Mind, Meaning and Mechanism:

The role of representation in explanations of cognition

Jonny Lee

PhD Philosophy

University of Sussex

May 2019

Declaration

I hereby declare that this thesis has not been and will not be, submitted in whole or in part to another University for the award of any other degree.

Signature: J Lee

Abstract

Representation remains ubiquitous in scientific explanations of cognition. At the same time, philosophers continue to question what, if anything, representation contributes to cognitive science. Whilst some practically define cognition in terms of operations performed over representations, others take the very concept of subpersonal representation to be incoherent. Despite the longstanding debate, this thesis argues that we now possess the resources needed to provide a satisfactory account of cognitive representation, taking the challenges raised by eliminativists as an opportunity to refine our understanding of its explanatory role.

I defend a 'mechanistic approach' that presents representational explanations as a kind of mechanistic explanation. This approach has three main parts which pull together several promising threads in the literature to form an original account. The first part is a mechanistic interpretation of explanations in cognitive science. This interpretation provides insight into the sort of *explanans* cognitive science offers, and the sort of theoretical entity a cognitive representation might be. The second part is acceptance of the increasingly popular notion of 'structural', 'simulation' or 'surrogate representation' ('S-representation'). This notion provides an empirically plausible and well-defined set of functional criteria for a genuinely representational mechanism, drawing an illuminating analogy between the functional role of a possible cognitive mechanism and a type of ordinary representation. The third part is a 'mechanistic account of content'. This provides a naturalistically respectable foundation for representation's paradigmatic semantic properties at the subpersonal level. Overall, the mechanistic approach ensures that representation ascriptions play a robust role anchored in our dominant explanatory framework. From this perspective, the future of cognitive representation looks bright.

Contents

Acknowledgements

I would like to extend thanks to the many people who generously contributed to this thesis. Above all, I am grateful to Sarah Sawyer for her invaluable supervision. This PhD would not have been possible without her incredible support over the years. My gratitude also extends to Ron Chrisley, for his constructive discussions and for the many academic opportunities he offered me beyond this thesis. I am also hugely appreciative of Mark Sprevak for his continued guidance, and for his encouragement to pursue a PhD in the first place.

I have had the great pleasure of working with many other graduate students and researchers throughout this project. Particular thanks go to Dan Calder, Joe Dewhurst, Adrian Downey, Becky Millar, Kate Nave, Nina Poth and Henry Schiller for their crucial conversations on all things representation related, and for their constructive feedback on earlier drafts.

This thesis was made possible by an award from the Consortium of the Humanities and the Arts South-East England, and from the Royal Institute of Philosophy (Jacobsen Studentship).

Finally, but by no means least, thanks go to my family and my partner Abbey for their unwavering support. They are the most important people in my world.

This thesis is dedicated to the memory of my grandmother Margaret Lee. She was an extraordinarily kind woman and her gentle encouragement of my academic studies is a continued inspiration.

Thesis Introduction

Familiar forms of representation play a crucial role in our daily lives. A successful road trip to a foreign city depends on a map correctly mirroring the highways and byways that we intend to drive. A successful cinema outing depends on the film listings correctly reflecting the start time of the film we wish to see. A successful soufflé depends on our recipe correctly identifying the culinary instructions we need to follow. In each case, representations help us to smoothly coordinate with each other and the world.

Less familiar forms of representation are also sometimes thought to play a crucial role in our daily lives. Scientists frequently appeal to the idea that the brain constructs, stores and manipulates internal 'cognitive representations' to explain the success (and failure) of essential cognitive capacities like perception, object recognition and navigation. And yet, philosophers continue to dispute what exactly, if anything, representation contributes to cognitive science. Whilst some practically define cognition in terms of operations performed over cognitive representations, others argue that the concept of cognitive representation is incoherent and should be exorcised from scientific practice. This discrepancy is largely driven by a lack of clarity over the criteria for the justified ascription of cognitive representation at the 'subpersonal level'—roughly, the level of brain processes and events. A vital question thus remains: what role, if any, does cognitive representation play in explanations of cognition?

This thesis addresses that question. I argue that we now possess the resources needed to provide a satisfactory account of cognitive representation, taking sceptical challenges as an opportunity to refine our understanding of its explanatory role. I acknowledge that

many historical justifications for ascribing representation are unsatisfactory. However, I maintain that when characterised appropriately, cognitive representation informatively describes the underlying causes of cognition according to some of our most promising scientific theories. To this end, I defend a 'mechanistic approach' that characterises cognitive representation as a type of cognitive mechanism. Accordingly, representational explanations of cognition are a type of mechanistic explanation of cognition.

My mechanistic approach has three main parts which pull together several promising threads in the existing literature. The first part is a mechanistic interpretation of explanations in cognitive science. This interpretation refines our understanding of what sort of *explanans* cognitive science strives for, and what sort of theoretical entity a cognitive representation might be. The second part is an argument in favour of the increasingly popular account of 'structural', 'simulation' or 'surrogate representation' ('S-representation'). This account provides an empirically plausible set of functional criteria for a genuinely representational mechanism. In doing so, it draws an illuminating analogy between the functional role of a cognitive mechanism and a type of ordinary representation that includes cartographic maps and scientific models. Both ordinary and cognitive S-representations are characterised by the way they mirror the structure of the world and guide the actions of their consumers accordingly. The third part is a mechanistic account of 'representational content' (what a representation 'is about'). This provides a naturalistic foundation for the semantic properties of representation at the subpersonal level.

Overall, the mechanistic approach ensures that representation ascriptions play a significant role in our explanations of cognition. Though not an empirical theory itself,

the mechanistic approach does suggest that cognitive representation will play a part in some of our best theories in contemporary cognitive science, including 'predictive processing' as well as cognitive neuroscience more generally.

The thesis proceeds as follows. **Chapter 1** discusses 'representationalism' and 'eliminativism'—where, roughly, the former position affirms the explanatory value of cognitive representation whilst the latter position denies it. I offer two important distinctions that help to clarify the commitments of these broad positions. The first distinction is between two kinds of eliminativism. One kind takes cognitive representation to be eliminated by our best available theory of cognition whilst the other takes cognitive representation to be eliminated because it involves a category error. I argue against the latter, leaving open the possibility that cognitive representation plays an explanatory role in cognitive science. The second distinction is between the explanatory role of cognitive representation on the one hand and the ontological status of cognitive representation on the other. Assessing representation's explanatory role comes apart from evaluating whether cognitive representations are real in any strong, metaphysical sense. I argue that the former is of primary importance to cognitive science.

