective theory of implementation, then it has to be committed to the idea that medium-independent properties are actual properties of the world. If it is not, then it fails as a theory of implementation.[64]

What this shows, I think, is that the nature and significance of the abstraction problem does not really turn on the concept of medium independence per se. While both sets of authors are aware of the two senses of abstraction, neither is moved by the other's considerations – each thinks that MAC needs to be committed to the other sense of abstraction to properly function as a theory of implementation (i.e. either the epistemological or metaphysical sense). What I think is actually going on is that both sets of authors are implicitly adopting a more general framing for MAC, and this is colouring how serious the abstraction problem is seeming and why MAC is taken to be able to function with only one or the other interpretation of abstraction.

Consider two such framings, then.[65] One is as a *realist* view of implementation. According to this perspective, a physical system implements a computation only in virtue of possessing the right properties. What makes digital computers and neural systems both computing systems is that they possess the right physical/causal structure. A physical system must actually possess the relevant properties – in this case, abstract, medium-independent properties

[64] Both sets of authors, for example, appeal to Polger and Shapiro (2016) in making the distinction,

[65] These two conceptualisations are not unique to MAC, they could just as easily apply to other accounts of implementation. They are nonetheless interesting, however, because they point to the way the nature and function of a theory of implementation can change depending on its wider framing.

or states – in order to qualify as a computing system – this is how MAC is usually interpreted (Milkowski, 2013; Piccinini, 2015).

The other is as an *anti-realist* view of implementation. According to this perspective, a physical system implements a computation not in virtue of the properties it possesses, but in virtue of how useful it is to interpret it as such; as we will see, Schweizer (2019) adopts something like this view. Because computers have physical activity that can be observed, predicted and manipulated, it makes sense to treat them *as if* they were implementing a variety of computations; while, conversely, since walls and rocks cannot conveniently be exploited to do computational work, it does not make sense to treat them as if they were implementing computations. The computational identity of a system reflects the interaction of the system's properties and human-centred interests (Sprevak, 2018b).

Notice that when framed as a realist view of implementation, as suggested by Hutto et al.'s reading, MAC fails to explain how medium-independent properties are concretely manipulated. It fails to reconcile how a state or activity can be abstract yet causally manipulated. But, when framed as an anti-realist view, as suggested by K&R's reading, the abstraction problem dissolves. MAC is simply not in the business of providing metaphysically abstract medium-independent descriptions, and so there is no abstract/concrete relationship to reconcile.

Both sets of authors are willing to give up, to a greater or lesser extent, one or more of the adequacy conditions because of their meta-theoretic commitments. Hutto et al., in adopting the realist framing, view metaphysical adequacy as indispensable. They see it as a requisite of MAC being a theory of implementation that it makes there a matter of fact about whether a physical system actually computes. K&R, on the other hand, in emphasising the epistemological interpretation of abstraction, downplay the significance of metaphysical adequacy. They see no issue with MAC being a purely epistemological tool. What matters first and foremost is that a theory of implementation explains a system's capacities.

The real trouble, however, is that regardless of which framing MAC chooses it still faces a dilemma. Notice that if the anti-realist framing is adopted, then this undermines one of MAC's major motivations (metaphysical adequacy); but if the realist framing is adopted, then MAC loses part of what makes it distinct as an approach to implementation (medium-independence). Accepting the realist conception carries with it the risk of flouting the abstraction problem, while accepting the anti-realist conception seems to entail the loss of MAC's claim to objectivity.

Consider the first horn. As mentioned, what initially made MAC desirable to some as a theory of implementation was that it made computing objective (Piccinini 2015). What united digital computers and neural processes, for instance, was that both could be implemented in variety of media in virtue of their degrees of freedom. If the anti-realist conception is adopted, however, then MAC no longer provides such an objective basis for computation. It no longer makes a system computational in virtue of possessing certain properties. Rather, what is relevant is how a physical system fits with our explanatory interests.

By way of contrast, consider that according to Schweizer's (2019) view, computational ascriptions should only be judged relative to how *useful* they are.[66] Whether or not a particular formalism ultimately proves interesting and useful comes down to how well it services our normative and descriptive needs. As Schweizer (2019) puts it: "[I]t's the only practical motivation for playing the interpretation game in the first place." Any account which satisfies these needs suffices as a theory of implementation.

Consider the second horn. If MAC accepts the realist framing, then it seems to have to give up the concept of medium independence. If MAC is committed to medium independence, and the abstraction problem casts doubt on medium independence's coherence, then MAC cannot remain committed to medium independence without jeopardising its wider plausibility. As Isaac (2018a, p.428) points out in a review of Piccinini's (2015) account: "medium

[66] Schweizer's (2019) account is generally representative of the anti-realist view.

independence does crucial work, establishing multiple realizability (122–23), while ruling out putative counterexamples to nonsemantic accounts, namely noncomputational input-output systems (e.g., digestive tracts [146] or mouse traps [153]).” Medium independence adds an important further constraint on implementation. Accepting the realist framing seems to come at the expense of giving up medium independence.

So, MAC seems to either have to give up what makes it initially desirable as a theory of implementation (metaphysical adequacy) or it has to concede the abstraction problem, and so reintroduce triviality concerns (extensional adequacy). Needless to say, neither horn is very attractive.

## 4.6 A Potential Solution

The abstraction problem put pressure on several of the roles MAC was supposed to fulfil as a theory of implementation. It played the metaphysical, explanatory and extensional adequacy conditions against each other via questioning the general coherence of medium independence’s status within MAC. Given this, a potential solution needs to find a way to retain the various adequacy conditions but also articulate a non-mysterious way for medium-independent vehicles to be manipulated by physical processes. This, I suggest, can be done by appealing to the ‘illata-abstracta’ distinction. While the original distinction is owed to Reichenbach (1938), in what follows I focus on Dennett’s (1981, 1987, 2000) presentation.

According to Dennett, there are two types of referents for terms in a theory: (i) there are the theory’s *illata*, which are its theoretical posits, and (ii) there are the theory’s *abstracta*, which are its calculation-bound entities. For example, in physics, terms such as mass, weight, position or velocity, might form a theory’s illata; while terms such as ‘centre of gravity’ might form its abstracta. What is interesting, according to Dennett, is that abstracta are both precisely definable in terms of illata, e.g., mass and position, and can be used to formulate interesting generalisations within a theory. For example, to calculate the moon’s movements, a point slightly off geometric centre can be used to account for the earth’s gravitational effects, and

such a point is defined precisely in terms of the earth's mass and position. While Dennett deploys the illata/abstracta distinction to develop his particular approach to intentionality and folk psychology, I am interested more generally in how the distinction might be used in what follows.

To that end, there is a potential misreading of the distinction I want to avoid. One way the distinction should not be read is saying that illata correspond to physical entities in the world, while abstracta function as purely theoretical, non-referential terms. This 'ontological' reading would presuppose the very full-blooded realism we need to avoid to escape the first horn of the dilemma. It would be making a claim about the reality of certain entities in the world (e.g., physical versus mental), and how our theoretical terms describe such entities. Rather, what distinguishes illata from abstracta is the role they play within a theory's explanatory practices. I will come back to this in more detail later, but the point to stress is that both entities are theoretical posits within a theory. Viger (2000) phrases the point quite nicely when he says: "[T]he difference between illata and abstracta *just is* their role in an explanatory practice, and not their ontological status. Illata and abstracta are, after all, equally theoretically posited entities" (original emphasis, p.134). As I construe it here, the illata-abstracta distinction advances an explanatory rather than an ontological thesis.

For present purposes, what is interesting about the illata-abstracta distinction is that it offers a third way to think about MAC's status as a theory of implementation. Recall the two previous framings. The anti-realist framing claimed that MAC's terms failed to refer and its sentences failed to be literally true or false. This led to issues accommodating metaphysical and extensional adequacy. The realist framing, on the other hand, claimed that MAC's terms did refer and that its sentences were, in fact, literally true or false. This, however, led to troubles articulating a notion of medium independence that avoided the abstraction problem.

The illata-abstracta distinction offers a third option: 'abstracta realism'.[67] According to abstracta realism, a theory's terms refer to calculation bound entities yet its sentences are literally true or false. Viewed through abstracta realist lens, MAC's computational descriptions address abstracta and its statements are truth apt depending on the predictive and explanatory utility they provide. The physical/causal states and activities of mechanisms, the ones described by structural explanations, form the illata of the theory, while computations and their vehicles, the ones described by the abstract, functional descriptions, form the abstracta.

There are two reasons to think computational states and activities should be treated as abstracta. The first is that such states and activities are robustly predictive and explanatory. As the history of cognitive science amply shows, treating cognitive systems as computational systems reliably generates a number of precise and accurate predictions. This is not just a heuristic. Explanatory practice is our best guide to what exists. If the success of physical explanations guarantee the reality of theoretical entities such as black holes or gravitational forces, then, by the same token, the explanatory success of computational explanations should bolster the case for the reality of computational entities.

The second is that computational descriptions overwhelmingly involve idealisations. In Dennett's classic example, beliefs and desires are ascribed to a system only under idealising *rationality* assumptions. It is only through the filter of rational considerations that intentional patterns become visible (i.e. using the intentional stance). Similarly, as numerous authors have pointed out, almost by definition computational descriptions involve some form of idealisation (see, e.g., Chalmers [2011]). It is only when one attends to a subset of the regularities produced by a system's activities that computational processes begin to take shape. Systems are only ascribed computations when their behaviour can be treated in terms of satisfying a particular input-output function.

[67] Dennett doesn't explicitly endorse this label, mostly because of his long standing worries about 'isms', but it suffices for present purposes because it nicely contrasts my view with the realist and anti-realist positions. I borrow the term from Yu and Fuller (1986).

More interesting, the move to abstracta realism offers a route through the previous dilemma. With respect to the first horn, if computational descriptions address abstracta, then unlike its realist counterpart, which conceives of medium-independent properties as illata, MAC's descriptions are literally true or false. Abstracta realism simply requires that medium-independent properties are definable in terms of the theory's illata. This is largely how MAC already views the relationship. Ritchie and Piccinini (2018) write, for instance: "degrees of freedom [medium-independence] reflect the independent dynamics of a physical system, and a system will only implement a computation, in a non-trivial sense, if the microstate transitions support counterfactuals that mirror the computational state transitions" (p.200). Physical systems do not actually have to possess medium-independent properties in order to make implementational statements true or false. Rather, the truth aptness of such statements depend on the precision and accuracy of the predictions and descriptions the theory offers using the abstracta.

With respect to the second horn, unlike anti-realism, tying computational descriptions to abstracta does not involve conceding the abstraction problem, and by extension re-introducing triviality concerns. Abstracta can, in fact, figure into causal explanations. They can explain why various changes take place in a physical system.

Consider the following example from Dennett:

> What did Connor do overnight to *cause* his boat to be so much faster? He lowered its center of gravity. Of course he did this by moving gear, or adding lead ingots to the bilge, or replacing the mast with a lighter mast, or something—but what *caused* the boat's improvement was lowering its center of gravity. (2000, p.358)

For Dennett, it makes as much sense to appeal to a boat's centre of gravity to explain its increase in speed as it does to appeal to more straightforwardly material factors, such as changing the boat's mast or rigging. Since it is the location of the boat's centre of gravity that makes a difference to the boat's performance, this is as good a reason as any to cite it as a cause of the boat's behaviour.

More to the point, what this shows is that explanations citing abstracta are not just causal shorthands, they can actually explain certain generalisations. In the same way, computations and their vehicles facilitate a whole host of interesting generalisations, such as those about retrieval rates in memory or processing speeds of novel stimuli. Differences in cognitive performance are better explained by citing how information is stored and transformed (computational explanations) as they are talking about how synaptic depolarisations affect action potential rates (physical explanations). Saying that medium-independent vehicles are manipulated by physical processes is no more mysterious than saying that centres of gravity are manipulated by changing the mass and position of a set of bodies.

The reason for why is also not particularly mysterious: abstracta are defined relative to illata. It is in virtue of being dependent on illata that abstracta can figure into causal explanations. The physical states and activities which constitute the illata of the theory are fully determinate of its abstracta; wherever the illata go, the abstracta go. Not unlike how a centre of gravity can be precisely formulated with respect to a set of bodies with mass and position, there are physical facts to which abstracta are similarly dependent.[68] Where the abstraction problem goes wrong is thinking that only illata figure into causal explanations. While this is true of the realist framing, it is not true of abstracta realism. Abstracta realism offers a way through both horns of the dilemma.

The question is, does abstracta realism also allow MAC to accommodate the adequacy conditions? Consider each again. First, with respect to metaphysical adequacy, abstracta realism ties the truth of implementation claims to abstracta. If a physical system possesses medium-independent properties, in the sense of being definable in terms of illata, then it

[68] One speculative suggestion here for how the relationship might be further fleshed out is to use Tyler Millhouse's 'simplicity criterion'. Roughly put, the simplicity criterion says that for any physical system, P, and a pair of computational models, C1 and C2, P implements C1 rather than C2 if there is an interpretation of C1 that is informationally more compressed than C2. The simplicity criterion provides a way of quantifying the degree of similarity between a physical system and a computational model using Kolmogorov complexity. Of course, more would be need to be said, but Millhouse's simplicity criterion might provide a useful way of formalising the relation between illata and abstracta (for details see Millhouse [2019, p.161-165]).

qualifies as a computing system. While its terms do not, strictly speaking, refer to features of the world (simpliciter), its sentences are nonetheless literally true or false. Second, in terms of explanatory adequacy, because abstracta are defined relative to illata, explaining a system's capacities requires not only describing its abstract functional structure but also its concrete composition and organisation. Functional and structural descriptions are equally important in trying to explain what a system computes and why. Finally, because it articulates a sense in which medium-independent properties can be manipulated by physical processes, abstracta realism not only ensures that paradigmatic cases will count as computing but also that non-paradigmatic cases will not. It leaves the taxonomic dimension of the theory untouched, preserving a notion of medium independence.

Not everyone is a fan of the illata-abstracta distinction, though. Ross (2000), for example, argues that there is no defensible basis for the distinction as it relies on dubious metaphysical assumptions. Dennett's particular interpretation, he argues, derives from Quine (1953), and is based on an outmoded 'Democritean faith' that we will be able to decompose everything into elementary particles and their relations (Ross, 2000, p.152). Besides not comporting well with Dennett's own non-reductionistic commitments, this type of reductionistic metaphysics is no longer supported by current scientific practice. Thus, the distinction is, at best, misguided and, at worse, patently false (see Ladyman et al. [2007, ch.4] for a similar worry).

Putting aside interpretative questions about Dennett's view for the moment, Ross's worry fails for the simple reason that it falls into the 'ontological reading' I warned against earlier. It misunderstands the source of the illata-abstracta distinction. The illata-abstracta is not 'read off' an ontological division in the world. It is not that physics describes the 'real' structures of the world (illata), while the special sciences re-classify higher-level patterns (abstracta). Rather, the distinction finds its basis in the explanatory practices of specific theories. In Dennett's hands, the distinction is used to elucidate the nature of intentional explanations. Characterising beliefs and desires as abstracta allows Dennett to show how our ontological commitments follow from the explanatory and predictive success of intentional explanations.

The same holds true of computation. As I have tried to show, computational explanations are abstract in just the same way as intentional explanations: they are idealisations that are predictively and explanatorily useful. Computational entities (medium-independent vehicles and processes) can be understood as abstracta because they are tied up in the predictive and explanatory success of computational explanations. Again, Viger (2000) makes the point admirably clear: "The abstractness of the entities we posit depends on the idealisation we make in positing them, but their being abstract is not relevant to our ontological commitment to them, which depends entirely on the success of the explanatory practices from which we posit their existence" (p.138). Rather than following from any outmoded metaphysical assumptions, the illata-abstracta distinction simply follows from the explanatory practices of different theories such as computation.

So, in a sense K&R were right, MAC is not committed to attributing metaphysically abstract properties to concrete physical processes. There is an important sense in which MAC is not making metaphysically spooky claims about medium-independent properties. However, unlike K&R this does not mean that medium independence is simply a feature of our explanatory practices. There is an important sense in which medium-independent properties are attributable to physical systems. This is the benefit of moving to abstracta realism. Thinking of MAC in terms of abstracta realism allows computational explanations to remain truth apt, while avoiding strenuous ontological commitments. It articulates a sense in which medium-independent vehicles can be manipulated by concrete physical processes. One way to see the present proposal, then, is as an elaboration of K&R's proposal, one which provides a wider framing in which to situate and expand their basic insight, but one which does not give up on any of the desiderata for a theory of implementation.

## 4.7 Conclusion

So, to recap, I set out to tackle a recent challenge to the mechanistic approach to computation, the abstraction problem. I did this by outlining two of the view's key features, constitutive

explanation and medium independence, and articulating how these features helped to satisfy a number of desiderata on a theory of implementation. Next, I outlined the abstraction problem and showed how it challenged MAC's use of medium independence, motivating the seriousness of the problem by articulating its effect on the adequacy conditions. Then, I examined a recent response to the problem from Kuokkanen and Rusanen (2018). I argued that K&R's proposal, while interesting, came up short, as it subtly reframed MAC as an anti-realist view of implementation, and so made problematic trade-offs among the various adequacy conditions. This revealed a general dilemma for MAC: either give up what makes it desirable as a theory of implementation (metaphysical adequacy) or concede the abstraction problem and give up medium independence, which reintroduced triviality worries (extensional adequacy). I responded by proposing a middle path: thinking of computations and their vehicles as abstracta. This, I argued, not only provided a way through the dilemma, such that it avoided the abstraction problem, but it also provided a way of retaining all three of the adequacy conditions. Hopefully, then, in answering the abstraction problem, I have helped to set MAC on a surer theoretical footing going forward.

# Introduction to Chapter 5

## Key ideas

There are two central ideas to Chapter 5.

The first is the notion of a 'Markov blanket'. Simply put, a Markov blanket is a formal, statistical tool that describes the probabilistic relations between a set of random variables. For a random set of variables, $S = \{X_1 \ldots X_n\}$, the Markov blanket of some random variable Y is any subset $S_1$ of S that is conditioned on other variables that are independent of Y. For example, in a Bayesian network, the boundary of a given node A includes all the 'parent' and 'children' nodes (those conditionally dependent on A), and the parents of all of their children. Any nodes not conditionally dependent on node A, or the parents of all of its children, are outside the boundary of the Markov blanket (Pearl, 1988).

In Ramstead et al. (2019a) hands, Markov blankets are said to open up the possibility of extended cognitive systems. This is because Markov blankets, in being hierarchically nested and fluid (i.e. Markov blankets nested within blankets and shifting in probabilistic relations), might come to include elements of the surrounding environment at higher temporal and spatial scales. Markov blankets are thought to reveal a novel route to cognitive extension in virtue of revealing the negotiable boundaries of cognitive systems.

The second is the notion of a 'transient extended cognitive system' (or TECSs). As Wilson and Clark (2009) describe them: "TECSs are soft-assembled wholes that mesh the problem-solving contributions of the human brain and central nervous system with those of the (rest of the) body and various elements of the local cognitive scaffolding." Because organisms often need to solve specific, context-dependent problems, such as catching a fly ball or navigating

with compass and map, they sometimes create temporary neural-body-world ensembles driven by action-perception cycles. These soft-assembly systems depend on the reliability and durability of the relationship between the agent and the external resource (e.g., body or environmental structure).

Clark's (2016, 2017) suggestion is that it is a short step from the notion of TECSs to a predictive processing based account of extended cognition. This is because predictive processing accounts help to flesh out the computational story behind why TECSs are formed. In the process of trying to minimise the difference between incoming sensory signals and prior expectations, organisms often engage in bouts of action-perception cycles with their surroundings. These circular, casual interactions help to offload computational work from the brain to the non-neural body and world. They sculpt the flow of information available for real time, on-going use. This creates a 'motor-informational weave', as Clark calls it, which is necessary to sustain the formation of TECSs. Larger problem-solving wholes are formed out of a need to structure and regulate information via exploiting external resources.