Chapter 2 examines what sort of theoretical entity cognitive representation might be if it is to play an explanatory role in cognitive science. I argue that for an entity to count as a cognitive representation in any interesting sense, it must function in a way that is distinctly representation-like. I present a mechanistic framework for understanding explanations in cognitive science and suggest that we understand cognitive representation as a type of mechanism characterised by its distinctly representation-like role in realising a cognitive capacity. I close by evaluating three notions of cognitive representation that

are common (implicitly or explicitly) within the scientific and philosophical literature. These are: 'receptor representation', 'action-oriented representation' and 'intentional stance representation'. I argue that all three are unsatisfactory as far as securing representation's explanatory role in cognitive science is concerned.

Chapter 3 continues to evaluate common ways of understanding cognitive representation, turning to the relationship between computation and representation. It is often supposed that a computational approach to cognition implies representation. I argue that this is false because computation and representation are distinct functional kinds. Physical computation does not presuppose representation and, furthermore, there is nothing special about the structure of computational explanations of cognition that straightforwardly implies representation. Along the way, I defend a mechanistic view of computational individuation.

Chapter 4 begins the constructive part of the thesis. I outline and defend a version of the S-representation account. Once embedded within a mechanistic framework, this account suggests that a cognitive representation is a cognitive mechanism whose functional role resembles a type of ordinary representation characterised by the way a vehicle structurally corresponds to its 'target' on behalf of its consumer. I argue that such 'representational mechanisms' find plausible vindication in the empirical literature. I also respond to several objections, and in doing so, bolster what I take to be the strongest version of the S-representation account.

Chapter 5 continues the constructive project by discussing how to think about content in representational mechanisms. I articulate two 'problems of content' that any complete

account of representation must overcome. The first problem is the ‘hard problem of content’ which concerns the legitimacy of positing content at the subpersonal level. I argue that if a mechanism meets the functional criteria for S-representation then there is sufficient justification for positing ‘correctness conditions’, and thus content, at the subpersonal level. The second problem is the ‘content determination problem’ which concerns how a given representation comes to have the content that it does: what makes a token representation about *x* and not *y*? I defend the ‘mechanistic account of content’. According to this account, the content of a token representation refers to the state of affairs that would need to be actual for it to realise a cognitive capacity and generate behavioural success. This provides an alternative to traditional ‘causal-historical theories’ that focus on the evolution, learning history or other aspects of a system’s etiology, in favour of a view that places the action-affordances of a mechanism front and centre. Ultimately, I argue that the S-representation account, especially once embedded within a mechanistic framework, dissolves worries surrounding the semantic properties of cognitive representation, securing a firm footing for representation to play an effective role in explanations of cognition.

Chapter 1

Representationalism and Eliminativism[1]

1. Introduction

This thesis concerns the longstanding debate over the role of representation in cognitive science. The issue is often framed as a conflict between representationalists on the one hand and eliminativists on the other, where representationalists are understood as those who affirm the value of subpersonal cognitive representation in scientific explanations of cognition, and eliminativists are understood as those who deny it. However, characterising the debate in these terms fails to fully capture the complexity of the conceptual landscape. This chapter elucidates the conceptual landscape by drawing attention to two distinctions that are typically overlooked. This sets the scene for the remainder of the thesis by clarifying its primary focus.

The first distinction is between two varieties of eliminativism about subpersonal cognitive representation. I call these 'a posteriori eliminativism' and 'a priori eliminativism'. According to the former, we should be eliminativists about cognitive representation because our best scientific theory of cognition makes no reference to such entities. According to the latter, we should be eliminativists about cognitive representation because the notion involves a category error, and hence the ascription of cognitive representation could not be explanatory, no matter what an empirical investigation into the nature of cognition reveals. This chapter argues against a priori eliminativism, thereby

[1] Portions of this chapter appear in Lee (2017).

leaving open the question of whether appealing to cognitive representation is explanatory in our best theory of cognition.

The second distinction is between the explanatory role of cognitive representation and the ultimate ontological status of cognitive representation. The question of whether and how cognitive representation contributes to explanations in cognitive science is distinct from the question of whether there are *really* any such things as cognitive representations. The former question concerns the theoretical posits of scientific theories and whether such posits count as representations. The latter question concerns the ontological commitments of scientific theories and depends on wider suppositions about how to view the existence of unobservables. For example, it may be that talk of cognitive representation *is* explanatory, but that would not by itself tell us whether such entities exist in a strong, metaphysical sense. The explanatory role of subpersonal cognitive representation will be my primary concern from **chapter 2** onwards.

The chapter proceeds as follows. **Section 2** introduces the notion of 'cognitive representation' and outlines the difference between 'representationalism' and 'eliminativism'. **Section 3** develops this discussion by establishing our first major distinction—that between a posteriori eliminativism and a priori eliminativism. It then raises two objections to a priori eliminativism. **Section 4** establishes our second major distinction—that between the explanatory role of cognitive representation and its ultimate ontological status. It then builds on this distinction by noting the difference between the 'scientific path' and the 'metaphysical path' toward cognitive representation, borrowing from a suggestion by Dennett (1991). The section concludes with a reflection on the

consequences of separating these paths for a mechanistic approach to representation, of the sort pursued in subsequent chapters.

2.1 What is representation?

Maps, models and portraits are just some of the familiar representations that play a significant role in our everyday lives, as are the written sentences that you will read throughout this thesis. But what do these objects have in common? Philosophers are often quick to note two crucial features of familiar representations: (i) they are physical things, and (ii) they have semantic properties, intentionality or content (for recent examples see, Egan, 2014, p. 115; Fresco, 2014, p. 16; Shea, 2018, p. 5). Maps, models and portraits are all physical entities, and like other physical entities, they have causal powers. At the same time, maps, models and portraits are also *about* something. Moreover, these two crucial features are related: familiar representations appear to have causal powers in virtue of their semantic properties. This is evident when we observe the importance of 'correctness' or 'incorrectness' in determining the outcome of behaviour caused by a representation.

The success or failure of behaviour in ordinary life is often explained, in part, by appeal to the correctness (truth/accuracy etc.) or incorrectness (falsity/inaccuracy etc.) of a representation that is involved in causing behaviour. For instance, a mountaineer's successful ascent of a mountain might be explained, in part, by the fact that their map accurately mirrors the topology of the mountain they were climbing. In other words, the mountaineer's behavioural outcome is partly explained by correct representation. Likewise, a mountaineer's failure to ascend a mountain might be explained, in part, by the fact that their map inaccurately mirrors the topology of the mountain they were climbing. In other words, the mountaineer's behavioural outcome is partly explained by

incorrect representation. In each case, the map is required to meet certain conditions to count as correct and cause the mountaineer to succeed in climbing. All familiar representations have such 'correctness conditions': states of affairs under which they count as correct (true, accurate etc.). I use 'correctness conditions' here as an umbrella term for different semantic measures of success, encompassing the common notions of 'truth conditions' and 'accuracy conditions' (truth conditions suggest that correctness is all or nothing, whilst accuracy conditions suggest that correctness is a matter of degree). We will return to the idea of correctness conditions throughout this thesis.