## The argument

The chapter makes two main claims. The first is that neither of the recently articulated predictive processing inspired proposals, Clark (2016, 2017) and Ramstead et al. (2019a), suffice to establish extended cognition. While both accounts make attempts at establishing cognitive extension via predictive processing based considerations, neither proposal, I suggest, succeeds. This is because, in virtue of relying on existing externalist assumptions, neither proposal provides anti-extensionists (those opposed to the idea of cognitive extension) with independent, well-motivated reasons for adopting extended cognition. The argument is not that PP makes no contribution or is completely orthogonal to discussions of extended cognition but, rather, that it is not *decisive* in settling the question of cognitive extension one way or the other as currently stated.

The second claim is that despite the shortcomings of existing attempts there is a route available to extended cognition via predictive processing. If one could show that there are certain features that are distinctive of cognitive processes that follow directly from accepting PP, and that extended systems, as a matter of fact, possess such features, then this should suffice to show that extended systems are genuine cognitive systems. While by no means decisive or uncontroversial, this, I suggest, provides one plausible route to cognitive extension. It shows how PP and extended cognition might be connected without relying on the contestable externalist assumptions used by Clark (2016, 2017) and Ramstead et al. (2019a). Cognitive extension, according to this argument, follows from the truth of PP. It offers a direct line to extended cognition from PP.

There is a worry that I want to address at this point. I do not develop it in the chapter, but it seems important within the larger context of the discussion. One issue with Ramstead et al.'s (2019a) proposal is that it seems to assume that appealing to a particular formalism (namely, Markov blankets) helps to settle ontological or metaphysical questions. They write, for instance:

> We cast ontological pluralism in terms of a multiscale formal ontology of cognitive systems. In the sense we are using the term, to produce a formal ontology means to use a mathematical formalism to answer the questions traditionally posed by metaphysics; i.e., what does it mean to be a thing that exists, what is existence, etc. Our formal ontology is effectively in the same game as statistical physics, in that it treats as a system sets of states that evince a robust enough form of conditional independence.

The claim is that 'formal' tools can help to provide insight into the ontological structure of certain phenomena, such as cognitive systems.

However, this assumption, if not deeply controversial, is, at least, too strong for present purposes. It requires the extended theorist to accept a specific metaphysics that they might well reject. The extended theorist might be wary of taking on such an assumption without further argumentation. Of course, there may be ways to justify this position – for example, Ladyman et al.'s (2007) 'structuralist' metaphysics might offer support. But it is nonetheless an extremely substantive assumption to make at the outset. Regardless of what one thinks of

the rest of Ramstead et al.'s (2019a) proposal, this initial assumption seems like one the extended theorists, not to mention the anti-extensionist, might understandable reject.

## Broader context

The broader context for Chapter 5 is the growing literature around predictive processing and 'radical' views of cognition, particularly the so-called '4Es' (see, e.g., Friston, 2011; Clark, 2015). In its own modest way, Chapter 5 tries makes a small contribution to this literature by sounding a cautionary note about predictive processing's relationship to one of the Es: extended cognition. The chapter casts a critical eye on the handful of attempts that have been made connecting predictive processing and extended cognition, but it also makes a positive contribution in the form of proposing what it would take to establish extended cognition via predictive processing.

Moreover, it might seem surprising that more attention has not been devoted to analysing the relationship between predictive processing and extended cognition given that the considerations for and against extended cognition are, for the most part, distinct from those concerning embodied and enactivist views. For example, questions about whether or not the body acts as a constraint or regulator on cognitive processing, considerations central to embodied cognition, are largely secondary to discussions of extended cognition; one might accept both points, for example, and yet still wonder if cognition is extended (see, e.g., Wilson and Foglia, 2015). Since the thinking that cuts in favour of extended cognition might differ substantively from the kind that proves helpful to embodied cognition or enactivism, a sustained, separate treatment seems necessary.

# Chapter 5 – How to Ride the Waves: Predictive Processing and Extended Cognition

## 5.1 Introduction

Predictive processing (henceforth PP) seems to be everywhere these days. From accounts of vision (Hosoya, Baccus, and Meister, 2005) and attention (Clark, 2016b) to consciousness (Wiesse, 2018) and imagination (Kirchhoff, 2017), PP has been put to work on almost every front. Yet the story is always the same: cognitive and perceptual processes are fundamentally engaged in a bid to try to reduce the mismatch between incoming sensory signals and prior expectations of the world, what is known as *predication error minimisation*.

Recently, some have started to explore connections between PP and more 'radical' views of cognition, such as enactivism or embodied cognition (Friston, 2011; Clark, 2015). Two proposals, in particular, have been recently offered connecting PP and extended cognition, the view that cognitive systems and processes sometimes extend beyond the boundary of the individual (Chalmers & Clark, 1998; Wilson, 2004; Clark, 2008; Chemero, 2009).

The first, articulated by Clark (2016a, 2017a), says that cognitive systems are extended insofar as the brain creates assemblies of neural, bodily and environmental elements in the service of minimising prediction error. On this proposal, extended cognitive systems follow from the brain's need to offload computational work via iterative, dynamic perception-action cycles. The second, offered by Ramstead et al. (2019a), says that cognitive systems are extended in virtue of the fact that the boundaries of cognitive systems are delineated by Markov blankets. On this proposal, extended cognitive systems follow from the fluid and negotiable boundaries revealed by Markov blankets.

The goal of this paper is to sound a cautionary note about these proposals. I argue that neither proposal provides a conclusive argument for extended cognition. Neither proposal provides anti-extensionists (those opposed to the idea of cognitive extension) with independent, well-motivated reasons for adopting extended cognition. Rather, the proposals, in their current state, rely on existing externalist assumptions that the anti-extensionist might reasonably reject. To be clear, the argument is not that PP makes no contribution or is completely orthogonal to discussions of extended cognition but, rather, that it is not *decisive* in settling the question of cognitive extension one way or the other.

I begin by laying out Clark (2016a, 2017a) and Ramstead et al.'s (2019a) proposals, alongside identifying their key features (Section 2). I then argue that both proposals come up short as accounts of extended cognition, as both remain committed to 'first' and 'second' wave approaches to cognitive extension (Section 3). This, I suggest, leads to a dilemma, either the two proposals: (i) fail to conclusively establish extended cognition or (ii) they function as demonstrations of a conceptual compatibility between PP and extended cognition, in which case they do not to provide any novel insight into extended cognition (Section 4). Finally, I conclude by drawing two lessons and suggesting what it would take for PP to conclusively establish extended cognition (Section 5).

There are three brief qualifications to make before proceeding. First, I am not concerned here with questions of representation. While there has been discussion elsewhere about how a notion of representation fits within the PP framework, I do not take stance on such matters here (Gładziejewski, 2015). Second, I do not make any claims about the relation between PP and other radical views of cognition, such as enactivism or embodied cognition. I focus here exclusively on the relation between PP and extended views of cognition (see, e.g., Clark, 2015; Allen and Friston, 2016; Hutto and Myin, 2017). Finally, my concern here is with *subpersonal* rather than *personal* level views of extended cognition. While some have recently argued for cognitive extension at the level of consciousness, and I do mention these views in passing, I

restrict my focus in what follows to those views which operate below the level of conscious awareness (see, e.g., Kiverstein and Kirchhoff, 2019).

## 5.2 Predictive Processing and Extended Cognition

In this section, I outline two recent proposals for how to connect PP and extended cognition.

### 5.2.1 Predictive Processing and TECs

The first proposal comes from Andy Clark (2016a, 2017a). Clark's proposal is formed largely as a response to Hohwy's (2013, 2016) 'neurocentric' vision of PP, so I spend a bit of time unpacking Hohwy's position first.

According to Hohwy, the brain/central nervous system is constantly trying to provide its best 'hypotheses' for incoming sensory data. In line with the standard PP story, Hohwy points out that there are two general ways it can do this. One is to update its expectations about the world (change its internal predictive or *generative* model); the other is act on the world so as to bring it in line with its expectation (*active inference*). As Hohwy interprets the two strategies, the former reduces the mismatch between prior expectations and incoming sensory signals, while the latter reduces the difference between the predicted and incoming proprioceptive signals. If Hohwy is right in his interpretation, then the brain is trying to constantly infer the shape and structure of a distal realm on the basis of partial and fragmentary information. The brain's access to the world is limited by the flow of information from an individual's sensory surfaces, whether that's visual, auditory or motor. Prediction error minimisation takes place behind what Hohwy calls an 'evidentiary boundary'.[69]

In response, Clark (2017a) suggests that Hohwy is ambiguous in how he uses the term 'inference'. One way to understand inference, he suggests, is as a *reconstructive* process. On

[69] Hohwy (2013, 2016) has two other arguments that Clark (2017a) criticises, one dealing with 'global scepticism', and another dealing with 'explanatory-evidential circle'. However, the inference argument is the most pressing, I think, and also forms the jumping off point for Clark's own proposal, so I focus on it here.

this interpretation, the brain is trying to piece together an internal model of the world based on impoverished incoming sensory data. This, Clark suggests, is what gives rise to the impression that the brain is secluded from the world, one has to infer the world from limited sensory input. However, another way to interpret inference is as a process that enables different strategies for co-ordinating behaviour. On this *non-reconstructive* interpretation, long term prediction error minimisation is actually an action-involving process. It is often simpler and more efficient to enact certain behaviours and shape the sensory flow than it is to try and create costly internal representations. As Clark (2017a) puts it: "The task of PP systems is not to infer the best description of the world given the sensory evidence. The fundamental task, using prediction errors as the lever, is to find the neuronal activity patterns that most successfully accommodate (in action, and in readiness for action) current sensory states" (p.734). Talk of inferring the right 'hypotheses' as Hohwy does is unnecessarily reconstructivist. An equally viable interpretation, one compatible with the PP story, is to view inference as a way of structuring on-going interactions with the world.

This idea of structuring on-going, fluid interactions with the world in order to minimise prediction error is ultimately what leads Clark to his positive proposal. Organisms are often engaged in iterative, dynamic perception-action cycles with their environment in order to offload computational work, and so minimise prediction error more cost effectively. Clark points to several example cases to illustrate.[70] For simplicity, I focus here on just one.

Consider, then, a staple of Clark's writings, the outfielder problem (see, e.g., Clark 2003, 2008, 2016). The outfielder problem asks us how a baseball player (an outfielder) is able to catch a fly ball while on the run. One suggestion, in line with reconstructive views, is to say that they engage internal replica building, estimating and tracking the ball's trajectory using rich, internal representations. However, an alternative, more cost efficient method, one in line with non-reconstructive views, is to say that they keep the image of the ball stationary on their

[70] Clark also appeals to Pezzulo, Rigoli, and Chersi's (2013) work on 'Mixed Instrumental Controller'.

retina in order to keep the flow of sensory information within a certain range. On this second approach, behavioural success is achieved not by constructing costly inner replicas of the ball's movements but simply by maintaining certain invariant relations between the organism and the world.

For Clark, the key insight from such examples is that organisms often use low-cost, action-driven strategies to minimise prediction error. This results in the creation of what he dubs 'transient extended cognitive systems' (TECs) (Clark 2008; Wilson and Clark 2009; Clark 2016a, Clark 2017a). Organisms often create temporary neural-body-world ensembles, driven by action-perception cycles, to solve context-dependent problems, such as catching a fly ball. These temporary problem-solving ensembles emerge and dissolve out of local necessity. They are a function of context and estimations of uncertainty. As Clark (2017a) makes the point: "Organismically salient (high precision) prediction error may thus be the glue that, via its expressions in action, binds elements from brain, body, and world into temporary problem-solving wholes" (p.747).

There are two key ideas to Clark's proposal. The first is that in order to sculpt the flow of information for real time, on-going use, organisms often engage in bouts of action-perception cycles. There are circular, casual interactions between organisms and their environment in order to offload computational work from the brain to the non-neural body and world. These create the 'motor-informational weave', as Clark calls it, necessary to sustain TECs.

The second is that many of the same basic rules and principles that govern inner neural coalitions, such as efficacy and efficiency, also govern TECs – Clark points to 'Optical Acceleration Cancellation' (the strategy of keeping the ball on the retina) as an example of how systems trade-off efficacy and efficiency. There is no real difference, according to Clark, in how coalitions of internal neural assemblies and world-spanning ones select actions to solve specific problems. Crucial to both is the ability to make task-specific information available for fast, fluid use. As Clark (2016a) puts it: "The formation and dissolution of extended (brain-body-world) problem solving assembles here obeys many of the same basic rules and

principles…as does the recruitment of temporary inner coalitions bound by effective connectivity" (p.261).

What these ideas suggest to Clark is that PP specifies the computational story behind why TECs are formed: they are one strategy amongst others for minimising prediction error. Sometimes organisms can get by with purely internal neural assemblies, and sometimes they need to exploit external non-biological resources to form larger problem-solving wholes. Extended cognitive systems follow quite naturally from several basic features of the PP framework.

### 5.2.2 Knitting Extended Markov Blankets

The second proposal comes from Ramstead et al. (2019a). Ramstead et al.'s aim is to show that there is no unique or privileged boundary to cognitive systems.[71] Cognitive systems, particularly the ones that prove interesting to cognitive science, are said to have multiple and nested boundaries; boundaries which are relative to the explanatory interest of researchers. They advocate for a form of what they call *ontological* and *methodological* pluralism about cognitive boundaries.[72]

To tackle the boundary question, Ramstead et al. deploy the concept of a *Markov blanket.* Simply put, a Markov blanket is a formal, statistical tool that describes the probabilistic relations between a set of random variables. For a random set of variables, $S = \{X_1 \ldots X_n\}$, the Markov blanket of a random variable Y is any subset $S_1$ that is conditioned on other variables that are independent of Y. For example, in a Bayesian network, the boundary of some node A includes all the 'parents' and 'children' nodes of A (those conditionally dependent on A), and the other parents of all of its children. Any nodes not conditionally dependent on A or its parents or children are outside the boundary of the Markov blanket (Pearl, 1988).

[71] Ramstead et al. label such views 'essentialist'.
[72] In passing Ramstead et al. are also interested in debunking Hohwy's (2016) 'neurocentric' view of PP. However, since I have discussed Hohwy's view in relation to Clark's proposal, I leave this discussion to one side and focus on Ramstead et al.'s boarder motivation.

In Ramstead et al.'s (2019a) hands, Markov blankets are said to carve out a system's systemic states (internal states) and non-systemic states (external states). They are said to provide one way of formalising the intuition that for something to exist it must be separate from the system in which it is embedded.[73]

Ramstead et al. (2019a) think that Markov blankets reveal a novel route to cognitive extension.[74] This is because the hierarchically nested and fluid nature of Markov blankets are said to open up the possibility of investigating cognitive systems that extend into their environment at different temporal and spatial scales. Ramstead et al. (2019a) write, for instance: "The Markov blanket formalism might allow us to study the transient assembly of cognitive boundaries over time…The variational framework, then, might allow us to model how organisms extend their Markov blankets into the environment, at a host of different spatial and temporal scales."

There are two key ideas to Ramstead et al.'s proposal. The first, as mentioned, is the concept of a Markov blanket. A Markov blanket, as previously stated, consists of probabilistic relations between a set of random variables. As Ramstead et al.'s (2019a) interpret them, Markov blankets carve out a system's systemic states (internal states) and non-systemic states (external states). Between the internal and external states are what are called the *active* and *sensory* states – the blanket itself consists of these intervening states. The active and sensory states form the intermediary links through which the internal and external states interact. The internal states influence the external states via the active states, while the external states influence the internal states via the sensory states. The internal and external states relate via 'conditional independencies'.[75]

---

[73] It is worth noting that there is nothing about systems or their internal or external states *per se* in the definition of a Markov blanket. Rather, it is merely about relationships between random variables. Using Markov blankets to delineate boundaries of systems is something that Ramstead et al. add to the definition. This was pointed out helpfully by Mark Sprevak (personal communication).

[74] This argument is inspired by Clark (2017b).

[75] Informally, conditional independence can be defined as follows: for some variable A, A is conditionally independent of B, given some variable, C, if and only if knowing A provides no additional information about B given C.

More importantly, because there is a continuous and reciprocal interaction between the internal and external states via the active and sensory states, the active and sensory states help to maintain the structural and functional integrity of the Markov blankets. Because systems are continuously trying to minimise (on average) their free energy, the active and sensory states construct and maintain an existential/evidential boundary with the environment.[76] A cell, for example, creates a semi-permeable membrane out of its surrounding molecular soup in order to regulate and sustain itself over time. Markov blankets are a response to trying to reduce or minimise prediction error (or variational free energy).

The second idea is that the statistical structure of Markov blankets is 'scale free'. This means that, in principle, Markov blankets can hold at any number of temporal and spatial scales, e.g., from macromolecules and organelles to organs and humans. There is no *one* Markov blanket for a system, but an overlapping, nested set of blankets. Ramstead et al. label this *multiscale integration*, writing of it: "[A]ll the subsystems that are individuated by their own Markov blanket are integrated as one single dynamical system through the system dynamics (i.e., adaptive action)". Successively larger and slower scale dynamics arise from, and are constrained by, the dynamics of smaller and faster scales.

What is particularly relevant for present purposes is that at higher temporal and spatial scales Ramstead et al. define the boundaries of a system in terms of 'joint phenotypes'. Joint phenotypes, according to Ramstead et al., are shared extended phenotypes – i.e. traits that enhance fitness, such as a beaver's disposition to build a dam. Joint phenotypes are coextensive traits consistent with two or more species' genetic makeup. The key point for Ramstead et al. is that joint phenotypes do not necessarily have to include a genetic component, although this is usually how they are understood. Rather, in principle, any trait that has an impact on fitness, whether it is biotic or abiotic, can be included in a joint phenotype. This means that in order to study extended organism-niche systems one need only analyse the statistical relationship

[76] I am glossing some of the technical details, but for present purposes this descriptions suffices to introduce the key elements of Markov blankets as Ramstead et al. use them.

between non-genetic traits and states of an organism using the formalism provided by Markov blankets.

One illustrative example is the case of a spider's web. A spider's web, Ramstead et al. (2019b) point out, extends an arachnid's sensory surfaces in all sorts of ways – for example, it allows the spider to detect food or threats by acting as an early detection system, or increases the spider's speed and mobility. The web is a biotic extension of the spider's body, produced and sustained over time in order to minimise long-term prediction errors. The web acts as the mediating active and sensory states. For this reason, the Markov blanket can be placed around the entire spider-web system. On this occasion, conducting extended research involves creating a generative model that simulates the effects of external factors on error minimisation (e.g., the spider-web system). Or, as Ramstead et al. (2019a) put it: "One can study organism-niche complementarity that obtains through phenotypic accommodation and niche construction over development (i.e., adaptation) using variational free energy."

What we have, then, are two proposals for how to connect PP and extended cognition. One says that PP reveals how non-neural resources are often used in real-time, dynamic processes to create extended cognitive systems; while the other says that the formal apparatus offered by Markov Blankets helps to settle thorny boundary drawing questions, and so reveals the fluid and negotiable nature of cognitive systems. Both proposals offer novel attempts at marrying various aspects of the PP story to extended cognition. However, in the next section, I raise some critical worries about the proposals.

## 5.3 Extended Waves

According to Sutton (2010), there are three overlapping but distinct 'waves' to extended cognition.[77]

The first wave says that cognitive extension follows largely from considerations of functional equivalence or 'parity' (Clark, 2003, 2008; Wilson, 2004). According to this

[77] Sutton (2010) does not explicitly address the third wave but more gestures in its direction.

approach, when external elements make similar functional contributions to behaviour as internal ones then they qualify as part of the underlying cognitive process/system. For example, in the now classic Otto-Inga case, Clark and Chalmers (1998) argue that since Otto's notebook, because he suffers from memory issues, plays an equivalent functional role in helping Otto find his way to the museum as would otherwise be played by biological memory, then it should qualify as part of Otto's extended memory system. Proponents of first-wave thinking usually employ a form of common-sense or psychofunctionalism, focusing on the functionally salient component parts or processes underlying a task or process (see, e.g., Sprevak, 2009).

The second wave moves past parity considerations and instead focuses on 'complementarity' (Sutton, 2010). According to this approach, internal and external elements often combine to form self-sustaining systems with new sets of functional properties. For example, as Wilson (2004) points out, many notational systems often help to augment and transform a user's mathematical reasoning capacities in ways otherwise unavailable to purely neuronal resources. In such cases, a new system emerges, one with functional properties over and above the unaugmented, purely brain-based systems. Second-wave thinking focuses on the complementary nature of external tools and technologies in driving cognitive extension.

Finally, the third wave, only recently articulated, builds on the second wave, in that it emphasises complementarity, but it also adds further notions of reciprocal, on-going dynamic causal flow and diachronic constitution. The key insight for third wave approaches is that extended systems follow not only from the transformative power of external resources, such as in the case of mathematical reasoning, but also from the way in which those resources distribute and coordinate behaviour in ongoing, dynamic ways over time.

Kirchhoff and Kiverstein (2019) lay out what they take to be four key 'tenets' of the third wave:

1. *Extended Dynamic Singularities*: Cognitive processes are constituted by causal networks with internal and external nodes comprising singular cognitive system.
2. *Flexible and Open Ended Boundaries*: The boundaries of mind are not fixed and stable but fragile and hard-won, and always up for negotiation.
3. *Distributed Cognitive Assembly*: The task and context-sensitive assembly of cognitive systems is driven not by the individual agent but by a nexus of constraints, some neural, some bodily, and some environmental elements (e.g., cultural, social, material).
4. *Diachronic Constitution:* Cognition is intrinsically temporal and dynamical, unfolding over different but interacting temporal scales of behavior.

For Kirchhoff and Kiverstein (2019), third wave thinking heralds a new approach to cognitive extension. While first- and second-wave approaches adopt atemporal and asynchronic notions of realisation and constitution, the third wave reconceptualises the underlying metaphysics of extension, offering a new vision of extension based on temporally dynamic and synchronic notions of realisation and constitution.

The question is, which of the three waves do Clark and Ramstead et al.'s proposals belong to? At first blush it might seem plausible to suggest they are instances of third-wave thinking. Both proposals, after all, seem to make much of the dynamic, on-going role of action and perception in free energy minimisation, and both proposals seem to suggest that the boundaries of cognitive systems are flexible and open-ended. However, what I want to suggest is that both proposals, in fact, reflect instances of first and second wave thinking.

Consider, first, Clark's proposal. While Clark's proposal makes much of the dynamic, on-going causal flow of action and perception in error minimisation, the significance of such loops largely rests in how they help to sustain and regulate the various neural and non-neural assemblies. While the causal dynamics are important in Clark's proposal what seems to be ultimately driving extension is the functional salience of non-neural resources in creating temporary problem solving assemblies.

Recall, for example, that in the outfielder case Clark saw no important difference between the strategies employed by neural and non-neural coalitions. Both assemblies were engaged in error minimisation, and both were formed in ways constrained by efficacy and efficiency – for example, the Optical Acceleration Cancellation strategy. What matters is not the location or composition of non-neuronal resources but their functional role, particularly with respect to how they help sustain and regulate different cost-cutting computational strategies over time.

What this shows, I think, is that so long as the non-neural resources occupy the right functional role within a wider task economy, such as in the case of trying to catch a fly ball, then they qualify as part of transient extended cognitive systems for Clark. He writes, for instance: "[T]he recruitment of task-specific inner neural coalitions within an interaction dominated PP economy is entirely on par with the recruitment of task-specific neural-bodily-worldly ensembles (2016, p.261). This seems to be a case of first wave thinking. The computational story might be elaborated in all sorts of interesting ways by PP, but the basic driving force behind extension remains, at core, considerations of functional equivalence or parity.[78]

Consider, next, Ramstead et al.'s proposal. What generates extension in Ramstead et al.'s proposal are not Markov blankets per se but, rather, how the Markov blankets apply to joint phenotypes at higher spatial and temporal scales. Ramstead et al. (2019a) write, for instance: "[T]he point we want to motivate here is that—especially in humans—many traits of the constructed niche defining the human joint phenotype increase state-trait complementarity by smoothing the attunement process, or variational free energy minimising process." The driving

[78] There is a potential second wave gloss on Clark's argument. That is, rather than seeing the embodied, embedded process responsible for ball catching as an external process that is functionally equivalent to internal processes, one might instead view the process as an external complementary process that functionally augments internal resources. On this reading, the action-oriented, embodied strategy is an external process that helps creates a system with functional abilities over and above what the internal resources could achieve on their own. However, as I already examine Ramstead et al. (2019a)'s proposal as an example of a second wave argument, I leave such interpretative questions to one side for the moment.

force behind extension is not the Markov blanket formalism but how that formalism is applied to external tools and technologies on the basis of complementarity during developmental or perceptual processes.

Recall, for example, the spider case. What allowed for extension in the spider case, according to Ramstead et al. (2019b), was not only the scale-free nature of Markov blankets but, also, the complementary role played by the web in enhancing the spider's fitness, and by extension, its perceptual abilities. As an external artefact, the web extended the spider perceptual reach beyond what it would otherwise be, e.g., by enhancing its ability to detect food or external threats.

What this shows, I think, is that the web's status an extension tool rests largely with its role as a complementary external biological resource. The Markov blanket applies to the entire organism-niche system, and not just some subpart, because the web functions as the system's intermediary active and sensory states. This is what grounds treatment of the entire organism-niche complement as warranting a Markov blanket. Again, the computational story might be neatly formalised by Markov blankets, but the basic driving force behind extension is still second wave considerations of complementarity.

However, one might remain convinced that Clark and Ramstead et al.'s proposals are better thought of as instances of third wave thinking. To see why not, recall again Kirchhoff and Kiverstein's (2019) key 'tenets' of the third wave thinking.

First, notice that talk of dynamic causal flow and reciprocal feedback is already part and parcel of first wave thinking. If Otto's notebook is not tightly paired via action perception cycles with the rest of his cognitive apparatus, then it fails to occupy the right functional role. The causal looping between organism and environment is what grounds claims about functional equivalence. Talk of dynamic singularities is not unique to third wave thinking. There is no reason to assume that because Clark's or Ramstead et al.'s proposal invoke such a notion that they therefore qualify as instances of third wave thinking.

Second, note that both Clark and Ramstead et al.'s proposal might embrace the idea that the boundaries of cognitive systems are fluid and negotiable, but this tenet, strictly speaking, is a feature of any view on extended cognition. First and second wave approaches also accept the idea that cognitive boundaries are not fixed. This is a conclusion but not a premise of extended thinking in general. So again, there seems to be no reason to assume that a commitment to the open ended boundaries of cognitive systems uniquely identifies Clark and Ramstead et al.'s proposals as cases of third wave thinking.

Third, notice that the concept of disturbed cognitive assemblies also falls out of first and second wave approaches. Both approaches argue for the possibility of distributed cognitive assemblies, only they do so via different considerations than third wave thinking: namely, considerations of complementarity and functional equivalence. Talking about distributed cognitive assemblies, again, does not single Clark or Ramstead et al.'s proposals out as instances of third wave thinking. Non-PP frameworks, such as functionalism, can also accommodate this idea (see, e.g., Wilson, 2004; Clark, 2008).

Finally, if constitution is diachronic as Kirchhoff and Kiverstein suggest, then this would form an independent and novel basis for cognitive extension. Assuming that cognitive systems are realised by complex, temporally extended causal dynamics, then such systems may, on occasion, extended into the surrounding environment (see, e.g., Kirchhoff, 2015; Kirchhoff and Kiverstein, 2019).[79] But again, neither Clark nor Ramstead et al.'s proposals rely on such considerations to motivate their extension claims. While both sets of authors acknowledge the importance of dynamic causal looping and information flow in their accounts, neither takes such facts to require a rethink of the underlying metaphysics of realisation or constitution. There is no suggestion that Markov blankets or active inference require diachronic realisation or constitution to function. In short, then, Clark and Ramstead et al.'s proposals are not

[79] See Hurley (1998) and Wilson (2004, ch.6) for earlier expressions of this idea.

instances of the third-wave thinking, because either they do not rely on the tenet or the tenet is one shared by first and second wave approaches.

## 5.4 Crashing Waves

I have argued so far that Clark and Ramstead et al.'s proposals, despite some appearances to the contrary, are instances of first and second wave approaches to cognitive extension respectively. Given this, in this next section, I raise concerns about the value of these proposals. I do so in the form of a dilemma.

First, notice that if Clark and Ramstead et al.'s proposals are interpreted as *arguments* for cognitive extension, then they to fail to conclusively establish extended cognition. What I mean by this is that if Clark and Ramstead et al.'s proposals are interpreted as first and second wave arguments for cognitive extension, then they seem to be vulnerable to a number of existing anti-extensionist objections. To be clear, the claim is not that PP makes no contribution to discussions of extended cognition. Rather, the claim is that PP is not *decisive* in settling the question of cognitive extension one way or the other, relying instead on existing externalist assumptions that the anti-extensionist might reasonably reject.

For example, according to Adams and Aizawa's (2001, 2008) 'coupling-constitution' fallacy, extended theorists often make an implicit move from causal coupling (or some form of dynamic entanglement) claims to claims about constitution. After citing examples of how cognitive processes are causally coupled to elements in the world, extended theorists then shift to claims about how such external elements constitute part of the cognitive process itself. Clark's proposal seems to reflect such thinking. It moves from claims about the tight causal coupling via action perception cycles of organism and world during prediction error minimisation to claims about the formation of transient extended cognitive systems. However, as Adams and Aizawa point out, there seem to be a number of homely examples, such as air-conditioning systems, where we do not concede the existence of a new, hybrid system based on causal-coupling considerations.

To be clear, I am not suggesting that Adams and Aizawa's are right in their critique. Rather, I am pointing out that functionalist-inspired parity arguments, ones based on coupling considerations, arguably never overcame traditional anti-extensionist objections, such as the coupling-constitution fallacy, and so it is unclear why a PP-inspired version of the argument should fare any better. The anti-extensionist objections, such as the coupling-constitution fallacy or the mark of the cognitive, seem to apply equally to PP-inspired accounts of extended cognition as they do traditional functionalist-based accounts, such as ones offered by Clark and Chalmers (1998) or Clark (2008).

Second wave complementarity arguments, such as the one offered by Ramstead et al. (2019a), seem to fare little better. As Rupert (2004) and Wilson (2002) point out, insofar as researchers are interested in making empirically testable projections across a variety of contexts, it seems dubious that a science of the mind can be constructed around systems that involve fast-dissolving assemblies. The social, cultural and material tools used by cognisers in niche-construction, for example, seem too transient to form a stable scientific kind for sustained investigation. Again, the point is not that Rupert or Wilson are right in their critique, but that PP-based versions of second wave arguments do not, by themselves, surmount such worries.

The more general point, then, is that there seems to be a dialectic question mark looming over PP-inspired arguments for extension. PP-inspired accounts seem to be at argumentative stand-still with respect to anti-extensionist objections. Despite being couched in the PP framework, the argumentative heavy lifting of Clark and Ramstead et al.'s proposals is still being done by traditional first and second wave parity and complementary considerations. There seems to be little reason why an anti-extensionist should be any more moved by these arguments. As stated, Clark and Ramstead et al.'s arguments do not surmount or vault the outstanding anti-extensionist worries, so much as bury them under a wealth of PP detail.

This is particularly problematic when it is noticed that, some select comments of Clark's notwithstanding, both sets of authors, in fact, see themselves as furthering the case for extended cognition. Clark (2017a), for instance, writes:

> This [PP] neatly accommodates frugal 'sensing-for-coupling'-style solutions of the kind celebrated by work in ecological psychology. But better still, it accommodates those solutions within the systematic and empowering context of a fluid, re-configurable economy in which the use of rich, knowledge-based strategies and the use of fast, frugal procedures are merely different expressions of a common uncertainty-estimating mechanism. (p.748).

Not only does PP accommodate insights near and dear to the hearts of embodied and enactivist theorists, but it also weaves together these insights into something resembling extended cognition.

Ramstead et al. (2019a), on the other hand, summarising the insights of their approach, write: "This [approach] speaks to the necessity of methodological pluralism in cognitive science; and to the importance of developing new interdisciplinary research heuristics to determine and study, for any phenomenon, the relevant levels of description that are necessary to account for it." For Ramstead et al., Markov blankets reveal not only an ontological pluralism about cognitive systems but also a wider methodological pluralism that should be taken up in cognitive science. In both cases, PP is said to provide something novel and interesting to discussions of extended cognition.

Second, notice that if the two proposals are not interpreted as *arguments* for extended cognition but, rather, as *demonstrations* of a general conceptual compatibility between PP and extended cognition, then it is unclear why this conclusion should prove particularly surprising. As Wilson and Clark (2009) already point out: "Suppose we grant the assumption of computationalism that has structured much of the work in cognitive science. If the kind of computation that at least parts of cognition involve are extended, then those parts or aspects of cognition will also be extended." Far from being incompatible with computationalism, extended cognition follows from it. Because of its indifference to a system's realising components, computational analysis can always, in principle, apply to world-spanning

cognitive systems. But given that PP is a specific form of computationalism about the mind (i.e. a form of Bayesianism), it should not be particularly surprising that it is compatible with extended cognition. Thus, if the two proposals are not *arguments* for cognitive extension but *demonstrations* of a general compatibility, then it is unclear why they offer any new insights not already available to extended theorists.

So, it seems, then, that Clark and Ramstead et al.'s proposals face a dilemma, either: (i) they fail to conclusively establish extended cognition or (ii) they function merely as demonstrations of a general conceptual compatibility between PP and extended cognition, in which case they do not offer something fundamentally new to the existing debate about extended cognition. The first option leaves PP-inspired accounts at an argumentative stand-still; while the second option reduces PP-inspired accounts to insights already available to extended theorists. There seem to be some serious questions looming as to the value of the two proposals.

That said, this does not mean that PP offers no insights whatsoever. PP does, after all, help to elaborate the computational story underlying extended cognition. It helps explain what drives the creation and dissolution of extended cognitive systems: namely, prediction error minimisation; and this is undoubtedly an improvement over previous accounts of extended cognition which leave such questions blank. As Clark (2016a) puts it, PP offers a "highly 'extension-friendly' proposal concerning the shape of the specifically neural contribution to cognitive processes" (p.260). But what PP does not do, at core, is change the arguments for or against extended cognition. [80]

## 5.5 How to Ride the Waves

There are two lessons that I think can be drawn from the preceding discussion.

[80] To his credit, Clark does acknowledge this at one point, writing: "Nothing the PP framework materially alters, as far as I can tell, the arguments previously presented both pro and con, regarding the possibility of genuinely extended cognitive systems" (2016a, p.260).

### 5.5.1 Lesson #1

The first is that Markov blankets do not delineate the boundaries of cognitive systems in the way Ramstead et al. (2019a) propose. Markov blankets, recall, are a purely statistical formalism. They describe probabilistic relations between random sets of variables. There are two ways to interpret what this means. One is in terms of 'subjective' Bayesian networks, where the probabilistic relations are defined over a set of variables in terms of estimations of uncertainty or prior probabilities. On this interpretation, the probability distributions range over an agent's prior beliefs. The other is in terms of 'objective' Bayesian networks, where the probabilistic relations are defined over random sets of variables in terms of their physical magnitudes. On this interpretation, the probability distributions range over sets of states in the world.[81]

Notice that Ramstead et al.'s proposal equivocates between these two senses. Return to the spider case to see why. On the one hand, talk of the spider reducing *its* prediction error by updating its priors using the web invokes the subjective sense of Bayesian networks. The probability distributions relevant to defining the Markov blanket are cast in terms of the agent's estimations of uncertainty about the world. On the other hand, the Markov blanket is also defined in terms of the conditional dependencies between internal and external states. The probabilistic relations are defined over physical variables in the world: namely, the relations between the spider, web, and its surrounding environment. While talk of the web enhancing the spider's perceptual capacities invokes the first sense, talk of the conditional dependencies between states of the spider and the world invokes the second.

The trouble is that this slippage gives the misleading impression that because the web helps to reduce uncertainty in the first sense (i.e. estimations of uncertainty), and there are probabilistic relations obtaining between the web and the world in the second sense (i.e. probabilistic relations between states of the world), that the boundary of the system can be

[81] This distinction was helpful pointed out to me by Mark Sprevak in personal communication.

drawn around the entire organism-niche complement. But this simply does not follow. There are two qualitatively distinct types of Markov blankets one can draw: one for the subjective random variables and one for the objective random variables. The Markov blankets can be defined in terms of both, but they should not be conflated.

Worse still, settling on one or the other interpretation does not help Ramstead et al.'s proposal. If the subjective interpretation is adopted, then the question of where to draw the cognitive boundaries remains open. If the agent's estimations of uncertainty are defined over their prior beliefs, then the question still remains as to where these prior beliefs are located. Are they internal or external to the agent?

If the objective interpretation is adopted, then the formalisation is trivial. For any random set of variables, there will always be *some* variables that are probabilistically independent of one another, whether it is particles in a cloud or neurons in the brain. Markov blankets formalise these relations. But simply because certain variables are probabilistically independent does not mean that the formalising relations reveal the *principled* boundaries of a system. Markov blankets can formalise principled and arbitrary boundaries alike. For example, just because neurons in the fusiform gyrus are conditionally independent of neurons in the pre-frontal cortex does not mean that there is a functionally or anatomically important boundary between the two. There will often be many probabilistic relations that do not reveal anything informative about a system.

Contrary to Ramstead et al.'s (2019a) claims, Markov blankets are not capable, even in principle, of arbitrating questions about the boundaries of cognitive systems. To do so would be to confuse properties of the model with properties of the thing being modelled. PP-inspired accounts need to avoid assuming that an appeal to Markov blankets alone settles questions of cognitive extension.

### 5.5.2 Lesson #2

The second lesson is that progress on cognitive extension may also require further work undermining some of the underlying intuitions and assumptions that drive anti-extensionist worries. That is, if the intuitions and assumptions that drive the anti-extensionist position are allowed to remain intact, then the anti-extensionist can continue to press their claims against PP-inspired accounts. As I've tried to show, PP-inspired accounts seem to be at an argumentative stand-still when framed in terms of first and second wave approaches. They do not seem to supply the anti-extensionist with novel grounds for accepting extension claims. One way to make progress in the debate, then, might be to chip away at the underlying intuitions and assumptions that sustain the anti-extensionist objections.

For example, Ross and Ladyman (2010) point out that Adams and Aizawa's alleged coupling-constitution fallacy seems to be based on a misplaced 'containment' metaphor, one that a mature metaphysics of mind will no longer support. Whether or not Ross and Ladyman are right, I leave to others to decide. The point for present purposes is that until these types of underlying intuitions and assumptions are addressed, couching extension arguments in dynamical, connectionist, or Bayesian terms may do little to advance the dialectic – other examples of such undermining work include Wilson and Clark (2009), Clark (2010), and Wheeler (2010). Moving the discussion forward about extended cognition may require excavating and undermining the underlying assumptions and intuitions that sustain the anti-extensionist worries, such as the coupling-constitution fallacy (Adams and Aizawa, 2009) or the motley-crew argument (Rupert, 2009).

### 5.5.3 Predictive Processing and Extended Cognition 2.0

More positively, here is one possibility for how PP might establish extended cognition, and by that I mean provide strong, independent grounds for accepting extended cognition. If one could show that there are certain features that are distinctive of cognitive processes that follow

directly from accepting PP, and that extended systems, as a matter of fact, possess such features, then this should suffice to show that extended systems are genuine cognitive systems. Here is a sketch of the argument, along with a brief gloss on each of its premises:

1. The distinctive processing undertaken by cognitive systems is described by PP (namely, prediction error minimisation).
2. If extended systems engage in prediction error minimisation, then they qualify as cognitive systems.
3. Extended systems engage in prediction error minimisation.
4. Therefore, some cognitive systems are extended.