We are starting to see that familiar representations, like maps, models and portraits, share a distinctive kind of 'functional role' characterised by their semantic nature (Haugeland, 1991; Ramsey, 2007). In this way, representation is a 'functional kind'; a genus defined by the functional role of its members (akin to pumps, pulleys, filters or indicators). Function is intended in the teleological sense, whereby, if *x* has the function to *y*, it has the 'purpose' or 'end' to *y* (more on function in **chapter 2**). In other words, maps, models and portraits are all representations because they all share the purpose or end of representing.

We can be a little more precise in capturing what is distinctive about representation as a functional kind, keeping our examples of familiar representation in view. At first pass, to represent is to stand-in for, simulate or replicate some entity (object, state, process, activity etc.) on behalf of some 'user', 'interpreter' or 'consumer' (the latter term is common in philosophical parlance), in a way that implies correctness conditions. Ordinary maps, models and portraits all meet this description. For instance, a map might stand-in for a mountain range on behalf of a mountaineer, such that the map counts as

correct when it appropriately mirrors the topology or other features of the mountain range and incorrect when it does not. Following this description, representation can be thought of as a triadic relation (Peirce, 1998, p. 478). The relation holds between a 'representational vehicle' (for example, a map), a 'represented entity' (for example, the Himalayas), and a 'representation consumer' (for example, a mountaineer). Future chapters will build on this understanding of representation (for related discussion see Haugeland, 1991; von Eckardt, 1993; Menary, 2007; Ramsey, 2007).

Maps, models and portraits are familiar to everyday life, but cognitive science frequently appeals to a special kind of representation: 'cognitive representation'. Cognitive science, in its broadest sense, is the interdisciplinary scientific study of cognition encompassing psychology, artificial intelligence, cognitive neuroscience, and other sciences of the mind. Across these disciplines, representation is thought to help explain a diverse range of 'cognitive capacities'—roughly, those capacities associated with the intelligent behaviours of minded creatures. These include visual-spatial perception, object recognition and categorisation, self-relative position tracking, attentional control, and language comprehension.[2]

Unlike ordinary maps, models and portraits, cognitive representations are thought to be located internally to a cognitive system, usually within the brain. It is often supposed that a cognitive system is able to operate within an uncertain and complex world by performing operations over these internal representations. In this way, cognitive

[2] The literature often refers to 'cognitive tasks'. I take 'capacities' and 'tasks' to be synonymous, each reflecting the primary *explananda* of cognitive science encompassing, as Lakoff & Johnson would have it, 'any mental operations and structures that are involved in language, meaning, perception, conceptual systems, and reason' (1999, p. 12). I opt for the term 'capacities' where possible. This is because the term is common in the mechanism literature with which I wish to stress continuity.

representation is thought to help 'mediate' between the world and the actions of a cognitive system, much like ordinary representations help mediate between the world and the actions of an agent. According to one popular characterisation, a cognitive representation 'is any internal state that mediates or plays a mediating role between a system's inputs and outputs in virtue of that state's semantic content' (Dietrich & Markman, 2003, p. 97).[3] Malafouris offers a similar characterisation:

> The idea of representation furnished a simple mechanism by which we could feed our cognitive apparatus with facts and information from the "external world"; it also suggested how we materialize and externalize our mental contents by way of behavioural output to the world. (2013, pp. 25-26)

For instance, representation has been thought, by many, to help explain how the brain processes visual percepts of distal objects: given some initial, relatively impoverished retinal stimulation, the brain employs internally coded rules and representations in order to process a 3D image that reconstructs the causes of that stimulation (for example, see Rescorla, 2015). Or to take another example, representation has been thought, by many, to help explain how an organism stores, plans and rehearses navigational routes through its spatial environment, specifically, by constructing, storing and exploiting internal 'maps' of its surroundings (for example, see Bechtel, 2016). Such representations are typically posited at the 'subpersonal' level.

Representation can be ascribed to cognitive systems at both the 'personal' and 'subpersonal' levels. In the book in which he coins the terms, Dennett summarises the personal level as the 'level of people and their sensations and activities' (1969, p. 93). Personal level representations are those that are ascribable to individual agents, and in

[3]Although, we shall see later that the notion of 'semantic content' is somewhat ambiguous, and mere 'mediation' is not sufficient for representation (see **chapter 2**).

some cases, by extension, to groups of such agents. These include linguistic utterances, explicit beliefs and desires, and conscious percepts. There is some dispute over just what counts as representational at the personal level. For instance, some question whether conscious percepts are representational in nature (for example, see Hutto & Myin, 2013, chapter 6). Nonetheless, few would question all paradigmatic personal level representations.[4] More controversial, as we shall soon see, is the notion of subpersonal representation. As Dennett summarises, the subpersonal level is the 'level of brains and events in the nervous system' (1969, p. 93). Subpersonal representations are those thought to be ascribable to parts of agents, usually parts of the brain, with underlying neural structures and activity serving as representational vehicles.[5] Subpersonal processes that are thought to involve representation are paradigmatically non-deliberative, and not accessible to conscious awareness.

The exact meaning and value of the personal vs. subpersonal distinction is controversial (for discussion, see Drayson, 2012, 2014). Nonetheless, the difference between personal and subpersonal representation is sufficiently transparent and useful for our purposes. This thesis focusses on explanation at the subpersonal level, and so, 'cognitive representation' will hereafter refer to internal representation posited at the subpersonal level. Note that 'cognitive representation' is often synonymous with 'mental

[4]Miłkowski (2015a) proposes 'semantic nihilism' as a hypothetical position that denies all representation at every level. He observes that semantic nihilism is self-defeating, as an argument for its truth would rely on premises with correctness conditions (hence, representation): 'it is a minimal requirement for rational argumentation in philosophy; one has to assume that one's statements can be truth-bearers.' (p. 74). Therefore, at least natural language representations exist. Miłkowski raises semantic nihilism because he thinks 'radical enactivist cognition', a candidate case for what I label 'a priori eliminativism' in **section 3** below, threatens to slide into this absurd position by raising the standards for correctness conditions exceedingly high.

[5] The '4E cognition' movement has widened the potential scope of cognitive vehicles to encompass aspects of both body and world (we will return to 4E in **section 3.2** below and in future chapters). 4E invites the possibility that some cognitive representations are realised, in part, by vehicles external to the brain. This thesis will primarily concern theories that posit cognitive representations realised in the brain, but nothing major hinges on extending the scope of possible vehicles to include extra-neural entities.

representation'. However, I use the former term where possible as 'mental representation' sometimes carries connotations of consciousness, folk psychology or the personal level. The term 'cognitive' also emphasises the relationship between representation and theories in cognitive science, which is of central importance to this thesis.

2.2 Representationalism and eliminativism

Despite the ubiquity of representation-talk throughout cognitive science, there remains considerable disagreement over whether and how representation contributes to explanations of cognition. The debate is often framed as a clash between 'representationalists', who affirm the explanatory value of appeals to subpersonal cognitive representation, and 'eliminativists', who deny the explanatory value of appeals to subpersonal cognitive representation. In what follows, I will unpack this dichotomy further before turning to examine its limitations for expressing the conceptual landscape.

Representationalism states that representation ascriptions are of value in (at least some) scientific explanations of cognition, taking our best theory of some phenomenon to posit subpersonal cognitive representation. Different versions of representationalism vary in their scope. Some representationalists take representation to play a key role in explaining all or almost all of cognition. For example, the traditional 'cognitivist' paradigm models all or almost all of cognition in terms of computational operations performed over symbolic representations realised within the brain (Fodor, 1975). Such positions fall under the bracket of what I dub 'global representationalism'. Other representationalists hold only that representation plays a key role in explaining certain domains of cognition. For example, a representationalist might grant that basic motor control in response to one's present environment does not require representations, whilst maintaining that more

complex capacities such as counter-factual reasoning do (for related discussion, see Clark & Toribio, 1994). Such positions fall under the bracket of what I dub 'local representationalism'. 'Global representationalism' and 'local representationalism' mark two ends of a continuum, with different proponents asserting the scope of representation's value in scientific explanations to a greater or lesser extent.

Eliminativism states that representation ascriptions are not of value in (at least some) explanations of cognition, taking our best theory of some phenomenon to 'eliminate' subpersonal cognitive representation. This understanding of eliminativism is broader than traditional 'eliminative materialism' which specifically targets the mental states of folk psychology, like beliefs and desires, suggesting such entities have no place in our best scientific explanations of cognition (for example, see Churchland, 1981; see **chapter 3** for more on folk psychology). Eliminativism, as I present it here, encompasses eliminative materialism but also includes positions that seek to eliminate cognitive representation in other guises (such as map-like representations; see **chapter 4**).

Mirroring representationalism, different versions of eliminativism vary in scope. Some eliminativists deny that cognitive representation plays a key role in any explanation of cognition. For example, some proponents of dynamical systems theory, such as van Gelder (1995) and Chemero (2009), claim that representation has no role to play within our best (dynamical) theories of cognition (more on this view below). Such positions fall under the bracket of what I dub 'global eliminativism'. Other eliminativists deny only that representation plays a key role in explaining certain domains of cognition. For example, Orlandi (2014) advocates a non-representational 'ecological view' of visual processing but does not necessarily deny the importance of representation in explaining

other phenomena (more on this view below). Such positions fall under the bracket of what I dub 'local eliminativism'. 'Global eliminativism' and 'local eliminativism' also mark two ends of a continuum, with different proponents denying the efficacy of cognitive representation to a greater or lesser extent. Notice that if one is only a representationalist with respect to certain domains of cognition then one is an eliminativist with respect to those other domains of cognition which are thought not to involve representation. In this way, local representationalism implies local eliminativism (and vice versa). We will return to this point in **section 3.2** below.

Though the representationalism/eliminativism dichotomy does reflect the broad contours of the debate over representation's role in cognitive science, we must be careful to respect two further distinctions that it fails to underscore: the distinction between a priori and a posteriori eliminativism, and the distinction between the explanatory role and ontological status of subpersonal cognitive representation. Visiting these two distinctions will help to clarify the conceptual landscape and the topic of this thesis. The first of these two distinctions captures the fact that the global and local categories of eliminativism are not the only significant ones for appreciating diverging attitudes held across the eliminativist continuum.

3.1 Two kinds of eliminativism

There are two emblematic tendencies that fall under the broad banner of eliminativism. 'A posteriori eliminativism' is the view that cognitive representation is eliminated by our best available scientific theory. This can take the form of either local eliminativism or global eliminativism because the a posteriori eliminativist can hold that subpersonal cognitive representation is eliminated from our best scientific theory of either some or all

of cognition. ‘A priori eliminativism’ is the view that cognitive representation is eliminated because it involves a category error. This is a form of global eliminativism because the a priori eliminativist holds that the very notion of subpersonal cognitive representation is incoherent and so could not explain any cognitive phenomenon, no matter what an empirical investigation reveals.

The distinction between a posteriori eliminativism and a priori eliminativism is idealised and blurs at the edges. However, it remains informative. The distinction helps to frame different accounts of eliminativism that are fuelled by different sets of arguments, not all of which are susceptible to the same objections. By locating a given version of eliminativism in relation to these two idealisations, the eliminativist can better articulate the nature of their anti-representationalist commitments whilst their opponent can better formulate their response. As such, this underexamined distinction is important. The remainder of this section will explore both versions of eliminativism in greater detail, before raising objections to a priori eliminativism.

3.2 A posteriori eliminativism

A posteriori eliminativism is characterised by an evaluation of the role of representation in our best theories in cognitive science. This tendency is evident in much of the history of anti-representationalism. For example, in the early days of embodied, embedded, extended and enactive cognition (so-called ‘4E cognition’), proponents were often concerned with demonstrating that some cognitive capacities—traditionally understood in terms of brain-based computational operations performed over discrete, symbolic representations—were best explained by non-representational processes. Amongst other

things, 4E draws attention to the power of bodily morphology, body-environment coupling, environmental offloading, and frugal environment-based heuristics.[6]

Partly inspired by developments in robotics and the sciences of artificial life (for example, Brooks, 1991; Chiel & Beer, 1997), much of the 4E movement has appealed to the best explanations of cognition given the data when appraising the need for cognitive representation (for example, see Chemero, 2009). Discussing anti-representationalist 'embodied' approaches to visual perception, Orlandi writes, 'As embodied theorists themselves concede, whether we can do without representations […] is an empirical issue. It is generally believed that we have to wait and see how things turn out' (2014, p. 14). It is also worth noting that many approaches falling within 4E were (and are) concerned with toppling the hegemony of a particular sort of representational explanation—one that assumes cognitive representation is discrete, language-like, wholly brain-bound and entirely descriptive (for related discussion, see Clark, 1997). Only a minority of the 4E movement has been hostile to cognitive representation *tout court*.