Premise 1 reflects the current state of opinion among proponents of PP (e.g., Clark, 2013, 2016; Friston, 2009, 2011; Hohwy, 2013). It says that the brain is fundamentally a Bayesian prediction machine, and that cognitive systems and processes are, at core, engaged in a form of prediction error minimisation. To clarify, this is prediction error minimisation at an 'algorithmic level'. It is not simply that the system's behaviour conforms to a general description of error minimisation but, rather, that the procedures used by the system actually carry out prediction error minimisation – for example, in the case of the brain, this might involve approximating a form of Bayesian inference. To say that something is cognitive is to claim that it is engaged in a process of prediction error minimisation.

Premise 2 makes a structural point about cognitive systems. It says that any system that employs prediction error minimisation should be included as a genuine cognitive system. Notice that this not simply a restatement of parity considerations or functional equivalence. It is not a claim about the functional role played by external resources. Rather, it is claim about what functional profile cognitive systems must have. The claim is that if extended systems exhibit the requisite fine-grain functional structure distinctive of cognitive processing, then they qualify as genuine cognitive systems alongside internal ones.

Finally, premise 3 says that the empirical facts bare out the PP story about cognitive extension. It claims that extended systems, as a matter of fact, do exhibit the fine-grained functional profile required to qualify as cognitive systems. Plausibly, this is what Clark (2016, 2017a) was gesturing at when he argued that TECs followed the same rules and principles as internal neural assemblies. The assumption is that there is enough empirical evidence to support the claim that extended systems also involve algorithmic level prediction error minimisation.

Of course, one can take issue with any of the three premises. One might be sceptical, for instance, of the truth of PP, or whether the empirical case that Clark (2016, 2017a) presents really is sufficient to establish extension.[82] However, while these worries are legitimate and will need to be addressed at some point, for now I want to focus on the argument as a whole. This is because I think the argument presented provides a novel, PP-based reason for thinking that cognitive systems might be extended. I think it shows *what it would take* for PP to provide a strong argument for extended cognition.

First, notice that the argument by-passes the previous anti-extensionists worries. It does not require that one accept additional externalist assumptions, such functional equivalence or complementarity, in order to establish extension. Cognitive extension, according to this argument, follows from the truth of PP, and also an assumption about the empirical case shaking out appropriately. It outflanks the anti-extensionist worries by eschewing first and second wave considerations in favour of a direct line from PP.

Second, notice that the argument does not rely on claims about particular formalisms to establish extension, such as Markov blankets. The truth of extended cognition depends the truth of PP. It thereby avoids the previous trap encountered by Ramstead et al.'s (2019a) proposal.

[82] To be fair, PP does have a fair amount of weight behind it, including work on retinal predictive coding (Hosoya, Baccus, and Meister, 2005), binocular rivalry (Hohwy, Rospstorff, and Friston, 2008), perceptual completion (Raman and Sarkar, 2016) repetition suppression (Summerfield et al., 2009), and multisensory (Talsma, 2015), to name only a few areas.

Finally, unlike Kirchhoff and Kiverstein's (2019) view, the argument does not require a rethink of the entire underlying metaphysics of mind in order to establish extension. One does not have to buy diachronic notions of realisation or constitution to accept the argument. Rather, all one needs to accept is that PP reveals the distinctive feature of cognition (predictive error minimisation), and that extended cognitive systems might possess such a feature.[83] Insofar as a theoretical coherence is an explanatory virtue, the argument seems to have the advantage of requiring minimal distribution to our wider network of beliefs about the mind and cognition.

In short, then, what the argument shows, I think, is what it would take for PP to provide a strong argument for extended cognition. While it is by no means decisive or uncontroversial, it reveals one possible, and I think plausible, route to cognitive extension. It shows how PP and extended cognition might be connected without relying on first- or second-wave style arguments and also while doing justice to the insights delivered by PP.

## 5.6 Conclusion

Where does this leave us going forward? What it suggests, I think, is that in its current state PP does not settle the internalist/externalist debate. But this does not mean that it cannot play an important role if suitably developed. What I have tried to show is not only the limits of existing PP-inspired accounts of extended cognition but also what it would take to make progress on extended cognition debates using PP. I have sketched a general argument for how PP might move the needle forward in discussions of extended cognition. This requires, I suggest, showing that there are distinctive features of cognitive processes revealed by PP, and that such features can be found in extended systems. The argument requires further elaboration and defense, but hopefully it shows how extended cognition might ride the PP wave going forward.

[83] In this way, the argument is reminiscent of some 'mark of cognitive' style arguments for cognitive extension (e.g. Wheeler, 2010).

# Introduction to Chapter 6

## Key ideas

The key idea of the Chapter 6 is that there is a particular conception of levels that, often implicitly, exerts influence over how some in cognitive science approach thinking about and studying cognition, what I label the 'Hierarchical Correspondence View of Levels'. The HCL is a particular view about how different ideas about 'levels' relate. It not a conception *of* levels, in the sense that 'levels of analysis' or 'levels of organisation' are view of levels. Rather, it is a conception of how different notions of levels fit together, such as 'levels of analyses' or 'levels of organisation'. To be specific, the HCL is a combination of three distinct ideas about levels.

First, the HCL maintains that the world is organised into a hierarchy of ontological levels. This is the rather intuitive, but often implicit, idea that the world is arrayed into a hierarchy of organised levels. For example, on Oppenheim and Putnam's (1958) famous layer cake model, higher order entities, such as cells, are assembled or composed out of lower order entities, such as molecules, via part-whole relations.

Second, the HCL holds that there are specific modes of analysis for describing the ontological levels. For example, according to Marr's classic tripartite framework, cognitive investigations can be separated into three distinct modes of analysis: the computational level, the algorithmic level, and the implementational level. For the HCL, there are different vocabularies or languages for analysing various types of phenomenon, such as cognition, and these can be arranged into a mutually informing hierarchy.

Third, the HCL claims that there is a specific relationship between the modes of analysis used to describe the world and the way the world is hierarchically organised. There is a strict correspondence between a specific mode of analysis, i.e. a descriptive vocabulary or language, and specific level of the ontological hierarchy, i.e. a set of stratified structures in the world.

## The argument

To show that the HCL has, in fact, found its way into cognitive science, I focus on two example cases. The first is Alan Newell's (1990) classic study of cognitive architecture. I argue that a pair of related concepts for Newell, the concepts of 'system levels' and 'bands', are suggestive of a tacit commitment to the HCL, and that such a commitment has implication for how Newell thinks about functional analysis and explanatory reduction. The second example comes from Ron Sun. Sun et al. (2005) set out a vision of computational cognitive modeling in which cognitive phenomena are organised and modelled into four distinct levels. Here, again, I argue that Sun et al.'s picture gives way to an underlying commitment to the HCL, and that this colours the way the authors' think about cognitive modeling and model construction.

My argument is that the HCL is deeply problematic when applied to cognitive science because it fails to appreciate an important distinction between two types of shifts: *shifts in analysis* and *shifts in grain*. Shifts in analysis address shifts in the mode of analysis or language used to describe a complex system. For instance, one might talk about a particular function, such as (car(list)), in terms of a sequence of machine language instruction (e.g., a programming language) or a local characterisation of the operation (i.e., the machine language). Shifts in grain deal with shifts in the generality of the explanations used to describe a complex system. For example, most of the functions within a low-level assembly programming language can be specified without reference to the machine language – that is, the code that directly implements actions in the computer's CPU.

I argue that the HCL fails to accommodate this distinction, instead viewing the shifts as overlapping in every case – that is, every shift in analysis involves a shift in grain. To illustrate, I draw attention to several examples of where there are shifts in grain but no shifts in analysis, and cases where there are shifts in analysis but no corresponding shifts in grain. This failure, I suggest, leads the HCL to impose two overly restrictive assumptions about cognitive investigation. On the one hand, it means that the HCL imposes an unnecessary methodological constraint on investigation. In envisaging the shift between modes of analysis as a shift between levels of organisation, the HCL restricts how many times a given mode of analysis might apply. On the other hand, it means that the HCL makes for a questionable empirical hypothesis about complex systems. The HCL assumes that complex systems will not have a nest structured within a given level of ontology.

## Broader context

Assuming this all makes sense, why do we need this investigation? And where does it fit with rest of the thesis?

With respect to the first question, one reason to pursue the inquiry is that it helps to show that conceptual metaphors, such as those about levels, are not simply explanatorily or methodologically neutral. As Lakoff, and Johnson (1980) famously point out, the conceptual metaphors we use can hold considerable sway over how and what we think. A good portion of Chapter 6 is dedicated to showing not only that the HCL exists, but that it is has made its presence felt specifically in cognitive science in virtue of shaping some key theoretical work.

A second reason is that it can potentially shed light on a number of existing debates in cognitive science. Although I do not take the point up in the chapter, I think having the HCL explicitly and precisely stated (I also provide a formalisation of the view) helps to mediate some interpretative debates in cognitive science.

For example, consider briefly a relatively recent issue of *TopiCS* in which Marr's (1982) seminal 'tripartite' view of levels featured as the focus. Several additions to the issue had some

choice criticisms of Marr's view. French and Thomas (2015), in particular, worry about the lack of a concept of emergence, writing:

> But Marr's levels are—and were meant to be—descriptive, rather than interactive and dynamic. For this reason, we suggest that, had Marr been writing today, he might well have gone even farther in his analysis. He would, we believe, have included a discussion of the following issues… (c) the emergence of structure—in particular, explicit structure at the conceptual level—from lower levels, and (d) the effect of explicit emergent structures on the level (or levels) that gave rise to them. (p.207).

For French and Thomas, a concept of how dynamic interactions emerge across levels is missing from Marr's account.[84]

But notice that this type of criticism really only makes sense when framed against the backdrop of the HCL. If Marr's view is interpreted not only as an explanatory idea about the modes of analysis one can use to describe cognition but also as an ontological thesis about how cognitive phenomena are structured in the world, then the account's failure to include a concept of emergence begins to make sense. Marr's account is lacking, because his descriptive modes fail to capture the appropriate ontological structure in the world. In other words, there is a particular mapping relation between levels of ontology and modes of analysis that Marr's account fails to respect. Details aside, the point for present purposes is that interpretative debates about levels, such as Marr's, are brought, at least to some degree, into sharper focus by having the HCL ready to hand.

With respect to the second question, Chapter 5 stands as a cautionary tale. It advises against a particular way of interpreting talk of computation. As Sun et al.'s (2005) proposal suggests, there is sometimes a tendency to view computation as residing at a 'level' below that of 'intentional' mental states but above that of physical entities. Partly this is encouraged by the organisation of levels of analysis into a hierarchy. The intuitive thought is that because 'levels of analysis' are hierarchically organised, so too must the cognitive world to which they apply. As Floridi (2008) puts the point writing about Dennett and Pylyshyn's views: "the common

[84] A similar sentiment is echoed in Love (2015).

reason seems to be: this is the right level of analysis because that is the right LoO [level of organisation]".

The trouble, as we will see, is that not only is this a misleading way to frame talk of levels but it also fails to capture something important about the type of explanatory practices we use to study complex systems in cognitive science: namely, the role of functional contextualisations. One way to see Chapter 6's place within the thesis, then, is as an argument against an altogether not uncommon interpretation of computation, one that should be avoided where possible. It is pre-emptive move against thinking of computation as a referentially exclusive mode of analysis.

# Chapter 6 – The Hierarchical Correspondence View of Levels: A Case Study in Cognitive Science

## 6.1 Introduction

There is a general conception of levels in philosophy which says that the world is arrayed into a hierarchy of levels, and that there are different modes of analysis that correspond to each level of this hierarchy, what I label the 'Hierarchical Correspondence View of Levels" (or HCL for short).[85] The HCL is a combination of three distinct ideas about levels: (i) that the world is organised into a hierarchy of ontological levels; (ii) that there are specific modes of analysis for describing these levels; and (iii) that a specific relationship obtains between the modes of analysis used to describe the world and the way the world is hierarchically organised. The HCL is a specific conception of how various ideas about levels fit together.

Oppenheim and Putnam's (1958) 'layer cake model' offers the classic example. According to the layer cake model, there is a neat mapping between a given 'level' of science, such as chemistry, and a given 'level' of nature, such as the molecular level. The constituents comprising an organised strata of phenomena in the world are neatly mapped to the predicates and theories linked with describing those constituents. For each mode of analysis in science, there is a distinct level of ontology to which it applies.[86]

However, despite its considerable lineage and general status in philosophy of science and metaphysics, the HCL has mostly escaped analysis in specific domains of inquiry. For

[85]The view has sometimes been called the 'correspondence views of levels'. However, this doesn't capture the important role of notion of hierarchy plays in the view. What makes the view particularly attractive for many, as we'll see, is not only the correspondence relation the view posits but also the tidiness of hierarchies.
[86] For an extended analysis of the layer cake model, see Kim (2002), Baxendale (2016) or Brooks (2017).

instance, while Floridi (2008) provides an interesting and critical examination of the HCL, he does so with an eye to developing a more general concept of 'levels' for philosophy as a whole. However, as many are now concerned, there may be no such precise, general concept of levels for all of philosophy and science. The growing consensus is that any explication of the concept of levels has to be relative to a specific domain of inquiry (Brooks, 2017; Brooks & Eronen, 2018; Potochnik forthcoming).

The goal of this paper is to take up this recent call to domain-specific analysis. My aim is to examine how the HCL applies specifically within cognitive science. I aim to show, first, that the HCL is, in fact, a conception of levels that has been employed in cognitive science; and second, that cognitive scientists should avoid its use where possible. I argue that the HCL is problematic when applied to cognitive science because it fails to respect two important kinds of shifts often used when analysing information processing systems: namely, *shifts in grain* and *shifts in analysis*. I also propose a revised version of the HCL which accommodates the distinction.

I take this project to be an extension of existing attempts to clarify and refine the concept of levels. However, unlike previous investigations, such as Craver's investigation (2007), which focuses on the mechanisms in cognitive neuroscience, or Baxendale's (2016), which focuses on non-reductive physicalism in philosophy of mind, the current discussion focuses specifically on the information processing systems which form the core of cognitive science.

The paper is divided into five sections. In Section 6.2, I unpack the main three ideas of the HCL. I then formalise the view, and provide a contrast with Craver's (2007) 'levels of mechanisms' view. In Section 6.3, I examine two case studies: Newell (1990) and Sun et al. (2005). I argue that these case studies not only show that the HCL has been employed in cognitive science, but that it has also actively driven several types of explanatory inferences. Then, in Sections 6.4, I argue that the HCL fails to appreciate a distinction between *shifts in grain* and *shifts in analysis*, and that this failure stems from an inability to appreciate the important role played by functional contextualisations in cognitive analysis. The result is an

impoverished and restrictive conception of levels for cognitive science. Section 6.5 responds to possible objections to the argument. Finally, to conclude, in Section 6.6, I briefly sketch a modified version of the HCL, one which accommodates the shift distinction and makes room for the role of functional contextualisations.

## 6.2 The HCL

As mentioned, the HCL pulls together three distinct ideas about levels.

First, the HCL assumes that natural phenomena, such as cognition, are organised into a hierarchy of ontological levels. Kim (2002) provides a clear statement of the idea: "what often seems implicit is a certain overarching ontological picture of the world according to which the entities of the natural world are organised in an ascending hierarchy of levels, from lower to higher, from simpler to more complex" (p.3). The rather intuitive, but often implicit, idea is that the world is arrayed into a hierarchy of organised levels. For example, in Oppenheim and Putnam's (1958) layer cake model, higher order entities, such as cells, are assembled or composed out of lower order entities, such as molecules, via part-whole relations.

As Potochnik (forthcoming) points out, a hierarchy of ontological levels can be arranged according to several different organisational schemes. For example, one common scheme is to organise levels in terms of compositional relations. On this scheme, higher levels are a stepwise function of the compositional relations obtaining between smaller and slower entities and processes (Wimsatt, 1976). Compositional relations are the most appealed to, but others schemes have been proposed, including: part-whole relations (Oppenheim & Putnam, 1958), realisation (Fodor, 1974), and temporal scales (DiFrisco, 2017). To clarify, I am not taking a stance on the plausibility of this ontological notion of levels, I am simply laying out its basic structure in order to show how it combines with other ideas about levels.[87]

[87] For critical discussion, see Wimsatt (1994), Kim (2002), Craver (2015), Eronen (2015), Potochnik (2010), or Baxendale (2016).

Second, the HCL assumes that there are various modes of analysis one can offer about the world, which we can assign to different levels. Floridi (2008) calls this the 'epistemological' conception of levels.

For example, according to Marr's classic tripartite framework, cognitive investigations can be separated into three distinct modes of analysis: the computational level, the algorithmic level, and the implementational level.[88] Each of these three 'levels' attempt to answer a different question about cognition – for example, the computational level answers questions about a cognitive system's function and constraints on the system, while the algorithmic level answers questions about the procedures and representations by which a cognitive system achieves its function.

The answer given to each mode of analysis constrains, to a greater or lesser extent, analysis at the level below. For example, depending on the nature of the constraints and function identified at the computational level, the set of procedures and representations posited at the algorithmic level can change. If vision has the function of representing objects rather than facilitating movement, then a whole host of different representations and algorithms are needed to explain this higher-level function. According to Marr's picture, there are different vocabularies or languages for analysing cognition (e.g., computational versus physical) and these can be arranged into a mutually informing hierarchy. In other words, the HCL also assumes that different modes of analysis can be arranged into a hierarchy of explanatory levels.

Finally, the HCL makes a particular claim about the relation between the previous two notions of levels: namely, it says that for each mode of analysis there is a corresponding level of ontology. That is, there is strict correspondence between a specific mode of analysis, i.e. a descriptive vocabulary or language, and specific level of the ontological hierarchy, i.e. a set of stratified structures in the world. The claim is not that there is just *a* correspondence between hierarchically arranged modes of analysis and ontological levels. Rather, it is that for every

[88] For some, there is also an 'architectural' level that is separate from the algorithmic level, but I leave such interpretative questions to one side for now (see Dawson 1998).

mode of analysis there is a *unique* level of the ontological hierarchy to which that mode of analysis applies. The HCL assumes that there is a one-to-one mapping between mode of analysis and levels of ontology, what I call the 'strict correspondence relation'.

I hasten to add two caveats. First, the HCL is not a conception of a specific notion of levels, such as levels of organisation or levels of analysis. Rather, it is a conception of how different notions about levels relate, i.e. certain explanatory and ontological notions. One implication of this is that the HCL does not say how many levels there are, what they consist of, or what their target is. Instead, it simply spells out the relation between different notions of levels via a specific conception of correspondence and hierarchical ordering. As we will see, this means that authors can put forward a number of different ontological and explanatory levels, while nonetheless remaining committed to the HCL.

Second, the HCL is not committed to a conception of how the different notions of levels relate internally. What I mean by this is that the HCL is compatible with a range of positions on inter-level relations. For example, one can maintain that Marr's computational level is insulated from details at the algorithmic level – i.e. that it is explanatorily autonomous – while nonetheless maintaining that it addresses a distinct ontological level. As we will see, the only formal requirement of the HCL is that there is a weak mapping relation between elements of one level and those of another within the hierarchy. Such a constraint is compatible with stronger and weaker formulations, such as identity or multiple realisability (see List (2019, p.858) for discussion)

### 6.2.1 A Formal Definition

Following the framework laid out in List (2019), the HCL can be formally captured using a pair of concepts: 'systems of levels' and 'functor'. A system of levels is pair <L, S> defined as follows:

- L = a class of objects called *levels*.

- S = a class of mappings between levels, where each mapping σ has a *source level* l and a *target level* l' and is denoted σ: l → l'

While a system of levels can, in principle, capture any notion of levels, I apply the concept here to the ontological and explanatory notions of levels as employed in the HCL.[89]

First, when applied to the ontological notion of levels, levels denote a set of possible *level-specific* worlds with a *total ordering* – i.e. S is reflexive, antisymmetric, transitive and connex.[90] Each level is a set of encoded facts that uniquely obtain at a world. For example, worlds at a physical level are those that encode the totality of physical facts – some levels are 'thicker' or 'thinner' depending on the amount of facts they encode, e.g., physical versus psychological levels. The relations between level-specific worlds are captured by mappings. Each lower-level world determines facts at a corresponding higher level world – for example, lower-level chemical facts determine or settle higher-levels biological facts. For some non-empty set of level-specific worlds L ($\Omega_1 \ldots \Omega_n$), there must be a class of functions S of the form σ: Ω → Ω'.