This tradition of selective, scientifically driven eliminativism continues. Consider Orlandi's (2014) 'ecological' picture of visual processing. Orlandi argues that visual processing is best understood as the result of an 'embedded system', as opposed to a cognitive process. Orlandi defines a cognitive process as an essentially representational

[6] To illustrate, take the following problem: how does a baseball outfielder catch a fly ball? In contrast to traditional explanations that appear to posit rich internal representations that simulate forward projectile motion (e.g., Saxburg, 1987a, 1987b), some 4E proponents have argued that a better model eliminates the need for representation. For instance, according to Wilson & Golonka (2013), our best explanation appeals to the idea that outfielders move laterally so as to ensure that the ball appears to trace a straight line, exploiting a strategy known as 'linear optical trajectory' (McBeath, Shaffer, & Kaiser, 1995). As the story goes, this strategy exploits a basic relation between the perception of the ball and the organism, ensuring that the two are continuously coupled, and eliminates the need for the organism to construct any internal representation.

and inferential affair (Orlandi thus adopts a narrower definition of 'cognition' than the one used throughout this thesis). According to Orlandi's conception, visual processing has traditionally been understood by 'inferentialists' in terms of encoded rules operating over internal representations within the brain. Orlandi's alternative ecological view drops the appeal to cognitive representation. In its place, Orlandi depicts the perceptual apparatus as an evolved physical system with hardwired sensitivities to producing certain outputs. The take-home message of the ecological view is twofold. Firstly, we should not view the biases and constraints of visual processing in terms of rules encoded within the system operating over representations, but in terms of built-in physiological features of the system. These hardwired features are 'literally just connections, akin to wires or valves, that cause something to happen whenever something else happens' (2014, pp. 45-46). Secondly, these features are the result of evolutionary pressure for a system to become attuned to salient features of its environment; for instance, we detect edges when faced with discontinuities because we lived (and continue to live) in a world of edges, and edges were (and are) useful to detect.

According to the ecological view, our best theory of visual processing is, as a matter of fact, a non-representational theory. Orlandi is not against representation ascriptions *simpliciter*—in fact, they allow for representation ascriptions in domains outside of visual processing (2014, chapter 1). Orlandi selectively opposes a representational explanation of visual processing because they believe there is a better non-representational explanation on offer. Recent responses to Orlandi have picked up on this. Mole & Zhao write in their 'empirical refutation' of Orlandi's ecological view that,

> Because it is an inference to the best explanation, Orlandi's argument depends on the premise that our theories of vision are not able to give better

> explanations when they are allowed to postulate the use of "representational resources" to produce visual percepts. (2016, p. 365)

Adopting this kind of a posteriori eliminativism results in a debate that is analogous to other areas of disagreement in science—a debate over our best theory of some phenomenon. In their response to Orlandi (2014), Mole & Zhao (2016) draw on a 'visual merging' experiment (Zhao, Cakal, & Yu; submitted). Briefly, Mole & Zhao argue that within visual processing, certain information becomes encoded in an abstract way—information that influences processing which occurs at a significantly later time than when the system was exposed to that information. In the experiment in question, colour-patterning was shown to affect localization-responses. Participants were shown a series of arrays consisting of ten distinctly coloured disks. In one condition, the 'structured condition', there were rules governing the pairwise patterning of how these disks were positioned, based on their colour. For example, a red disk would always appear to the left of a blue disk. In the other condition, the 'unstructured condition', there were no such rules. The participants were unaware of the conditions. Following many iterations, participants were allowed a break for as long as they wished (averaging at two minutes) before finally being asked to make a visual judgement on the location of a single disk that appeared for 100ms. Participants pointed to where they believed the disk had appeared using a mouse pointer. The study showed that those participants who had undergone the structured condition tended to locate the single disk closer in space to where its colour partner had appeared. Mole & Zhao argue that,

> The observed influence of color-patterning on localization-responses indicates that, in the 15-minute exposure phase of these experiments, some information comes to be encoded concerning the regularity that governs the distribution of the colored disks. It is our contention that the best explanation of this phenomenon requires that one deny Orlandi's claim that the visual system uses no representations […] the visual system must be using this

> representation in a way that makes a behaviorally-relevant contribution to the experience produced by the subsequently flashed disk. (2016, p. 368)

Whether Mole & Zhao are correct in their interpretation of the study is of less significance for present purposes than the form of their argument. Mole & Zhao concur with Orlandi's conception of what sort of contingent evidence is required to demonstrate that visual processing involves representation, but argue that the empirical evidence, as it happens, casts doubt on the ecological view. They conclude by saying,

> Inference to the best explanation has respectable epistemic credentials only if the theories that it favors give us the best explanation for all of the relevant explananda [...] That is true of Orlandi's theory, but it is no less true of its cognitivist rival which we have been defending here [...] We have suggested that these data swing the balance of probabilities in favor of the idea that visual processes form and use representations in the course of generating our experiences. (2016, p. 372)

The structure of this debate highlights that Orlandi and their opponents rely on contingent empirical evidence to defend a limited form of eliminativism and representationalism respectively. The possibility of an empirically-driven anti-representationalism implied by such debate underscores the essence of a posteriori eliminativism.

Orlandi (2014) offers a version of local a posteriori eliminativism. In keeping with our earlier observation, if one is a local eliminativist, holding representation to be eliminated from some subset of cognitive science, then one is a representationalist when it comes to those other areas of cognitive science that are thought to remain representational. With this in mind, it is worth noting that the conclusions I reach in subsequent chapters could be utilised to support both local a posteriori eliminativism and local representationalism simultaneously. I argue that many traditional notions of cognitive representation are

unsatisfactory because they fail to identify theoretical entities that possess a distinctly representation-like function (see **chapter 2** and **chapter 3**). Nonetheless, there is at least one account, the S-representation account, that demonstrates how cognitive representation can play a robust explanatory role (see **chapter 4** and **chapter 5**). Though the S-representation account itself does not constitute a commitment to how prevalent representation ascriptions are in practice, it does (as it happens) appear to reflect the commitments of some of our most promising theories in cognitive science. To the extent that S-representations are posited as part of our best theories, the account that I defend supports a local, empirically-driven form of representationalism.

At this point, we should observe that my above characterisation of a posteriori eliminativism is too simplistic in one vital respect. Fortunately, the simplification is instructive. My characterisation implies that representationalists and eliminativists share a unified and satisfactory understanding of cognitive representation, merely disagreeing over whether our best theory posits entities that meet this understanding—akin to, say, cosmologists debating whether data on the velocity of gas clouds at the centre of the Milky Way indicates the presence of a black hole. However, unlike black holes, there is no agreed upon understanding of cognitive representation. On the contrary, the explanatory role of cognitive representation remains so controversial, in part, because different theorists hold widely varying standards. As such, we cannot assess the value of representation for cognitive science solely by examining whether theorists use the term ‘representation’ to describe entities in our best theory because it is unclear whether the standards by which they do so are satisfactory. In other words, it remains possible for scientists to talk about representation without referring to entities that meet reasonable

standards for genuine representation (we will return to this idea in **section 3.4** and **section 4.2** below).

As it happens, many eliminativists, including Orlandi (2014), do attempt to spell out reasonable standards for cognitive representation before arguing that these are not met in this or that arena of scientific investigation. This is the correct way to proceed. Substantive agreement or disagreement requires shared meaning, and to decide whether cognitive representation plays a role in our best theory, an approximate consensus on what counts as a cognitive representation is required. Again, future chapters will be devoted to articulating a set of reasonable standards for the justifiable attribution of cognitive representation that accords with the kind of explanations offered by contemporary cognitive science—namely, mechanistic explanations of cognitive capacities.