Second, when applied to the explanatory notion of levels, levels denote a set of languages – i.e. sets of formal expressions. A sentence is true at a given mode of analysis only if the object it denotes appears in the world. For example, the truth of a sentence, Φ, of a language, ℓ, is determined at a world, ω, if Φ is member of a maximally consistent subset of ℓ to which ω corresponds; otherwise it is false. The relations between modes of analysis are also captured by mappings, only this time the elements are those of a language rather than facts of a level-specific world – for example, sentences in a physical language and sentences in an intentional language. For some non-empty class of modes of analysis L ($\ell_1 \ldots \ell_n$), there is a class of functions S of the form σ: ℓ → ℓ'. [91]

[89] List (2019) uses the framework to address several debates in philosophy, such as those about consciousness and realisation.
[90] Note that this is specific to the HCL conception of levels; List allows for partial orders too in his system of levels.
[91] The following conditions hold under this definition:

Complementing the notion of a system of levels is the concept of a 'functor'. A functor is a structure-preserving mapping between different systems of levels. A *functor*, F, from one system of levels, <L, S>, to another, <L', S'>, is a mapping which assigns to each level l in L a corresponding l' = F(l) in l', and assigns to each mapping $\sigma : l \rightarrow l'$ in S a corresponding mapping $\sigma' = F(\sigma)$ in S', where $F(\sigma) : F(l) \rightarrow F(l')$.

We can use functors to explicate the notion of a one-to-one mapping for the HCL. For each level of <L, S>, the HCL claims that there is unique level in <L,' S'> to which it is assigned. Or, to be more specific, for two systems of levels, one of which is an explanatory system of levels <L, S> and one of which is an ontological system of levels <L', S'>, there is a functor from <L, S> to <L', S'>, and <L', S'> to <L, S>. The two systems are structurally equivalent, i.e. each level and mapping of <L, S> uniquely corresponds to a level and mapping of <L', S'>, and vice versa.

Consider Marr's (1982) tripartite framework as an example. According to the HCL, Marr's computational, algorithmic and implementation levels constitute three languages for describing cognition $\{\ell_1, \ell_2, \ell_3\}$ and an ontology of levels $\{\Omega_{1...}\ \Omega_n\}$.[92] Marr's levels form a non-empty class of modes of analysis and mappings, <L, S>, and a corresponding non-empty class of levels-specific worlds and mappings, <L, S>, such that there is a mapping between each mode of analysis and mappings in <L, S> and a level-specific world and mappings in <L', S'>. So, to say that a sentence $\Phi$ is true at the computational mode of analysis, $\ell_1$, is to say that $\Phi$ denotes a set of facts or entities at a corresponding level-specific computational world, $\Omega_{L1,}$ such that $\Phi \in \ell_1$ and $\Phi \in \Omega_{L1}$. I should be clear. I'm not claiming that this is

S1) *Transitivity*: S is closed under *composition* of mappings, i.e., if S contains $\sigma : l \rightarrow l'$ and $\sigma' : l' \rightarrow l''$, then it also contains the composite mapping $\sigma' \cdot \sigma : l \rightarrow l''$ defined by first applying $\sigma$ and then applying $\sigma'$ (where composition is associative).

S2) *Reflexivity*: For each level L, there is an *identity mapping* $1_L : l \rightarrow l$ in S, such that, for every mapping $\sigma : l \rightarrow l'$ , we have $1_L \cdot \sigma = \sigma = \sigma \cdot 1_L$

S3) *Uniqueness*: For any pair of levels l and l', there is at most one mapping from l to l' in S.

[92] An 'ontology' here simply means a minimally rich set of worlds, $\Omega_L$, such that each world in $\Omega_L$ settles everything that can be expressed in L.

actually what Marr holds. Rather, I'm simply using it to illustrate how the HCL would interpret Marr's view.[93]

### 6.2.2 Levels of Mechanism and the HCL

To get a better sense of the HCL, consider how the view contrasts with Craver's (2007, 2013, 2015) view of levels. According to Craver, the concept of levels should be defined relative to mechanistic investigation. For an entity or activity to count as residing at either a higher or lower level, it must be embedded within a given mechanistic investigation.

To use Craver's favourite example, NMDA receptors, which are types of post-synaptic receptors, form one component in the mechanism for Long Term Potentiation (LTP), the process of strengthening synaptic connections in the central nervous system. The NMDA receptor (an entity) is a component part of LTP (the activity). However, LTP is also a component in the mechanism for spatial map formation in the hippocampus, which is in turn part of the mechanism for spatial navigation. NMDA receptors are part-whole related to LTP mechanisms, which are, in turn, part-whole related to the spatial maps involved in the mechanism for spatial navigation.

The key idea for Craver is that explaining phenomena such as spatial navigation requires 'topping off'. The idea is that while NMDA receptors and LTP might be part-whole related within an investigation of spatial navigation, if focus shifts to explaining a different phenomenon, such as Huntington's disease or chronic pain (two phenomena also thought to involve NMDA receptors and LTP), then the two items might form parts of different mechanistic hierarchies, ones which might span completely different spatial and temporal scales. According to Craver's picture, what 'level' a given entity or activity resides at is only relative to the mechanism it is embedded within. To be at a 'lower level' just is to be one of the components organised into a mechanism as a whole, which constitutes the 'higher level'.

[93] That said, some have occasionally interpreted Marr this way. Churchland and Sejnowski (1990), for example, once argued that "[o]nce we look at them closely, Marr's three levels of analysis and the brain's levels of organization do not appear to mesh in a very useful or satisfying manner" (p.38).

Levels of mechanisms are investigation relative, and can span several different spatial and temporal scales.

For present purposes, the important point is that whereas Craver's view assumes that there is only a partial ordering to levels, the HCL assumes there is a total ordering. The difference lies in how each characterises the relation between levels within a system of levels. According to the HCL, levels have a total ordering. They are *well defined*. This means that for any two levels one can always say that one level is higher or lower than the other. For any two levels, Ω and Ω', of a system of ontological levels, <L, S>, there is a function, S, such that either σ: Ω → Ω' or σ: Ω' → Ω (but at least one). In contrast, in a partial ordering, it could be that either Ω or Ω' are higher (but at most one). Total orderings are more restrictive subset of partial orderings. Of course, a set of levels in L could be larger than a set of L' such that the mappings between S and S' are preserved, but the important point is that the mapping relations within L and L' are in both directions (i.e. one-to-one).

What the contrast with Craver's view nicely pulls out is that the HCL makes a specific claim not only about the relation *between* systems of levels (i.e. the strict correspondence), but it also makes a specific claim about how levels relate *within* a system of levels: namely, the hierarchy is a total ordering. As opposed to Craver's view, which takes a more liberal stance on how levels are defined within a system of levels, the HCL claims that there is total ordering to levels to both the explanatory system of levels and the ontological system of levels. As we will in section 6.4, this claim about total orderings gets the HCL into trouble when it is applied to certain types of complex systems such as cognition.

## 6.3 The HCL and Cognitive Science

However, even granting the existence of the HCL as described, one might still wonder whether the view has actually found its way into cognitive science. So, to motivate this further claim, I will consider two cases studies. A case study approach is useful here because conceptions of levels such as the HCL are often hidden in the background of various domains (see, Kim 2002,

p.3). To show that a particular view is actually committed to a certain conception of levels one often needs to do a fair amount of expositional work. That said, I will focus on not showing only that certain views within cognitive science are committed to the HCL, but also that the HCL actively drives the explanatory inferences such views make.

### 6.3.1 Case Study 1

The first view I want to consider comes from Alan Newell. Newell's work is interesting because it forms not only a seminal position in the history of cognitive science, but also because it represents the type of integrative research many cognitive scientists aspire to, bringing together experimental, computational, and philosophical thinking.

In his most developed and influential work, Newell (1990) argues that the human cognitive architecture is divided into distinct functional layers, what he calls 'system levels'. These system levels reflect the substantive divisions between different sets of components and processes in cognition. Each system level is composed of components from the level below. He writes, for example: "A system level is a collection of components that are linked together in some arrangement and that interact, thus producing behaviour at that system level. In a system with multiple layers, the components at one level are realised by systems at the next level below, and so on for each successive layer" (1990, p.119). As one moves from lowest to highest organisational layers in our cognitive architecture, distinct system levels emerge, such as the 'deliberate act' or 'neural circuit' system level.

But in addition to system levels, there are also what Newell calls 'bands'.[94] Bands are the different modes for describing the various system levels, they include in ascending order: the biological band, the cognitive band, the rational band, and the social band.[95] Each band captures distinct sets of components and processes of one or more of the system levels. A band

[94] Though Newell and Simon (1976) initially developed this account together, Newell has since developed it the furthest, so I focus on it here.
[95] Newell goes on to list the 'evolutionary' and 'astronomical' bands, but he speculatively says little about these so I leave them out for the movement.

is set of qualitative expressions that capture various aspects of different system levels at different time scales – for example, the rational band uses intentional terms such as beliefs and desires to describe actions at particular speed. He writes, for instance: "Several adjacent [system] levels group together in what are called bands. Different bands are quite different phenomenal worlds" (1990, p.122).

Movement between the different bands for Newell is contingent on the reliable predications one can make about the behaviour of the system levels at each band. He writes, for example: "[E]ach new level occurs, so to speak, as soon as it can in terms of numbers of components. For the levels of the biological band, this factor is a reading from what we know empirically, so it stands as a confirmation of the level of analysis" (1990, p.123). Confirmation of the biological band comes from making reliable predications about behaviour of various components at the corresponding system level. Thus moving from band one to another results from a failure to make reliable predications about the system levels described at the band either above or below.

Newell also envisages a particular relation between the system levels and the different bands:

> Starting at the bottom, there is the biological band of three levels—neurons; organelles, which are a factor of ten down from neurons; and neural circuits; which are a factor of ten up…Above the biological band, there is the cognitive band. Here the levels are unfamiliar—I've called them the deliberate act, cognitive operation, and unit task, each of which takes about times (~~10) as long as the act at the level beneath (p.122-3).

Not only do each of the bands address distinct system levels, but they do so because of the distinct times scales associated with the system levels, e.g., ~~10ms for the neural circuit system level. The time scales of action distinguish the various bands. In short, distinct system levels, which are hierarchically organised according to compositional relations and which are distinguished by different time scales, are targeted by distinct descriptive bands.

The fingerprints of the HCL are all over Newell's view. First, structures in the world (the system levels) are arranged hierarchically according to a specific organisational scheme

(compositional relations). There is an ontological system of levels <L', S'>, consisting of several level-specific worlds L' ($\Omega_1 \ldots \Omega_n$) and a class of functions S' of the form $\sigma$: $\Omega \rightarrow \Omega'$, such that each lower system level determines facts at a corresponding higher system level. Second, the various system levels are described by distinct descriptive modes (bands). There is an explanatory system of levels <L, S>, consisting of several 'qualitative vocabularies' L ($\ell_1 \ldots \ell_n$), and a class of functions S of the form $\sigma$: $\ell \rightarrow \ell'$, such that each element of a band can be mapped to an element of another band in the hierarchy. And finally, each band describes a unique set of system levels, all operating at distinct time scales. There is structural mapping from the explanatory system of levels to the ontological system of levels, and vice-versa. There is a strict correspondence between levels of each of the hierarchies.

Importantly, though, the HCL does not simply form a passive conception in Newell's view, it also actively drives several of the inferences Newell makes about cognitive investigation.

First, consider that Newell makes a specific claim about the link between the different bands. In describing the relation between the biological and cognitive band, for instance, he writes: "starting with neural circuits of minimal size, a system level can be built out of these and then, with the new systems as components, another system level can be built on top…The base level for the cognitive band must be where the architecture starts" (1990, pp.131-2). The connections between different modes of analysis (e.g., the cognitive and biological bands) are a function of the compositional relations among the various system levels (e.g., the neural circuit system level and the component system level).

This feature of Newell's view is interesting, because it points to the specific influence of the HCL. In embracing the HCL, Newell is led into a particular solution to what Bermudez (2005) dubs the 'interface problem', the challenge of how to integrate or connect different modes of analysis. Since there is a direct correspondence between the modes of analysis we offer about cognition and the hierarchical structure of our cognitive architecture, explaining higher level phenomenon in terms of lower modes of analysis effectively requires finding the corresponding components and activities at the appropriate lower level of organisation

responsible for realising the higher level phenomenon being described – Bermudez (2005) calls such level crossing explanations 'vertical explanations'. The HCL shapes how Newell conceptualises the relation between different levels of investigation.

Second, consider that Newell makes a specific claim about structure of cognitive investigation itself. He writes, for example: "In any reasonable sense, the cognitive realm is reducible to the natural science realm…all that the lower level systems of the biological band do is support the computational mechanisms" (1990, p.149). Again, this feature of Newell's view is suggestive. Talk of the explanatory reduction of different modes of analysis, such as the cognitive band, seems to follow from the strict correspondence of the explanatory and ontological hierarchies. Downward movement between the various bands reflects a functional decomposition of high-level entities and processes into low-level ones. Moving from higher level input-output processes to lower level symbolic representations and procedures involves a process of functionally decomposing higher level phenomena at one mode of analysis to lower-level phenomena at another mode of analysis. Again, in adopting a version of the HCL, Newell is led into a particular stance on the structure of cognitive investigation.

So, not only is Newell's view committed to the HCL, but the HCL also plays an active role in driving several of the inferences Newell makes about cognitive investigation. It shapes the way he (i) thinks about the relation between different levels of investigation and (ii) how he conceptualises cognitive investigations as a process of functional decomposition.

### 6.3.2 Case Study 2

The second view I want to explore comes from Ron Sun. Sun's work is interesting because it sits at the cross-roads of experimental, computational and philosophical thinking in cognitive science; Sun does work in both experimental psychology and computational modeling.

In a co-authored paper entitled "On Levels of Cognitive Modeling", Sun sets out his vision of computational cognitive modeling, what is labelled a 'new hierarchy' of four levels. As the moniker suggests, Sun et al. propose four distinct levels for cognitive modeling: the

sociological/cultural level, the psychological level, the componential level, and the physiological level.

Each level has three parts: (i) a type of analysis, (ii) an object of analysis, and (iii) a computational model. On this multi-level hierarchical approach, different computational cognitive models are constructed on the basis of the various types of analysis they use to address distinct types of phenomena. Models can be constructed in either direction from the lowest to highest levels, but they need to address the appropriate processes, entities and causal relations at each level. Sun et al. write, for instance:

> [A] scientific theory of cognition requires the construction of a hierarchy of different levels with consistent causal descriptions from a low level through a series of intermediate levels to high-level phenomenon […] Scientific understanding depends upon the selection of key elements of cognitive phenomena and the creation of models for such elements at appropriate levels (2005, p.634).

The spectre of the HCL looms large in Sun's view.

First, the objects and causal relations at higher levels of Sun et al.'s hierarchy are defined in terms of combinations of objects and processes at lower levels. Sun et al. write, for instance: "Higher-level entities would be made up of sets of more detailed entities, and causal relationships at higher level would be generated by the casual relationships amongst the equivalent entities at more detailed levels" (2005, p.624). Different entities and processes relate via part-whole relations on the basis of considerations of size and complexity. The objects of the componential level, for example, which include intra-agent processes and components, are composed by implementational level entities and processes, such as action potentials. There is an ontological system of levels <L', S'>, consisting of several level-specific worlds L' ($\Omega_1 \ldots \Omega_n$) and a class of functions S' ($\sigma$: $\Omega \rightarrow \Omega$'), such each element of one level can be mapped to an element of another level in the hierarchy.

Second, computational models are assigned to different levels of the hierarchy in virtue of the complexity of the processes or entities they describe. Sun et al. write, for instance: "The difference [between types of analysis] is mostly in the information density (level of details) and thus the length and the complexity of description, and in the (likely) smaller number of

entities needed by description at a more detailed level" (2005 p. 626). As one moves up the hierarchy of explanatory levels, one moves to descriptions of increasingly complex and causally dense phenomena. The sociological type of analysis resides at a higher level than that of the componential analysis, for example, because it addresses causally more complex entities, and hence offers informationally denser explanations. There is an explanatory system of levels <L, S>, consisting of several languages L ($\ell_1 \ldots \ell_n$), and a class of functions S ($\sigma$: $\ell \rightarrow \ell'$), such each element of one level can be mapped to an element of another level in the hierarchy.

Finally, the relation proposed between the different aspects of the hierarchy is one of strict correspondence:

> A theory on any level creates entities at that level of descriptive detail as well as causal relationship between those entities that correspond with a range of data. Entities at a higher level often tend to package sets of lower-level entities in such a way that the higher-level causal relationships can be specified without reference to the internal structure of the higher-level entities (2005, p.622).

The phenomena addressed by the different types of analysis, and thereby computational models, are based on what unique causal relationships they capture. For example, because sociological phenomena support different causal relations than componential level phenomena, models constructed targeting sociological phenomena reside at a different level of the hierarchy than componential ones. Each type of analysis addresses only certain types of entities and process at various organisational levels. There is structural mapping from the explanatory system of levels to the ontological system of levels, and vice-versa.

Moreover, similar to Newell's view, several of the explanatory claims Sun et al. make can be seen as being driven by the HCL.

First, consider that Sun et al. claim that their new hierarchy offers a clear research programme for cognitive modeling. First, researchers need to characterise a phenomena at a specific mode of analysis, such as the cognitive or social mode of analysis. Then, after collecting experimental data, certain causal relationships between entities at that particular level (or across several levels) can be identified. Using these casual relationships, researchers

can then develop a fleshed out computational model, one with all the requisite algorithms and data structures. Finally, the computational model can be implemented and tested against empirical data.

Notice that organising computational modeling in this way seems particularly appealing when explanatory and ontological hierarchies are aligned via a strict correspondence relation. When there is an ordered hierarchy of levels, it makes sense why modeling can involve identifying phenomena at particular levels of ontology, describing them using the appropriate mode of analysis and then fleshing them out as computational models. The ordered stages in such a process require the promise of being able to neatly bridge the aligned explanatory and ontological hierarchies. Without such a picture, it is unclear why what is being modelled at a given mode of analysis is situated at the appropriate ontological level of the hierarchy.

Second, consider that Sun et al. also suggest that their new hierarchy offers a clear way of creating level crossing models. That is, it offers a means of "mixing levels, or integrating models at different levels". Researchers can provide not only precise, detailed models at a specific mode of analysis but they can also offer interesting *cross-* and *mixed-level* analyses, so long as they are clear about which phenomenon are being targeted, which constraints from higher and lower levels are included, and which types of analysis are being used.

Again, the HCL seems to be driving Sun et al.'s reasoning. If ontological levels are hierarchically organised, each mapping to a corresponding mode of analysis, then the prospect of developing mixed level models seems much easier. But notice that this requires a total ordering to the hierarchy. It requires being able to identify where a given phenomenon stands in the hierarchy respective to other levels. Without knowing which levels are above or below in the hierarchy, it becomes less clear which facts about entities and processes at different levels need to be incorporated. Of course, this is isn't to say that such integrated models cannot be developed. Rather, it is to simply point out that the suggestion itself finds a good deal of its support from the conception of levels at play in the account.

So, again, not only can the HCL be found in a prominent view of cognitive science but it also can be seen to actively drive various aspects of the view's reasoning. It helps to justify both a distinct research programme in computational cognitive modeling and a method for integrating different types of models.

## 6.4 The Shifting Nature of Levels

The dangers of various parts of the HCL have been well documented elsewhere. For example, in philosophy of biology, Potochnik and McGill (2012) worry that the hierarchical notion of ontological levels does not make room for complex forms of realisation; while in cognitive neuroscience, Craver (2007) worries that the ontological notion of levels fails to make sense of the rich multi-layered structure of the brain.[96] These are important criticisms, and there is much to be said for them.

However, in what follows, I develop a novel worry. I argue that the HCL is uniquely problematic for cognitive science because it conflates two important kinds of shifts. It does so because it neglect an important part of our explanatory practices about complex systems: namely, the role of functional contextualisation. To be clear, I take this argument to apply specifically to cognitive science. It might be that the HCL fails in other domains for different reasons, but here I am concerned with how it applies in cognitive science. I begin by laying out the different kinds of shifts and then saying why the HCL neglects functional contextualisations.