Despite the limitations of the a posteriori eliminativism label, it remains important for capturing significant similarities in strategy shared by a subset of eliminativists. Furthermore, it highlights the difference between these theorists, who rely on assessing the contingent efficacy of cognitive representation in scientific practice, and those that believe the very notion of cognitive representation is incoherent. This will become clearer as we turn to examine a priori eliminativism next.

3.3 A priori eliminativism

Drawing on a broadly Wittgensteinian and Rylean heritage, the a priori eliminativist maintains that ascriptions of cognitive representation rest on a failure to understand the essential character of both representation and the subpersonal level. In turn, this misunderstanding begets a category error. I will focus on two versions of a priori

eliminativism, drawing on the contemporary literature. The first version provides perhaps the most explicitly a priori eliminativist argument and comes courtesy of Bennett & Hacker (2007). The second version is less clear-cut but offers informative (and to my mind more persuasive) a priori eliminativist concerns. This comes courtesy of Hutto & Myin (2013; 2017). Let's examine these in turn.

Bennett & Hacker (2007) provide the most straightforward argument for a priori eliminativism about cognitive representation. For Bennett & Hacker, representation-talk in neuroscience is part of the larger practice of attributing 'psychological predicates' to the brain. Other examples of these predicates include, 'believing', 'storing' and 'hypothesising'. They write,

> [T]his application of psychological predicates to the brain *makes no sense* [...] *The brain is not a logically appropriate subject for psychological predicates*. (*ibid*., p. 21. Original emphasis.)

The essential idea behind what I dub the 'nonsense view' of psychological predicates at the subpersonal level, is that psychological predicates refer to capacities or properties of whole persons by definition. For Bennett & Hacker, the 'mereological fallacy' is committed by scientists who attribute predicates that refer only to whole persons, to parts of persons (*ibid.*, p. 22). They write,

> The organs of an animal are parts of the animal, and psychological predicates are ascribable to the whole animal, not to its constituent parts. (*ibid.*, p. 22)

The nonsense view operates on the principle that certain predicates have necessary limits on the domains in which they successfully refer. Intuitive support for this idea comes from the consideration of ordinary examples that highlight the limitations on literal

reference. The sea does not literally roar, dance or devour, it only does these things metaphorically. Pencils do not themselves literally write, they only do so in a derivative sense (persons write using pencils). In much the same way, we might doubt that neurons can worry, hope, love and laugh. Such terms appropriately describe activities at the level of persons, but it is hard to imagine their literal application to neurons. The force of the nonsense view derives from a generalisation of these intuitive cases to include all psychological predicates. This establishes the psychological and the neural—roughly, the personal and the subpersonal—as unique realms that strictly demarcate the legitimate application of all psychological predicates. For Bennett & Hacker, eliminating psychological predicates at the subpersonal level is an exercise in the correct application of concepts. Psychological predicates apply 'essentially' (*ibid.*, p. 22) and 'paradigmatically' (*ibid.*, p. 23) to whole animals. To think otherwise is to be entranced by a 'degenerate form of Cartesianism' (*ibid.*, p. 20).

The nonsense view rests on two shaky assumptions. The first assumption is that *all* psychological predicates share the same fate—for example, that loving and predicting are both necessarily bound by the same limitations on reference—and so are all equally nonsensical when applied to the subpersonal level. The second assumption is that the domains of successful reference for a predicate can be determined from the armchair. And yet, there are good reasons to think that knowing whether a psychological predicate at the subpersonal level successfully refers depends on ascertaining, on a case-by-case basis, the properties implied by a predicate, and the results of attempting to extend that predicate into a new sphere of explanation. In lieu of a general argument that shows otherwise, though we may concede that many (even most) prototypically psychological predicates are likely to remain within the bounds of the psychological domain, we cannot rule out

the possibility that a subset of psychological predicates will successfully refer elsewhere. We will return to this point in the objections to a priori eliminativism below. For now, the central message of the nonsense view is that, just as only agents can worry, hope, love and so on, only persons can represent—brains cannot.

The idea that cognitive representation begets a kind of category error is also discernible in arguments for 'radical enactivist cognition' (REC). Articulating an a priori eliminativist version of REC is useful because, unlike Bennett & Hacker, Hutto & Myin avoid eliminating psychological predicates wholesale, instead highlighting a narrower problem with the concept of cognitive representation—namely, that it implies subpersonal correctness conditions (Hutto & Myin, 2013, 2014, 2017; Myin & Hutto, 2015).[7]

According to REC, there is no representation at the level of 'basic cognition'. Such basic cognition, I take it, encompasses all cognitive activity at the subpersonal level. Hutto & Myin agree with the characterisation of representation offered above, holding that representation involves the possession of content in a way that implies correctness conditions. However, for Hutto & Myin, this understanding reveals the 'hard problem of content': any purported case of representation must bear correctness conditions if it is to meet the requirements for genuine representation-hood; and yet, there is no satisfactory justification for attributing correctness conditions at the level of basic cognition.[8]

---

[7] There is a degree of ambiguity in the arguments for REC. One might contend that Hutto & Myin are not committed to a priori eliminativism because they only hold that no account to date has shown how subpersonal representation plays a legitimate role in scientific explanations (leaving open the future possibility of a successful account). This degree of ambiguity is unsurprising given that the categories of a posteriori and a priori eliminativism are fuzzy at the edges. In any case, Hutto & Myin are clearly concerned with the conceptual foundations of subpersonal representation, and from this we can tap an instructive a priori vein.

[8] The tradition of 'autopoetic enactivism' raises similar a priori worries about the compatibility of correctness conditions and the subpersonal level, though from a somewhat different angle (Varela, Thompson & Rosch, 2016). I touch on this tradition again in **chapter 3**.

For Hutto & Myin, basic cognition is entirely a matter of sensitively and selectively responding to information—more precisely, 'natural information' (see **chapter 2** and **chapter 5** for further discussion on information). Supposed content ascriptions at the subpersonal level reduce to the identification of 'informational relations'. This 'co-variation' or 'co-occurrence' based understanding of representation is common in the literature. For example, a frog's brain state might be said to represent 'fly' because of some reliable pattern of covariation between activation of the brain state and the presence of fly stimuli. And yet, information (understood as 'information-as-covariance') is logically distinct from representational content (Hutto & Myin, 2013, p. 67). This is because representational content implies correctness conditions, whereas covariation (at least by itself) does not. Therefore, mere information is not sufficient for representation.

Hutto & Myin further claim that genuine representational content only arises with the introduction of intersubjective norms for it is only these norms that produce standards for genuine correctness. Such norms emerge amongst socialised (possibly only language-using) agents with relatively complex cognitive capacities. In short, those adopting REC hold that only 'public representations' established by intersubjective, sociocultural practices enable the semantic properties required for representation (2015, p. 6). Tonneau expresses a similar assumption about the need for socialised agents to enable genuine representation. In discussing Ramsey's (2007) defence of cognitive representation's role within certain theories of cognition (discussed in later chapters), Tonneau writes:

> Ordinary representation is rooted in a set of social practices and contexts […] It would be absurd to suppose that these social factors are present in the case of neural states […] If social practices are needed to make of an entity a

> representation, however, Ramsey should conclude that there are no representations in the brain and that there cannot be any. (2011, p. 339)

Malafouris (2013), in defending a version of enactivism within the context of cognitive archaeology, expresses some related concerns about the possibility of genuine representation outside of human engagement with material culture (following Steiner, 2010). He writes,

> [T]he only representations with any substantial or real implication for human cognition are to be found outside the head. Internal representations are simply a misleading attempt to explain the unfamiliar intricate workings of the human mind and brain by way of a more familiar model: that of the external material symbol. (*ibid*., p. 31)

In summary, for those inclined towards REC-style eliminativism, social or cultural factors are necessary for genuine representation. By definition, such factors are not present at the subpersonal level. There is no content at the subpersonal level, therefore, there is no representation at the subpersonal level.

We are now equipped with an understanding of the important distinction between the two kinds of eliminativism. The remainder of this section surveys two objections to a priori eliminativism. Casting doubt on a priori eliminativism is important because if it is correct, and cognitive representation is an incoherent concept, then any hope for representationalism is dashed. As the target is wide, the objections will be broad. These objections are not intended as knockdown arguments. Rather, they are intended to cast enough doubt over a priori eliminativism that the idea of cognitive representation playing an explanatory role in cognitive science remains plausible.

3.4 Objection 1: A priori eliminativism and the ubiquity of representation

The a priori eliminativist faces a *prima facie* problem: they claim that cognitive representation results in a category error and so should be eliminated, yet cognitive science is replete with theories that invoke cognitive representation. In turn, one might reasonably suspect that pervasiveness of representation-talk indicates its explanatory utility for at least some of cognitive science. More narrowly, the thought goes, if representation ascriptions feature in our best explanations of some cognitive phenomena then we should believe that such ascriptions are explanatory for those phenomena. However, this conflicts with the a priori eliminativist's claim that the notion of cognitive representation results in a category error.

One strategy available to any eliminativist when tasked with addressing the prevalence of representation-talk is to deny that representation features in our *best* theory. The ubiquity of cognitive representation in scientific practice does not support representationalism unless our best scientific theory features cognitive representation. However, the opponent of a priori eliminativism could reasonably insist that this strategy only indicates the contingent eliminability of representation. This is because it is a strategy that relies on a conditional claim about what sort of entities are posited as part of our best scientific theory. Indeed, whilst this strategy highlights that the representationalist cannot appeal solely to the ubiquity of representation ascriptions—because such ascriptions do not necessarily feature in our best theory—one could still hold that this ubiquity indicates that representational theories of cognition are on offer. For instance, a proponent of Orlandi's (2014) ecological view might concede that legitimate representation ascriptions are common throughout theories of visual processing, whilst maintaining that our best theory of visual processing happens to be the

ecological view. In doing so, one would allow that there are legitimate representational theories of visual processing competing with the ecological view.

There is a further, subtly different strategy open to the a priori eliminativist. This strategy allows that entities referred to as 'representations' sometimes feature in even our best theories; however, it denies that those entities ever meet the minimum requirements for genuine representation. The eliminativist here draws attention to a real possibility: theorists can *talk* about representation without actually positing anything distinctly representation-like. Recall from our earlier discussion that talk of representation is not sufficient to prove the efficacy of representation ascriptions (*qua* representation) in our best theory of cognition. This is because scientists can misdescribe the entities that are posited as part of a theory (see **section 3.2** above). For instance, a theoretical entity might resemble representation—hence, the temptation to label it as such—but still, lack some core characteristic that justifies the representation label. Indeed, future chapters will show that whilst some purported cognitive representations share certain properties with familiar representations they do not resemble anything that we would intuitively classify as a full-blown representation. For instance, many so-called representations within cognitive science are more accurately described as 'casual relays', or otherwise possess a more basic function than representing (see **chapter 2**). This second strategy is consistent with a priori eliminativism. Nonetheless, a priori eliminativism does not straightforwardly follow: one can agree that some so-called cognitive representations are mislabelled without thinking that cognitive representation results in a category error. The question remains whether so-called cognitive representations are always, and by necessity, mislabelled. This brings us to our next objection.

3.5 Objection 2: A priori eliminativism and 'armchair boundary drawing'

The most serious objection to a priori eliminativism targets the idea that the armchair is a suitable location from which to ascertain the legitimate domains of reference for some or all psychological predicates, and suggests that this armchair adjudication lends itself to unresonable standards for cognitive representation. This is particularly problematic for Bennett & Hacker's version of a priori eliminativism though, as we shall see, it raises a challenge for Hutto & Myin's version of a priori eliminativism too.

A natural worry about a priori eliminativism is that it sets the bar for cognitive representation too high. In particular, the nonsense view implies that all psychological predicates are conceptually bound to whole persons. But it is this very person-centric understanding of representation that the representationalist finds unmotivated. Such a rejection of the personal level as (a priori) necessary for representation is supported by the reasonable intuition that determining the appropriateness of representation ascriptions involves examining their application on a case-by-case basis. Thus, we can assess whether activity in the new domain matches the paradigmatic activity that the predicate describes, such that the novel application is warranted. For example, we can assess whether a purported cognitive representation is performing a role that is analogous to familiar cases of representation, such that the representational label usefully describes its properties.

Figdor (2014, 2017, 2018) presents a similar scepticism about the armchair enforcement of restrictions on psychological predicates (Figdor, 2018, positions her view in opposition to Bennett & Hacker's). For Figdor, activity-referring terms originating in the psychological domain sometimes refer to activities in non-psychological domains too. Furthermore, it is only by examining the application of psychological predicates in novel

domains on a case-by-case basis that we discover what, if anything, justifies extending the scope of reference. This way of thinking about psychological predicates has, as I see it, two chief components. The first is a claim about what our default reading of psychological-predicate use in science should be; that is, we should understand scientists as using psychological predicates literally. The second is a claim about what that reading implies; that is, psychological predicates can, at least sometimes, successfully refer within non-psychological domains.[9] Let's examine Figdor's position further.