First, there are *shifts in analysis*. These are shifts in the mode of analysis or language used to describe a complex system. For instance, one might talk about a particular function, such as (car(list)), in terms of a sequence of machine language instruction (e.g., a programming program) or a local characterisation of the operation (i.e., the machine language). In such cases, the action's functional role is being re-interpreted relative to a new theoretical vocabulary or

[96] See also Kim (2002), Rueger & McGivern (2010), and Eronen (2013).

language (e.g., the language of computer programming versus machine code). When one switches from one theoretical vocabulary to another, one is shifting from one mode of analysis to another.[97]

Second, there are *shifts in grain*. These are shifts in the generality of the explanations used to describe a complex system. For example, most of the functions within a low-level assembly programming language can be specified without reference to the machine language – that is, the code that directly implements actions in the computer's CPU. But, to understand how the primitive functions of the low-level assembly program translate into the machine code, one occasionally needs to shift to the more rapidly executed units of the machine code (i.e., the 1s and 0s). Here, unlike the previous case, there is a shift not only in the mode of analysis (the theoretical vocabulary) but also in the grain of analysis (the actual size or speed of the components being described). As opposed to shifts in analysis, shifts in grain lessen the degree of abstraction in one's descriptions.

The problem with the HCL is that it runs together these two kinds of shifts. It assumes that shifts in analysis always occur with shifts in grain; that in any case where there is a shift in grain there is a corresponding shift in the mode of analysis.

To see this, recall the HCL's view on functional decomposition. As we saw with Newell's view, the HCL claims that shifts among modes of analysis can be interpreted as functional decompositions across ontological levels of a hierarchy. A higher-level phenomenon, such as spatial memory, might be functionally decomposed into lower-level components and mechanisms, such as LTP mechanisms or NMDA receptors, by moving from higher to lower modes of analysis, such as the cognitive to biological band. The reasoning was that because of the strict correspondence between a given mode of analysis and level of ontology, any shift among modes of analysis has to precipitate a change in the level of ontology. To speak of

[97] Here I am taking 'theoretical vocabulary' to simply mean the explanatory tools one uses plus the language in which those tools are couched. So, for example, computational analysis not only uses the language of computing but also couches its explanations in terms of how some phenomenon can be said to compute or not.

moving from higher- to lower-level explanations (e.g., cognitive to biological band) just is to move from higher- to lower-level organisationally layered components and processes (e.g., spatial maps and LTP mechanisms) within a system.

Cases where shifts in analysis do reflect shifts in grain, the HCL does well. However, where the view struggles is cases in which the relationship breaks down. Consider, for example, predictive processing accounts of perception. Predictive processing accounts treat perception as a top-down, error-minimisation task (Clark, 2013, 2016). The brain is said to be trying to continuously reduce the mismatch between incoming sensory signals and previous expectations, using computational principles approximating Bayes. In this case of vision specifically, the brain is thought to integrate information from low-level visual features into complex representations through top-down processing, creating a hierarchy of increasingly abstract representations in the visual system.

The point for present purposes is that this predictive hierarchy is said to map to the functional and anatomical structures of the brain, with the PFC at the top and sensory peripheries at the bottom (Hohwy, 2016). This is important because it means that the analysis of the bottom-up and top-down information flow involves shifts in grain (spatial and temporal activities), but nonetheless employs the same mode of analysis (i.e. computational analysis). There is a shift in grain but no corresponding shift in mode of analysis.

As a further example, consider Changeux's (2017) 'dynamical nesting model'. For Changeux, the brain is a nested assembly of functional structures at multiple levels of organisation, including: the level of genome, the TF–gene network level, the level of epigenetic action on synapse formation, and the level of long-range connectivity. Each level is reciprocally inter-regulated and in constant dynamic evolution with the others. What is again relevant here is that while the brain has a dynamic mutli-level structural organisation, all of the levels are governed by explanations using one mode of analysis: namely, the level of physics and chemistry. Contrasting his approach with others he writes, for example: "the models aimed at representing and/or simulating a process and/or behaviour on the basis of

minimal, yet realistic, architectures and activity patterns most often use a single level of organisation. To attempt a type of modelling that spans several levels, as proposed here, is in itself a theoretical position" (2017, p.169). So, again, while shifts in grain of analysis occur across multiple temporal and spatial scales, the actual mode of analysis remains constant.

But the reverse case also holds. It is possible to have shifts in analysis without having shifts in grain. To return to the program/machine code example, here we have a single function, (car(list)), that is characterised both in terms of a sequence of computer language instruction (i.e., the low-level assembly program) and as a local characterisation of the operation (i.e., the machine language). It represents a shift in the mode of analysis, but not the actual grain or size of the action. While the property or action's functional role is re-interpreted relative to a new theoretical vocabulary (e.g., the language of computer programming versus machine code), the actual property or action itself remains at the same grain of analysis. Thus, there is a change in the mode of analysis, but no change in the grain of analysis.

So, it seems that there are two types of shifts available for complex systems, which in turn lead to three possible options in conceptual space. Figure 3 provides illustration.

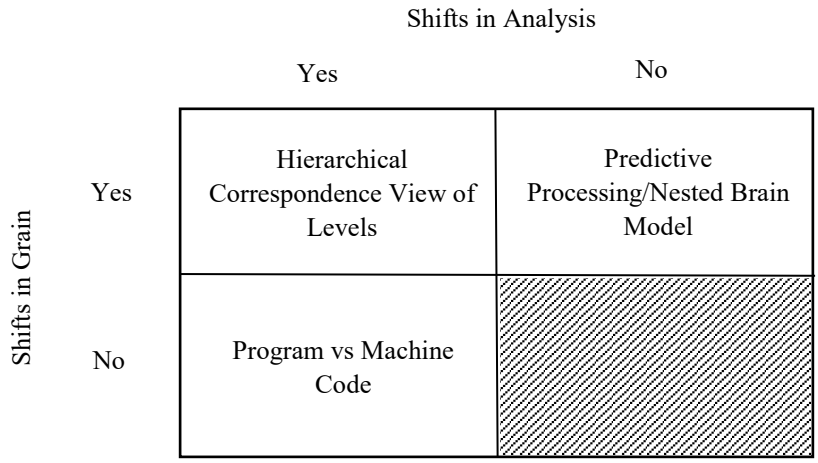


Figure 3. Two by two matrix of possible combinations of grain versus analysis shifts.

If shifts in grain (size of property) and shifts in analysis (mode of analysis) come together, then something like the HCL follows: shifts in grain accompany shifts in analysis at every turn. But if the two shifts come apart, then there are two further possibilities: either there are

shifts in analysis without shifts in grain (the computer language vs machine code case) or there are shifts in grain without shifts in the analysis (the predictive processing and brain model cases). As we have seen, the HCL, in conflating the two kinds of shifts, restricts the conceptual space so as to only include the first set of cases. This has two important consequences.

The first is that it means that the HCL imposes an unnecessary methodological constraint on investigation. In envisaging the shift between modes of analysis as a shift between levels of organisation, the HCL restricts how many times a given mode of analysis might apply. For example, as Bechtel (1994) notes, one can just as easily apply Marr's modes of analysis to low-level intracellular processes, such as oxidative phosphorylation, as one can to high-level cognitive processes, such as vision. The reason is that, at core, Marr's levels simply answer different types of questions. We can ask what purpose oxidative phosphorylation serves at one level, or what metabolites contribute to the process at another. Computational and algorithmic analysis do not in every case have to mark shifts in different sized properties. There is no reason a conception of levels should place such constraints on investigations ahead of time.

The second is that it means that the HCL makes for a questionable empirical hypothesis about complex systems. In over-stating the fit between explanatory and organisational levels, the HCL precludes the possibility of cognitive systems having a multi-nested structure. McClamrock (1991) foreshadows this worry talking about Marr's account: "[i]f we were to take the 'three levels' view as making a claim about the actual number of levels of organization (or stages of natural decomposition) in cognitive systems, it would be a very substantive (and I think false) empirical claim. It would be claiming that cognitive systems will not have any kind of multiple nesting of levels of organization" (p.191). The HCL assumes that complex systems will not have a nest structured within a given level of ontology.[98] However, again, as we saw, this seems like an overreach. There is no reason a conception of levels should make such a specific assumption about the structure of complex systems like cognition.

[98] To be clear, by 'nested structure' I mean here that fact that within a level of ontology there might exist several temporal and spatial grains.

### 6.4.1 Functional Contextualisations

So why does the HCL run the two types of shifts together, and so become overly restrictive? It does this because it fails to appreciate an important part of our explanatory practices about complex systems: namely, it neglects the role played by 'functional contextualisations'.

Simply put, functional contextualisations are analyses of the way in which a physical structure's functional role is often determined by situating or embedding it within a wider system. As Cummins (1983) puts the idea: "to ascribe a function to something is to ascribe a capacity to it that is singled out by its role in an analysis of some capacity of a containing system" (p.99). Functional contextualisations are not simply capacities picked out by their place in a system but, rather, analysis of the activity of some item in terms of how it is organised into the workings of a system (Craver, 2013).

For example, to say that the heart distributes oxygen or pumps blood, one has to not only describe the heart's local/intrinsic properties, such as its constricting and releasing activities, but also its position within the circulatory system. Without reference to the external objects and activities of the circulatory system, the functional role of the heart is unclear; the object's local/intrinsic properties fail to identify its functional role. If the contextual description changes, then so too does the heart's function. So, if, for example, the heart is described as a 'noise-maker', then it is embedded in a new system (a new causal nexus), and thereby implicates a new set of components and activities, such the resonance frequency of the heart's chamber walls or the capacity to use the heart in a three piece band. Characterised one way, the heart forms part of the mechanism for blood circulation; characterised another way, it forms part of the mechanism for sound generation.[99]

Functional contextualisations are equally important in cognitive investigations. This is because cognitive systems, in being information processing systems, are, in principle,

[99] Craver (2013) frames these upward and downward changes in terms of 'contextual' and 'constitutive' explanations but the idea is the same.

decomposable any number of times. As McClamrock notes: “the degree of functional decomposition (i.e. the true level of organization) might be done any number of times for some particular information processing system” (1991, p.191). The functional decomposability of a given cognitive or perceptual system is not only tied to its organisational structure, but also to how the system is initially functionally contextualised.

Take vision, for example. If characterised in terms of facilitating object recognition, a whole suite of properties and activities are salient to investigation. These include fine-grained properties, such as those in the primary visual cortex involved in edge detection, and coarse-grained properties, such as those involved in categorisation. However, if, in contrast, vision is characterised in terms of facilitating action-guidance, then an entirely new set of properties and activities are implicated, such as those involved in maintaining body images or reaching behaviour. The functional decomposability of vision depends not only on structural organisation of the phenomenon but also on how the activity is initially characterised.

Once it is appreciated that functional characterisations help fix organisational relations, it starts to make sense why shifts in grain and analysis come part. Shifts in grain are only possible relative a given functional contextualisation. They are what set the grain of analysis. Without either implicitly or explicitly contextualising a property or object, the levels of functional decomposability available for the system remain unclear, particularly for systems with nested structure.

For example, to return to the cell case, once a function has been specified, such as oxidative phosphorylation, a number of different forms of analysis can be applied. One might, in the spirit of Marr, ask algorithmic questions about what metabolites contribute to the phosphorylation process, or, alternatively, one might ask questions about what intracellular components and activities are responsible for implementing phosphorylation. Once the grain of analysis has been set shifts among modes of analysis can occur; not the other way round. In running the two kinds of shifts together, the HCL is unable to appreciate the role of functional

contextualisations in analysis. It wrongly assumes that the grain of analysis is set by the mode of analysis. Mechanists have long appreciated this point. Toulmin (1975), for example, writes:

> Indeed, the very organization of organisms – the organization that is sometimes described as though in simply involved a 'hierarchy' of progressive larger structure – can be better viewed as involving a 'ladder' of progressively more complex systems. All these systems, whatever their level of complexity, need to be analysed and understood in terms of the functions they serve, and also of the mechanisms they call into play (p.53).

The question of whether two properties are localisable to the same level is only answerable relative to a particular functional contextualisation. It is only once a property has been provided with a contextual function that we can begin to speak of parts being at a 'higher' or 'lower' levels. This does not mean we cannot say whether one part is larger or smaller, faster or slower than another (shifts in grain are still possible). Rather, it means that things like spatial or temporal relations do not by themselves mark organisational divides in the world.

## 6.5 Objections

Time to consider some objections. First, one might worry that the preceding argument relies too heavily on mechanistic considerations. If one rejects the mechanistic conception of levels, then maybe there is little reason to buy the argument on offer.

One reason to think otherwise is that while the argument does lean on several points made by mechanists, the argument as a whole relies on points distinct from the mechanistic conception of levels. While it is true to say that the mechanistic account of levels is embedded within a larger program – the notion of 'levels of mechanisms' is defined, for example, largely in terms of other technical terms, such as mechanistic explanation and mechanism – the general grain versus analysis distinction is one several authors have pointed to more generally. For example, Floridi (2008) builds his account of 'method of levels of abstraction' on just such a distinction. So while the mechanist might use the grain versus analysis distinction to particular ends, the distinction itself holds more generally.

Furthermore, functional contextualisations, while used by mechanists, are also not something intrinsic to the framework. Cummins (1983) and McClamrock (1991), for example, both invoke the notion in analysing psychological explanations more generally. Both argue that functional contextualisation are important for understanding the relationship between higher and lower levels of explanation. So, while the current argument does rely on points by the mechanist, is it not reducible to or highly dependent on such claims. It does not presume a mechanistic conception of levels.

Second, one might wonder why the advocate of the HCL cannot admit the grain/analysis shift distinction but still maintain that there is a correspondence relation. After all, both Newell (1990) and Sun et al. (2005) concede that there are several spatial and temporal scales associated with each level of ontology.

The problem is that while both authors can admit that there are shifts in grain *within* a level of organisation, neither can say that there are shifts *across* levels of ontology. The examples provided showed as much. Changeux's (2017) dynamical nest model, for example, showed that across quite notably spatial and temporal scales, e.g., genes to synaptic connections, the same modes of analysis could apply. There are several instances in which there can be quite extreme cases of grain shifts without shifts in analysis. The issue for the HCL is that in embracing a strict correspondence relation it commits itself to saying that at some point shifts in grain are going to reflect shifts in ontological levels, but there is good reason to think that this is not the case.

Finally, does not relying on the notion of functional contextualisations make explanations using different modes of analysis relative or perspectival? The answer is yes, but not in a vicious way. While it is true that, in principle, any an item can be functionally contextualised relative to a given system, only a subset of the contextual functions ascribed will be explanatorily interesting. For example, whether or not the heart functions as a noise-maker may be contingent on the larger system in which it is embedded, but this does not mean that explaining the heart's function in terms of being a noise-maker will be of much explanatory

interest to us. Only a subset of the causal nexuses we choose to describe will turn out to have any explanatory value.

Moreover, one of the benefits of making room for functional contextualisations is that it allows a new vantage on certain debates. Enactivist views, for example, are often keen to emphasis the action-guidance functions served by various perceptual or cognitive systems. This is in contrast to representationalists accounts, which often focus on describing how cognitive systems function to recover rich, structured information from impoverished sensory data.

The benefit of having functional contextualisation in hand is that rather than seeing this debate as one about the 'right' mode of analysis for cognition. The debate might be instead better understood as a disagreement about which functional contextualisations should take precedent using which modes of analysis. So, in the case of vision, for example, the representationalist is keen to emphasise things like object-recognition using a computational analysis, while the enactivist is focused on action-guidance behaviours using more algorithmic-type analysis. The point is that while relying on functional contextualisations to set the grain analysis used within a given cognitive investigation might be somewhat perspectival, this does not mean that it cannot be explanatorily productive.

## 6.6 The HCL Reconsidered

As the preceding argument demonstrates, the strict correspondence relation and the total ordering of the ontological and explanatory hierarchies offer overly restrictive assumptions when applied to complex systems such as information processing systems. Given this, what I want to do now is briefly sketch what an alternative version of the HCL might look like without these problematic elements.

First, consider what a version of the HCL might look like without the strict correspondence relation. If the HCL no longer claims that there is an isomorphic mapping from a system of levels <L, S> to <L', S'>, then while each level and mapping of <L, S> can be assigned to a

corresponding level and mapping in <L', S'>, there is a no mapping in the other direction – <L, S> and <L', S'> are no longer structurally equivalent. So while <L, S> and <L', S'> still denote distinct notions of levels – for example, levels of ontology and modes of analysis – there is now only a mapping from <L, S> to <L', S'>.

To illustrate, return to Marr's (1982) tripartite framework. According to the original version of the HCL, Marr's computational, algorithmic and implementation levels denoted three languages for describing cognition $\{\ell_1, \ell_2, \ell_3\}$ and an ontology of levels $\{\Omega_{1...}\ \Omega_n\}$. Marr's levels formed a non-empty class of modes of analysis, L, and a corresponding non-empty class of levels-specific worlds, $\Omega_L$, such that there was a mapping between each mode of analysis in L and a level-specific world in $\Omega_L$.

However, according to this weaker version of the HCL, while Marr's levels still denote a non-empty class of modes of analysis, L, and a corresponding non-empty class of levels-specific worlds, any level $\ell$ of <L, S> can now be mapped to one or more level-specific worlds $\Omega_L$ in <L', S'>. To say that a sentence $\Phi$ is true at the computational mode of analysis, $\ell_1$, for instance, is to say that $\Phi$ denotes a set of facts or entities in a corresponding level-specific world, $\Omega_L$, but it is no longer the case that $\ell_1$ needs to only be mapped to $\Omega_{L1}$. It can be mapped to any level-specific world, as long as the appropriate orderings are preserved.

Second, in terms of total ordering of the hierarchy, while there would still be a mapping of levels *between* <L, S> and <L', S'>, it would no longer be the case that the levels *within* <L, S> or <L', S'> could be given a total ordering. This does not mean that levels cannot still be organised – one could still, for example, map elements of $\ell_1$ to elements to level $\ell_2$ – but one can no longer tell whether for any two levels that one is necessarily higher or lower than the other. For any two levels in a system of levels, there is function S from either from $\sigma$: $\ell \rightarrow \ell'$ or from $\sigma$: $\ell' \rightarrow \ell$ but not both.

So, again, returning to Marr's view as an example, while one can say that a particular set of descriptions at the computational level $\ell_1$, ones that map to a level-specific world $\Omega_{L1}$, are 'above' another set of descriptions at the algorithmic level $\ell_2$, one cannot say that another set

of computational descriptions $\ell_3$, ones mapped to a different level-specific world $\Omega_{L2}$, are necessarily 'higher' or 'lower' than $\ell_2$. So, in the modified version of the HCL, there are two distinct systems of levels, but there is no strict correspondence or total ordering of levels.[100]

The real question is: can this weaker version of the HCL accommodate the previous worries?

First, with respect to the question of iteratively applying modes of analysis, if the mapping relation between systems of levels is no longer one-to-one, then a given mode of analysis can, in principle, apply to more than one ontological level. A particular language in <L, S>, $\ell_1$, no longer needs to be mapped uniquely to a level-specific world in <L', S'> such as $\Omega_{L1}$. It could just as easily map to $\Omega_{L2}$ or $\Omega_{L3}$. This modified version of the HCL can now accommodate applying the same mode of analysis, such as Marr's computational level, to both lower-level processes, such as oxidative phosphorylation, as well as higher-level processes, such as object recognition. Without the strict correspondence relation, different modes of analysis are no longer tied to specific grains of analysis. Investigation at a given mode of analysis can span several shifts in grain without a corresponding shift in mode of description.

Second, with respect to the question of nested structure, if systems of levels are no longer locked into a mapped one-to-one, then there are multiple mappings possible different between modes of analysis and levels of ontology. For some complex system, S, composed of ontological levels, $\Omega_L$, a language, $\ell_1$, can apply to some subset of $\Omega_L$, while another language, $\ell_2$, can apply to a subset of that subset. Unlike Newell or Sun's views, in eschewing the strict correspondence relation, the modified version of the HCL can accommodate the possibility of cognitive systems with nested structure. A given mode of analysis can now address entities and process at spanning multiple grains of analysis within shifts in the mode of description.

[100] Formally, this revised version of the HCL respects the previous three conditions S1- S3. The important different is that drops the total ordering property.

Finally, with respect to functional contextualisation, if a system of ontological levels is only partially rather than totally ordered, then how a given levels gets initially fixed is undetermined. Since one can only say on a partial ordering that a level is locally higher or lower, the HCL leaves the question open as to how a particular entity or activity at a given ontological level is initially assigned. In a dropping the commitment to total orderings the same entity or activity can be situated or embedded in different systems depending on the explanatory need.

## 6.7 Conclusion

So, what have I shown? And why is it important? First, I have shown that the HCL can, in fact, be found in cognitive science, and that it has it driven certain explanatory inferences. This reveals what many have already suspected: that talk of levels is not just a harmless metaphor but one which can actively structure how thinking proceeds on philosophical and scientific problems. Second, I have shown that the HCL is particularly problematic when applied to the types of systems studied in cognitive science. Unlike previous worries, which often derive from more general concerns about accommodating various outlier cases, the present argument derives its force from the subject matter of cognitive science itself: namely, information processing systems. Third, I have shown that while the HCL is generally unsuitable for cognitive science, there is a way to make the view serviceable. Following a number of other authors, such as Kim (2002), I have shown that a local approach can deliver a conception of levels that is responsive to the needs of a particular domain of inquiry – in the present case, accounting for the role of functional contextualisations, iterative applications of modes of analysis, and the possibility of nested structure. Finally, in applying and extending List's (2019) formal framework, I have hopefully been able to bring an additional level of precision to the discussion of levels in cognitive science.

What lessons are to be drawn? On the one hand, the lesson is quite general. It is one similar to others drawn elsewhere about levels: we do not need a monolithic and all-inclusive

hierarchy of levels to make progress on scientific and philosophical questions. A local, domain specific approach is preferable. On the other hand, the lesson is quite specific: cognitive systems not only require their own unique conception of levels, but such a conception has to pay special attention to various properties unique to the types of information processing systems studied in cognitive science.

# Conclusion

I started off the thesis by asking a simple question:

*What are the boundaries of computational mechanisms?*

I labelled this the boundary question. The answer, I have suggested, is that computational mechanisms – at least, the ones that form the proper units of analysis in cognitive science – are certain types of distributed or world-spanning functional mechanisms. I have proposed that the computational mechanisms underlying certain cognitive phenomena, such as bat echolocation or cricket phonotaxis, are, in fact, wide environment-involving mechanisms manipulating medium-independent vehicles. This view brings together two distinct strands of computational thinking: those views which emphasise the *location neutrality* of computation and those which view computation as a *mechanistic* process. In bring these strands of thought together, I have attempted to articulate a vision of computation wherein computational mechanisms are sustained and generated by the on-going, reciprocal feedback that often exists between brain, body, and world.

The argument of the thesis has been developed on multiple fronts. On the empirical front, I have tried to show that wide mechanistic computation garners a good deal of support from a general class of systems, active sensory systems, of which bat echolocation and sightless spatial navigation are instances. These examples, I suggested, plausibly fall under the rubric of wide mechanistic computing systems. On the explanatory and methodological front, I have attempted to show that wide mechanistic computation is an equal competitor to internalist accounts; that there is a plausible case for considering wide computational explanations superior to their rival internalist counterparts. And on the conceptual front, I have shown how

wide mechanistic computation might be used to resolve thorny questions in 4E cognition; that wide mechanistic computation has something important to add to conversations about 4E cognition.

Taken together, this motley of evidence forms a rich mosaic. It paints a picture of an approach supported and sustained by multiple lines of evidences from a variety of sources. As I mentioned in the Introduction, the thesis is better seen as an attempt to weave together different strands of empirical research into an independently plausible and well-motivated approach to computation, rather than as an extended inference to best explanation.

The significance of the thesis is fourfold.

First, in highlighting the limits of existing approaches, I have tried to show that the best hope for the wide approach to computation, in the spirit Wilson, Hutchins, Wells, and Losonsky hope, is to throw in with the mechanistic account. I have argued that if wide computation is going to offer something to cognitive science, then its fate should be tied up with the mechanistic framework. The mechanistic account offers the best means avoiding persistent explanatory and methodological questions that have plagued previous formulations of the view.

Second, in addressing the abstraction problem and the two challenges, I have tried to not only demonstrate the plausibility and viability of wide mechanistic computation but also place mechanistic computation itself on a surer theoretical footing. An additional benefit of advancing the case of wide mechanistic computation is that it has offered an opportunity to reflect on, and extend, the foundations of the mechanistic approach to computation more generally. Thus, even if one does not buy the mechanistic account of wide computation, there is still something to be gained from my defense of mechanistic computation.

Third, in examining recent proposals for how to connect predictive processing accounts and extended cognition, I have tried to sound a cautionary note about how we should think about this new and exciting approach to the mind. I have suggested that while in its current state PP does not settle the internalist/externalist debate, this does not mean that it cannot

potentially play an important role if suitably developed. What I have tried to show is not only the limits of existing PP-inspired approaches to extended cognition but also what it would take to make progress on extended cognition debates using PP.

Fourth, and finally, in unpacking and examining the role of the HCL in cognitive science, I have tried to minimise the effects of an attractive but ultimately misleading picture of levels. I have tried to counter a conception of levels that has influenced both what and how researchers have approached certain questions in cognitive science. At minimum, any story about the computational 'level' of analysis in cognitive science has to square with this fact going forward.

Where do we go from here?

First, I think more needs to be done to flesh out the empirical case. What I have done so far is, at best, a proof of concept. A good deal of the heavy lifting remains to be done. To really drive home the argument, further examples of wide mechanistic computing need to be identified and examined. As I have argued, I think this can be done. There already exists a robust class of phenomena amenable to wide analysis. Additional focus on such cases will only help to further the case for wide mechanistic computing.

Second, actual cases of extended autonomous computational systems need to be found. In Chapter 3 I suggested that a resolution to the second tension could be achieved via highlighting a shared set of conceptual resources between the wide computationalist and the enactivist, such as the concept of autonomy. And while this might be enough to rebut the claims of the anti-computationalist, it isn't enough to achieve full reconciliation. Full reconciliation requires finding actual cases. I have already pointed to some key features: functionally-closed systems. An extended computational autonomous system will be one that involves a functionally-closed system computing some function, such as certain types of Turing machines. Identifying such cases will only help to further bolster the case for integration and reconciliation.

Third, the preliminary argument sketched in Chapter 5 needs to be further fleshed out. I only said what it *would* take for predictive processing to move the needle forward in

discussions of extended cognition. Showing that the argument is actually true requires a further defence and elaboration each of the premises. For instance, while I think that a good deal of empirical evidence already speaks in favour of the unique role of prediction error minimisation in cognition (premise 1), showing that extended system actually trade in such features, as a matter of fact, will require more work (premise 3). Clark (2016a, 2017a) has usefully pointed in the general direction of where to look for such evidence, e.g., low-cost, action driven heuristics, but a thorough and extensive review of the empirical literature is still needed before a PP-inspired extension argument in the vein described can be fully vindicated.

Finally, while it is outside the philosopher's purview, I think actual studies need to be conducted. If what has been said is on track, then there should be a way forward when it comes to 'testing' for wide computing mechanisms. I say 'testing' because, as I have argued, any computational investigation is going to have a perspectival element. Cases of wide computation cannot simply be read off phenomenon in the world. A good deal will depend on how a phenomenon is initially functionally contextualised. What counts as computational is, in part, dependent on the explanatory needs of the investigator. That said, connecting wide investigations to interlevel experiments, as was done in Chapter 2, should provide at least some methodological guidance when it comes to making sense of wide mechanistic systems experimentally. There is lots to do but the stage is set.

Having looked forward to what is to come, I want to now look back briefly and reflect on where we have been. It has been over 25 years now since wide computationalism was first proposed. There have been spurts and starts, gestures and overtures, and yet the view has alluded a sustained treatment. While the view has been defended and mobilised, discovered and rediscovered, it has remained disjointed and untethered. In the conclusion to his initial presentation, for instance, Wilson (1994) writes: "I think, however, that the most interesting issue concerns not the coherence of wide computationalism but the extent to which a wide computational research strategy and could be employed within cognitive psychology" (370). For better or worse, I have tried to answer Wilson's call-to-arms. I have tried to show that

wide computationalism can offer an empirically interesting and explanatorily attractive research strategy to cognitive science, particularly when it draws on a network of concepts from mechanistic philosophy and empirical psychology. My only hope is that we do not have to wait another 25 years for the results of this effort to be taken up.

# Bibliography


Adams, F., and Aizawa, K. (2001). The bounds of cognition. *Philosophical Psychology*, 14, 43-64.

Adams, F., & Aizawa, A. (2008). *The bounds of cognition*. Oxford: Blackwell Press.

Allen, M, and Friston, K. (2016). From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese*. doi:10.1007/ s112 29-016-1288-5.

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

Baumgartner, M., & Wilsktzy. (2017). Is it possible to experimentally determine the extension of cognition? *Philosophical Psychology*, 30(8), 1104-1125.

Baumgartner, M., and Casini, L. (2017). An Abductive Theory of Constitution. *Philosophy of Science*, 84 (2), 214-233. doi: 10.1086/690716.

Baxendale, M. (2016). The Layer Cake Model of the World and Non-Reductive Physicalism. *Kriterion – Journal of Philosophy*, 30(1), 39-60.

Bechtel, W. (1994). Levels of Description and Explanation in Cognitive Science. *Minds and Machines*, 4, 1-25.

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

Bechtel W., and Richardson, R. (1993). *Discovering Complexity: Decomposition and Localization as Scientific Research Strategies*. Princeton: Princeton University Press.

Bermudez, J. (2005). *Philosophy of psychology: A contemporary introduction*. New York: Routledge.

Boone, W., and Piccinini, G. (2016). Mechanistic Abstraction. *Philosophy of Science*, 83, 686-697. doi: 10.1086/687855.

Brooks, D. (2017). In Defense of Levels: Layer Cakes and Guilt by Association. *Biological Theory*, 12(3),142–156.

Brooks, D, and Eronen, M. (2018). The significance of levels of organization for scientific research: A heuristic approach. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 68-69, 34-41. doi: 10.1016/j.shpsc.2018.04.003.

Chalmers, D. (1994). On implementing a computation. *Minds and Machines*, 4(4), 391–402.

Chalmers, D. (1996). Does a rock implement every finite-state automaton? *Synthese,* 108, 310–333.

Chalmers, D. (2011). A computational foundation for the study of cognition. *Journal of Cognitive Science*, 12(1), 323–357.

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

Chrisley, R L. (1995). Why everything doesn't realize every computation. *Minds and Machines*, 4, 403– 430.

Changeux, J. (2017). Climbing Brain Levels of Organisation from Genes to Consciousness. *Trends in Cognitive Sciences*, 21(3), 168-181.

Clark, A. (2003). *Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence*. Oxford: Oxford University Press.

Clark, A. (2005). Intrinsic content, active memory, and the extended mind. *Analysis*, 65, 1–11.

Clark, A. (2008a). Pressing the Flesh: A Tension in the Study of the Embodied, Embedded Mind? *Philosophy and Phenomenological Research*, 76(1), 37-59.

Clark, A. (2008b). *Supersizing the mind: Embodiment, action, and cognitive extension*. Oxford: Oxford University Press.

Clark, A. (2010). Coupling, Constitution and the Cognitive Kind: A Reply to Adams and Aizawa. In R., Menary (Ed), *The Extended Mind*. Aldershot: Ashgate.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 1–24.

Clark, A. (2015). Radical Predictive Processing. *The Southern Journal of Philosophy*, 53, 3-27.

Clark, A. (2016a). *Surfing Uncertainty: Prediction, Action and the Embodied Mind*. Oxford: Oxford University Press.

Clark, A. (2016b). Attention alters predictive processing. Behavioural Brain Science, 39. doi: 10.1017/S0140525002472.

Clark, A. (2017a). Busting out: Predictive brains, embodied minds, and the puzzle of the evidentiary veil. *Noûs*, 51(4), 727-753.

Clark A. (2017b) How to knit your own Markov blanket: resisting the second law with metamorphic minds. In T. Metzinger, W. Wiese (Eds.), *Philosophy and predictive processing: 3*. Frankfurt am Main, Germany: MIND Group.

Clark, A., & Chalmers, D. (1998). The Extended Mind. *Analysis*, 58, 7-19.

Clark, A., & Kiverstein, J. (2009). Introduction: Mind Embodied, Embedded, Enacted: One Church or Many? *Topoi*, 28, 1-7.

Churchland, P. S., and Sejnowski, T. (1990). Neural Representation and Neural Computation. *Philosophical Perspectives*, 4, 343-382.

Coelho Mollo, D. (2018). Functional individuation, mechanistic implementation: the proper way of seeing the mechanistic view of concrete computation. *Synthese*, 195, 3477-3497. doi: 10.1007/s11229-017-1380-5.

Cummins, F. (1983). *The Nature of Psychological Explanation*. Cambridge, MA: MIT Press.

Craver, C. (2006). When Mechanistic Models Explain. *Synthese*, 153(3), 355-76.

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

Craver, C. (2013). Functions and Mechanisms: A Perspectivalist View. In: Huneman P. (Ed.) Functions: selection and mechanisms. Synthese Library (Studies in Epistemology, Logic, Methodology, and Philosophy of Science), vol 363 (pp 133-158). Springer, Dordrecht.

Craver, C. (2015). Levels. In T. Metzinger & J. M. Windt (Eds.), *Open MIND:* 8. Frankfurt

am Main: MIND Group. 10.15502/9783958570498.

Darden, L. (2002). Strategies for discovering Mechanisms: Schema instantiation, modular subassembly, forward/backward chaining. *Philosophy of Science*, 69(3), 354-365.

Dawson, M. (1998). *Understanding Cognitive Science*. Blackwell Publisher.

DiFrisco, J. (2017). Time Scales and Levels of Organization. *Erkenntnis*, 82, 795–818.

Dennett, D. C. 1981. Three Kinds of Intentional Systems. In R. Healey (Ed), *Reduction, Time and Reality* (pp.37-61). Cambridge, England: Cambridge University Press.

Dennett, D. C. (1987). *The intentional stance*. Cambridge: MIT Press.

Dennett, D. C. (2000). With a Little Help from My Friends. In D. Ross, A. Brook, and D. Thompson (Eds.), *Dennett's Philosophy: A Comprehensive Assessment* (pp.327-388). Cambridge, Mass.: MIT Press.

Dretske, F. (1981). *Knowledge and the flow of information*. Cambridge: MIT Press.

Dewhurst, J. (2016). Computing Mechanisms and Autopoietic Systems. In Müller (Ed.), *Computing and Philosophy*. Heidelberg: Springer (Synthese Library).

Dewhurst, J. (2018). Individuation without Representation. *British Journal of Philosophy of Science*, 69: 103–116. doi: 10.1093/bjps/axw018.

Dewhurst, J., and M. Villalobos. (2017). The Enactive Automaton as a Computing Mechanism. Thought: *A Journal of Philosophy*, 6, 185–192.

Di Paolo, E. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences*, 4 (4), 429–452.

Di Paolo, E. A. (2009). Extended Life. *Topoi*, 28, 9-21.

Di Paolo, E. A. & Thompson, E. (2014). The enactive approach. In L. Shapiro, (Ed.), The *Routledge Handbook of Embodied Cognition* (pp.68-78). New York: Routledge Press.

Egan, F. (1992). Individualism, computationalism and perceptual content. *Mind*, 101, 443–59.

Egan, F. (2010). Computational Models: A Modest Role for Content. *Studies in History and Philosophy of Science*, (3), 253-259.

Egan, F. (2018). The nature and function of content in computational models. In M. Sprevak

& M. Colombo (Eds.), *Routledge Handbook of the Computational Mind* (pp.247-258). London: Routledge.

Elber-Dorozko, L., and O., Shagrir. (2018). Computation and Levels in the Cognitive and Neural Sciences. In M. Sprevak and Matteo Colombo (pp. 205–225). London: Routledge. doi:10.4324/9781315643670.

Eronen, M. (2013). No Levels, No Problems: Downward Causation in Neuroscience. *Philosophy of Science*, 80(5), 1042–1052

Eronen, M. (2015). Levels of Organization: A Deflationary Account. *Biology and Philosophy*, 30(1), 39–58.

Floridi, L. (2008). The Method of Levels of Abstraction. *Minds and Machines*, 18(3), 303–329.

Fodor, J. (1981). *Representations: Philosophical Essays on the Foundations of Cognitive Science*. Cambridge, Mass.: MIT Press. A Bradford Book.

Fodor, J. (1974). Special Sciences: The Disunity of Science as a Working Hypothesis. *Synthese*, 28, 97-115.

Fodor, J. (1975). *The language of thought*. New York: Crowell

Fodor, J. (1980). Methodological Solipsism considered as a research strategy in cognitive psychology. *Brain and Behavioural Sciences*, 3(1), 63-73.

Fodor, J. (1987). *Psychosemantics*. Cambridge: MIT Press.

Fresco, N. (2014). *Physical computation and cognitive science*. Berlin, Heidelberg: Springer-Verlag.

Frigg, R. (2012). Models in Science. Stanford Encyclopaedia of Philosophy. http://plato.stanford. edu/entries/models-science.

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends Cognitive Science*, 13, 293–301. doi:10.1016/j.tics.2009.04.005.

Friston K. (2011). Embodied inference: or 'I think therefore I am, if I am what I think'. In W. Tschacher & C Bergomi (Eds.), *The implications of embodiment (cognition and communication)* (pp. 89– 125). Exeter, UK: Imprint Academic.

Friston, K. (2013). Life as we know it. *Journal of the Royal Society, Interface*, 10. doi:10.1098/rsif.2013. 0475.

Froese, T. and Ziemke, T. (2009). Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial Intelligence*, 173, 466-500.

Gallagher, S. (2005). *How the body shapes the mind*. Oxford: Oxford University Press.

Gallistel, C.R. (1989a). Animal cognition: The representation of space, time and number. *Psychology Annual Reviews*, 40, 155–89.

Gallistel, C.R. (1989b). *The organization of learning*. Cambridge: MIT Press.

Garson, J. (2003). The Introduction of Information into neurobiology. *Philosophy of Science* 70(5), 926-936. doi: 10.1086/377378.

Gibson, J.J. (1966). *The senses considered as perceptual systems*. Boston: Houghton Mifflin.

Gibson, J.J. (1986). *The ecological approach to visual perception*. East Sussex: Psychology Press.

Gladziejewski, P. (2015). Predictive coding and representationalism. *Synthese*, *193*(2), 559–582.

Glass, D. H. (2007). Coherence measures and inference to the best explanation. Synthese, 157(3), 275-296.

Haimovici, S. (2013). A Problem for the Mechanistic Account of Computation. *Journal of Cognitive Science*, 14, 151-81.

Harbecke, J., & Shagrir, O. (2019). The role of the environment in computational explanation. *European Journal of Philosophy of science*, 9, 37. doi: 10.1007/s13194-019-0263-7.

Harman, D. (1988). Wide Functionalism. In S. Schiffer & S. Steele (Eds.), *Cognition and Representation*. Westview Press.

Haugeland, J. (1998). Mind Embodied and Embedded. In J. Haugeland (Ed.), *Having Thought: Essays in the Metaphysics of Mind* (pp. 233-267). Cambridge, MA: Harvard University Press.

Hohwy, J. (2013). *The predictive mind*. Oxford: Oxford University Press.

Hohwy, J. (2016). The self-evidencing brain. *Nous*, *50*(2), 259–285.

Hohwy, J., Roepstorff, A., Kriston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, 108(3), 687-701. doi: 10.1016/j.cognition.2008.05.010.

Hosoya, T., Baccus, S., Meister, M. (2005). Dynamic predictive coding by the retina. *Nature*, 436, 71-77.

Hurley, S. L. (1998). *Consciousness in Action*. Cambridge, MA: Harvard University Press.

Hutchins, E. (1995). *Cognition in the wild*. Cambridge: MIT Press.

Hutto, D., & Myin, E. (2013). *Radicalizing Enactivism: Basic minds without content*. Cambridge, MA: MIT Press.

Hutto, D., and Myin, E. (2017). *Radicalising Enactivism: Basic Minds without Content*. Cambridge, M.A: MIT Press.

Hutto, D., Myin, E., Peeters, A., and Zahnoun, F. (2019). The Cognitive Basis of Computation: Putting Computation In Its Place. In M. Colombo, and M. Sprevak (Eds.), *The Routledge Handbook of The Computational Mind* (pp.265-281). London: Routledge.

Irvine, E. (2014). Problems and Possibilities for an Empirically Informed Philosophy of Mind. In M. Sprevak & J. Kallestrup (Eds.), *New Waves in Philosophy of Mind* (pp. 185-207). Palgrave McMillian.

Isaac, A. (2013). Modeling without representation. Synthese, 190(16), 3611-3623. doi: 10.1007/s11229-012-0213-9.

Isaac, A. (2018a). Physical Computation: A Mechanistic Account. *The Philosophical Review* 127, (3), 426–431. doi:10.1215/00318108-6718882.

Isaac, A. (2018b). Embodied cognition as analog computation. *Reti, Saperi, Linguaggi: Italain Journal of Cognitive Sciences*. doi:10.12832/92298.

Kästner, L. (2017). *Philosophy of cognitive neuroscience, causal explanations, mechanisms and experimental manipulations*. Berlin, Boston: De Gruyter.

Kaplan, D. M. (2011). Explanation and description in computational neuroscience. *Synthese*, 183, 339-373.

Kaplan, D. M. (2012). How to demarcate the boundaries of cognition. *Biology and Philosophy*, 27, 545–570.

Kersten, L. (2014). Music and Cognitive Extension. *Empirical Musicology Review*, 9(3-4), 193-202.

Kersten, L. (2017a). A Mechanistic Account of Wide Computationalism. *The Review of Psychology and Philosophy*, 8(3), 501-517. doi: 10.1007/s13164-016-0322-3.

Kersten, L. (2017b). Extended music cognition. *Philosophical Psychology*, 30(8), 1078-1103.

Kersten, L., & Wilson, R. (2016). The Sound of Music: Externalist Style. *American Philosophical Quarterly*, 53(2), 139-154.

Kim, J. (2002). The Layered Model: Metaphysical Consideration. *Philosophical Explorations*, 5(1), 2–20.

Kirchhoff, M. D. (2015). Extended cognition & the causal-constitutive fallacy: In search for a diachronic and dynamical conception of constitution. *Philosophy and Phenomenological Research*, *90*(2), 320–360.

Kirchhoff, M. (2018). Predictive processing, perceiving and imagining: Is to perceive to imagine, or something close to it? *Philosophical studies*, 175, 751-767.

Kirchhoff, M. D., and Kiverstein, J. (2019). *Extended Consciousness and Predictive Processing: A Third Wave View*. Routledge.

Kirsch, D. (2009). Problem solving and situated cognition. In P. Robbins & M. Aydede (Eds.), *The Cambridge Handbook of Situated Cognition* (pp. 264–306). New York, NY: Cambridge University Press.

Kirsch, D., & Maglio, P. (1995). On distinguishing epistemic from pragmatic actions. *Cognitive Science*, 18, 513-549.

Klatzky, G.R. Golledge, G.J. Cicinelli, W.J. Pellegrino, and A.F. Fry. (1993). Nonvisual navigation by blind and sighted: Assessment of path integration ability. *Journal of Experimental Psychology: General,* 122(1), 73–91.

Krickel, B. (2018). Saving the mutual manipulability account of constitutive relevance. *Studies in History and Philosophy of Science Part A*, 68, 58-67.

Krueger, J. (2014). Varieties of extended emotions. *Phenomenology and the Cognitive Sciences*, 13(4), 533-555.

Kuokkanen, J. and Rusanen, A. (2018). Making Too Many Enemies: Hutto and Myin's attack on computationalism. *Philosophical Explorations*, 21(2), 282–294. doi:10.1080/13869795.2018.1477980.

Ladyman, J., Ross, D., Spurrett, D., and Collier, J. (2007). *Every Thing Must Go: Metaphysics Naturalized*. Oxford: Oxford University Press.

Lahav, O., and D. Mioduser. (2008). Haptic-feedback support for cognitive mapping of unknown spaces by people who are blind. *International Journal of Human-Computer Studies*, 66, 23–35.

Lakoff, G., & Johnson, M. (1980). Metaphors We Live By. Chicago, IL: University of Chicago Press.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought*. NY: Basic Books.

Leigh, R.J., and D. Zee. (2006). *The neurology of eye movements, 4th ed*. New York: Oxford University Press.

List, C. (2019). Levels: Descriptive, Explanatory, and Ontological. *Nous*, 53(4): 852-883.

Losonksy, M. (1995). Embedded systems vs individualism. *Minds and Machines*, 5, 357-371.

MacIver, M.A. (2009). Neuroethology: From morphological computation to planning. In P. Robbins and M. Aydede (Eds.), *The Cambridge Handbook of Situated Cognition* (pp.480–504). New York, Cambridge University Press.

Maturana, H. (1981). Autopoiesis. In M. Zeleny (Ed.), *Autopoiesis: A theory of living organization* (pp.21- 33). New York, Oxford: North Holland.

Marr, D. (1982). *Vision*. San Francisco: W.H. Freeman.

McClamrock, R. (1991). Marr's Three Levels: A Re-Evaluation. *Minds and Machines*, 1, 185-196.

Maturana, H., & Varela, F. J. (1980). *Autopoiesis and cognition: the realization of the living*. Dordrecht, Holland: Kluwer Academic Publishers.

Menary, R. (2007). *Cognitive integration: Mind and cognition unbounded*. Basingstoke: Palgrave Macmillan.

Menary, R. (2010a). Introduction to the special issue on 4E cognition. *Phenomenology and Cognitive Science*, 9, 459–463.

Menary, R. (ed.) (2010b). *The Extended Mind*. Cambridge, M.A.: MIT Press.

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

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

Milkowski, M. (2015). Computational mechanism and models of cognition. *Philosophia Scientiae*, 18(3), 1–14.

Milkowski, M. (2018). Morphological Computation: Nothing but Physical Computation. *Entropy*, 20(12), 942. doi:10.3390/e20120942.

Miłkowski M., Clowes R., Rucińska Z, Przegalińska A., Zawidzki T., Krueger J., Gies A., McGann M., Afeltowicz Ł., Wachowski W., Stjernberg F., Loughlin V. and Hohol, M. (2018). From Wide Cognition to Mechanisms: A Silent Revolution. *Front. Psychol.* 9,

2393. doi: 10.3389/fpsyg.2018.02393

Millhouse, T. (2019). A Simplicity Crieterion for Physical Computation. *British Journal for the Philosophy of Science*, 70, 153-178. doi: 10.1093/bjps/axx046.

Newell, A. (1990). *Theories of Unified Cognition*. Cambridge, MA: Harvard University Press.

Newell, A., and Simon, H. (1976). Computer Science as Empirical Inquiry: Symbols and Search. *Communications of the ACM*, 19(3), 113-126.

Noe, A. (2004). *Action in Perception*. MIT Press.

Noe, A. (2009). *Out of our heads*. New York: Hill and Wang.

Nowakowski, P. (2017). Bodily Processing: The Role of Morphological Computation. *Entropy*, *19*(7), 295.

Oppenheim, P., and Putnam, H. (1958). Unity of Science as a Working Hypothesis. In H. Feigl, M. Scriven, & G. Maxwell (Eds.), *Minnesota Studies in the Philosophy of Science*, (Vol. 2. pp.3–36). Minneapolis: University of Minnesota Press.

O'Regan, J. & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 25(4), 883–975.

Pezzulo, G., Rigoli, F., and Chersi, F. (2013). The mixed instrumental controller: Using value of information to combine habitual choice and mental simulation. *Frontiers in Psychology*, 4, 92. doi: 10.3389/fpsyg.2013.00092.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA: Morgan Kaufmann Publishers.

Piccinini, G. (2007). Computing mechanisms. *Philosophy of Science*, 74(4), 501–526.

Piccinini, G. (2008). Computation without Representation. *Philosophical Studies*, 137(20), 205- 241.

Piccinini, G. (2010). The Mind as Neural Software? Understanding Functionalism, Computationalism, and Computational Functionalism. *Philosophy and Phenomenological Research*, 81(2), 269-311.

Piccinini, G. (2015). *Physical Computation*. Oxford: Oxford University Press.

Piccinini, G. (2018). Computational mechanisms. In S. Glennan & P. Illari (Eds.), *The Routledge Handbook of Mechanisms and Mechanical Philosophy* (pp.435-446). London: Routledge Taylor & Francis Group.

Piccinini, G., and Scarantino, A. (2011). Information processing, computation, and cognition. *Journal of Biological Physics* 37(1), 1–38. doi: 10.1007/s10867-010-9195-3.

Piccinini, G., and S., Bahar. (2013). Neural Computation and the Computational Theory of Cognition. *Cognitive Science*, 37 (3): 453–488. doi: 10.1111/cogs.2013.37.issue-3.

Polger, T., and Shapiro, L. (2016). *The multiple realization book*. Oxford: Oxford University Press.

Potochnik, A. (2010). Levels of explanation reconceived. *Philosophy of Science*, 77(1), 59-72.

Potochnik, A, and McGill, B. (2012). The Limitations of Hierarchical Organization. *Philosophy of Science*, 79(1), 120-140.

Potochnik, A. (forthcoming). Our World Isn't Organized into Levels. In Dan Brooks Brooks, James DiFrisco & William C. Wimsatt (eds.), *Levels of Organization in Biology*. Cambridge, USA: MIT Press.

Putnam, H. (1975). Philosophy and Our Mental Life. In H. Putnam (Ed.), *Mind, Language and Reality: Philosophical Papers (Vol. 2)*. Cambridge: Cambridge University Press.

Psillos, S. (1999). *Scientific realism*: How science tracks truth. London: Routledge.

Pylyshyn, Z. (1984). *Computation and cognition*. Cambridge: MIT Press.

Ramstead, M., Kirchhoff, M., Constant, A., and Friston, K. (2019a). Multiscale integration: beyond internalism and externalism. *Synthese*. doi: 10.1007/s11229-019-02115-x.

Ramstead, M., Constant, A., Badcock, P., & Friston, K. (2019b). Variational ecology and the physics of minds. *Physics of Life Reviews*, 1–18.

Raleigh, T. (2018). Tolerant enactivist cognitive science. *Philosophical Explorations*, 21(2), 226–244.

Reichenbach, H. (1938). *Experience and prediction*. Chicago: The University of Chicago Press.

Ricciardi, E., D. Bonino, L. Sani, T. Vecchi, M. Guazzelli, V. L. Haxby, L. Fadiga, and P. Pietrini. (2009). Do we really need vision? How blind people 'See' the actions of others. *The Journal of Neuroscience* 29(31), 9719–9724.

Ritchie, B., and Piccinini, G. (2018). Computational Implementation. In M. Sprevak & M. Colombo (Eds.), *Routledge Handbook of the Computational Mind* (pp.192-204). London: Routledge.

Robinson, D.A. (1989). Integrating with neurons. *Annual Review of Neuroscience*, 12, 33–45.

Romero, F. (2015). Why there Isn't inter-level causation in mechanisms. *Synthese*, 192(11), 3731-3755.

Ross, D., and Ladyman, J. (2010). The Alleged Coupling-Constitution Fallacy. In R. Menary (Ed.), *The Extended Mind* (pp. 155–166). Cambridge, MA: The MIT Press.

Rowlands, M. (1999). *The body in mind*. Cambridge: Cambridge University Press.

Rueger, A., and McGivern, P. (2010). Hierarchies and Levels of Reality. *Synthese*, 176(3), 379–97.

Rupert, R. (2004). Challenges to the hypothesis of extended cognition. *Journal of Philosophy*, 101, 389–428.

Rupert, R. (2009). *Cognitive systems and the extended mind*. New York: Oxford University Press.

Rutkowska, J. C. (1993). *The Computational Infant*. Brighton: Harvester.

Scheutz, M. (1999). When physical systems realize functions …. *Minds and Machines*, 9(2), 161–196.

Scheutz, M. (2001). Causal versus computational complexity. *Minds and Machines*, 11(4), 534–566.

Schweizer, P. (2019). Computation in Physical Systems: A Normative Mapping Account. In M. Vincenzo, D. Alfonso, & D. Berkich, *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence,* (pp.27-47). Springer, Cham.

Searle, J. (1980). Minds, brains and programs. *The Behavioral and Brain Sciences,* 3(3), 417–24.

Searle, J. (1983). *Intentionality*. Cambridge: Cambridge University Press.

Segal, G. (1997). Review of: Cartesian psychology and physical minds: Individualism and the sciences of mind by Robert A. Wilson. *The British Journal for the Philosophy of Science* 48(1), 151–156.

Sekular, R., and R. Blake. (1990). *Perception*, 2nd ed. McGraw-Hill: New York.

Shagrir, O. (2001). Content, computation and externalism. *Mind, 110*(438), 369–400.

Shagrir, O. (2006). Why We View the Brain as A Computer. *Synthese*, 153, 393-416.

Shagrir, O. (forthcoming). In defense of the semantic view of computation. *Synthese*.

Shapiro, L. (2011). *Embodied Cognition*. Routledge Press.

Shapiro, L. (2019). Flesh matters: The body in cognition. *Mind and Language*, 34, 3-20.

Smart, P. (2018). Human-extended machine cognition. *Cognitive System Research*, (49), 9-23.

Sprevak, M. (2009). Extended Cognition and Functionalism. *Journal of Philosophy*, 106, 503-527.

Sprevak, Mark. (2010a). Computation, individuation, and the received view on representation. *Studies in History and Philosophy of Science,* 41, 260–270. doi: 10.1016/j.shpsa.2010.07.008.

Sprevak, M. (2010b). Inference to the hypothesis of extended cognitoin. *Studies in History and Philosophy of Science*, 41, 353–362.

Sprevak, M. (2018). Triviality arguments about computational implementation. In M. Sprevak & M. Colombo (Eds.), *Routledge Handbook of the Computational Mind* (pp.175-191). London: Routledge.

Sprevak, M., and Colombo, M (Eds.). (2018). *Routledge Handbook of the Computational Mind*. London: Routledge.

Stich, S. (1983). *From folk psychology to cognitive science*. Cambridge: MIT Press.

Sun, Ron, Coward, Andrew, and Zenzen, Michael. (2005). On Levels of Cognitive Modeling. *Philosophical Psychology,* 18(5), 613–637.

Sutton, J. (2010). Exograms and interdisciplinary: History, the extended mind, and the civilizing process. In R. Menary (Ed.), *The Extended Mind* (pp. 189–225). Cambridge, MA: The MIT Press.

Summerfield, C., Monti, J., Trittschuh, E., Mesulam, M., Egner, T. (2008). Neural repetition suppression reflects fulfuilled perceptual expectations. *Natural Neuroscience*, 11(9), 1004-1006.

Thagard, P. (2007). Coherence, Truth, and the Development of Scientific Knowledge. *Philosophy of Science*, 74, 28–47.

Thompson, E. (2005). Sensorimotor subjectivity and the enactive approach to experience. *Phenomenology and the Cognitive Sciences*, 4, 407-427.

Thompson, E. (2007). *Mind in Life: Biology, phenomenology, and the sciences of mind*. Cambridge, MA: Harvard University Press.

Thompson, E., & Stapleton, M. (2009). Making Sense of Sense-Making: Reflections on Enactive and Extended Mind Theories. *Topoi*, 28, 23-30.

Toulmin, S. (1975). Concepts of function and mechanism in medicine and medical science. In H. Tristram Engelhardt, Jr., & S.F. Spikers, *Evaluation and Explanation in the Biomedical Sciences* (pp.51-66). Dordrecht: D. Rediel.

van Gelder, T. (1995). What Might Cognition be if not Computation? *Journal of Philosophy*, 92 (7), 345–381.

van Fraassen, B. C. (1980). *The scientific image.* Oxford: Oxford University Press.

Varela, F., E. Thompson, and E. Rosch. (1991). *The embodied mind: cognitive science and human experience*. Cambridge: MIT Press.

Viger, C. (2000). Where Do Dennett's Stances Stand?. In D. Ross, A. Brook, & D. Thompson (Eds.), *Dennett's Philosophy: A Comprehensive Assessment* (pp.131-145). Cambridge, Mass.: MIT Press.

Villabolos, M., & Ward, D. (2015). Living systems: Autopoiesis, autonomy and enaction. *Philosophy and Technology*, 28(2), 225-239.

Villalobos, M., & Dewhurst, J. (2016). Cognition, Computation and Dynamic Systems: Possible Ways of Theoretical Integration. *Límite*, 11(36), 20-31.

Villabolos, M., & Silverman, D. (2017). Extended Functionalism, radical enactivism, and the autopoietic theory of cognition: prospects for a full revolution in cognitive science. *Phenomenology and Cognitive Science*, 17, 719-738.

Villalobos, M., & Dewhurst, J. (2018). Enactive autonomy in computational systems. *Synthese*, 195, 1891-1908.

Ward, D. & Stapleton M. (2012). Es are good: Cognition as enacted, embodied, embedded, affective and extended. *Consciousness in interaction*, 89-104.

Webb, B. & Harrison, R. (2000). Integrating sensorimotor systems in a robot model of cricket behaviour, sensor fusion and decentralized control in robotic systems III. *Proceedings of the Society of Photo-Optical Instrumentation Engineers*, 4196, 113-124.

Webb, B. (2008). Using Robots to Understand Animal Behavior. In H. J. Brockmann, Timothy J. Roper, Marc Naguib, Katherine E. Wynne-Edwards, Chris Barnard, & John C. Mitani (Eds.), *Advances in the Study of Behavior. Volume 38* (pp. 1–58). Elsevier.

Wells, A. J. (1998). Turing analysis of computation and theories of cognitive architecture. *Cognition*, 22(3), 269-294.

Wheeler, M. (2010a). Minds, things and materiality. In C. Renfrew, L. Malafouris (Eds.), *The cognitive life of things: recasting the boundaries of the mind*. McDonald Institute for Archaeological Research Publications, Cambridge.

Wheeler, M. (2010b). In defense of extended functionalism. In R. Menary (Ed.), *The Extended Mind* (pp. 245-270). Cambridge, MA: The MIT Press.

Wiesse, W. (2018). *Experience Wholeness: Integrating Insights from Gestalt Theory, Cognitive Neuroscience, and Predictive Processing*. MIT Press.

Wilson, R. A. (1994). Wide computationalism. *Mind*, 103(4), 351–372.

Wilson, R. A|. (1995). *Cartesian psychology and physical minds: Individualism and the sciences of the minds.* Cambridge: Cambridge University Press.

Wilson, R. A. (2004). *Boundaries of the mind: The individual in the fragile sciences*. Cambridge: Cambridge University Press.

Wilson, R. A. (2010). Extended vision. In (eds.) N. Gangopadhyay, M. Madary, and F. Spicer, *Perception, action and consciousness* (pp.277–290). New York: Oxford University Press.

Wilson, R. A., & Foglia, L. (2015). Embodied cognition. In E. N. Zalta (Ed.), *The Stanford Encyclopaedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, 9, 625–636.

Wimsatt, W. (1976). Reductionism, Levels of Organization, and the Mind-Body Problem. In G. Globus, G. Maxwell & I. Savodnik, *Consciousness and the Brain* (pp. 205-267). Spring US.

Wimsatt, W. (1994). The Ontology of Complex Systems: Levels of Organization, Perspectives, and Causal Thickets. *Canadian Journal of Philosophy*, 20, 207-274.

Yu, Paul and Fuller, Gary. (1986). A Critique of Dennett. *Synthese*, 66, 453-476.