For Figdor, science takes priority over commonsense intuition when determining whether the extension of a psychological predicate (the set of entities that satisfy the term's meaning) includes entities in a novel domain of scientific explanation. When scientists deploy psychological predicates like 'hypothesising', 'predicting' and 'representing' at the subpersonal level they are typically selecting those terms because they take them to appropriately describe the phenomenon in question (Figdor also examines the application of psychological predicates within biology). In this way, scientists are attempting to appropriately apply predicates in a literal fashion. This does not mean that scientists are infallible in their choice of description; the use of a psychological predicate may or may not accurately describe activity in the new domain. Rather, it is to say that the ascriptions of psychological predicates by scientists are typically best interpreted as literal—scientists intend the sameness of reference across psychological and non-psychological domains. It also does not mean that scientists never use psychological predicates in a

[9] Literalism should not be confused with scientific realism (see **section 4.1** below). Literalism, as intended here, offers a thesis about how to interpret descriptions within an arena of discourse. This is not the same as a thesis about the ultimate ontological status of theoretical entities. What matters is that literalism supports the potential explanatory value of psychological predicates by interpreting their scientific use at the subpersonal level as analogous to their use at the personal level: psychological predicates can contribute to an explanation within non-psychological realms, by referring in the same (literal) way they do in the psychological realm.

metaphorical or similar 'as-if' fashion. However, such uses are an exception, not the rule. By interpreting the standard scientific usage of psychological predicates in this literal way, we open ourselves to evaluating the appropriateness of psychological predicates on a case-by-case basis. Figdor (2018) suggests this can be done through qualitative comparison—judging whether activity in a novel domain is similar enough to activity in the psychological domain to warrant application of the same psychological predicate—or, if available, by testing whether the formal model of an activity that prototypically occurs within the psychological domain also applies to activity in a novel domain.[10]

Supporting Figdor's view that psychological predicates can refer in non-psychological domains, there is no obvious reason to assume that a psychological predicate necessarily encapsulates agent or human-only information. Psychological predicates capture activities that, in principle, can occur across different domains. As Figdor points out, superordinate-categories of objects (classes of objects across different domains) are often bound by activities that are common across those object domains (Figdor, 2014). There is no reason, in principle, that just because a term originally referred to activity in one domain (say, the psychological), it could not refer to activity in another (say, the neural). When scientists apply psychological predicates to non-personal level phenomena, they often do so because they have observed an interesting similarity between activities. For example, when a Bayesian psychologist talks of the brain 'estimating', they are not bewitched by Cartesian black magic but are selecting what they take to be the most accurate predicate for capturing the phenomenon, given the parallel between everyday

[10] An example of the latter is the Drift-Diffusion Model of two-choice decision-making (Ratcliff & McKoon, 2008) that applies, it transpires, to the behaviour of both humans and fruitflies (Figdor, 2018). This model accounts for relatively automatic two-choice decisions in terms of accumulating information over time at a certain rate (the 'drift rate') until evidence for one option exceeds a prespecified threshold and a response to the decision-making task is given.

estimation and activity performed by the brain. In turn, ‘estimation’ is revealed to be more prevalent throughout nature than first thought. Figdor writes of psychological predicates,

> Rather than think such uses are cognitively defective, it is more plausible that they pick out categories of which some of their human manifestations are familiar prototypes, but which have equally real, as-yet non-prototypical, members. We stand to be able to understand psychological capacities better by seeing the full range of ways in which they can be possessed and by developing formal models of them that are applicable across many domains. (2017, p. 4306)

In short, extending the use of psychological predicates into novel areas of science is sometimes a justifiable act of capturing natural phenomena in familiar and informative terms. By doing so, we arrive at a fuller, scientifically-driven appreciation of that psychological predicate and its place in nature. To borrow from Dennett, ‘there is no bright line between analysis and revision’ (2018, p. 59). In summary, there is no reason to suppose that representation could not refer to activity at the subpersonal level simply because it counts as a psychological predicate.

The present objection chiefly targets Bennett & Hacker’s nonsense view of psychological predicates at the subpersonal level. This is because Hutto & Myin do not seek to eliminate cognitive representation through its membership as a psychological predicate. Rather, they identify a particular feature of representation, namely content, and argue that *this* feature is essentially bound to the personal level, never mind what fate befalls other psychological predicates. Following Figdor, an important question remains: are the personal level norms that Hutto & Myin rightly identify as essential to prototypical, personal level representation necessary for all entities legitimately typed as representation; and does ascribing cognitive representation always subtract from, and never add to, the accuracy of our theoretical descriptions? Future chapters argue that in

contrast to what Hutto & Myin suggest, it is possible for a class of subpersonal 'cognitive mechanism' to sufficiently resemble a class of familiar representation without invoking the intersubjective norms of socialised agents. Theories that posits such mechanisms are representational theories because they posit entities with distinctly representation-like functions, in a manner that implies content, as a part of their explanations of cognition.[11]

This section has raised two reasons to be suspicious of a priori eliminativism. I have argued that we should not dismiss cognitive representation on the grounds that it involves a category error, leaving open the possibility that representation could play an explanatory role in cognitive science. And yet, even if subpersonal representation does turn out to play an explanatory role in cognitive science, it does not follow that such representations exist in any strong, metaphysical sense. This brings us to our second distinction that helps clarify the conceptual landscape surrounding representationalism and eliminativism.

4.1 Representationalism vs. realism about representation

There is an important difference between the explanatory role of subpersonal cognitive representation on the one hand and its ultimate ontological status on the other. Of course, the issues are related: if talk of subpersonal cognitive representation is not explanatory, then there is no reason to think that subpersonal cognitive representations exist. Nonetheless, the issues are distinct. Representationalism and eliminativism, as I have presented them, are chiefly concerned with the explanatory role of representation. It follows that 'representationalism' is different from 'realism about representation' in any strong metaphysical sense, and they should not be conflated.

[11]My thanks go to Carrie Figdor for discussing the points raised in this section with me (personal communication).

To appreciate the distinction under examination, observe that there are different 'ontological stances' that a representationalist can adopt and not all of these imply realism about cognitive representation. These ontological stances are informed by orthogonal assumptions about the ontological commitments of a scientific theory, as discussed within the realism/anti-realism debate in philosophy of science. A complete exploration of the scientific realism/anti-realism debate would steer us too far off course. For the purposes of this chapter, it will suffice to briefly survey the core differences between scientific realism and anti-realism, and their relationship to representationalism and eliminativism. This will accentuate the difference between a stance on the theoretical value of cognitive representation and a stance on the final ontology of cognitive representation.

For present purposes, the scientific realism/anti-realism debate can be viewed primarily as a debate over the ontological commitments of a scientific theory (following Quine, 1948). Scientific realism is a broad family of views that urge us to accept the existence of the theoretical entities featured in our best scientific theories. These theoretical entities are typically understood as 'unobservables': those entities posited by a theory (objects, properties, processes, mechanisms etc.) that cannot be observed directly. Exemplar unobservables include subatomic particles like quarks and microscopic biological units such as DNA. According to scientific realism, if our best theories in physics posit quarks, then we ought to believe in the existence of quarks; if our best theories in biology posit DNA, then we ought to believe in the existence of DNA.

Anti-realism is a broad family of views that reject scientific realism. Anti-realists hold that we are not compelled to believe in the existence of the unobservables featured in our

best scientific theories. According to one prominent strain of anti-realism, 'instrumentalism', the success of a scientific theory is measured by its capacity to help us navigate the observable world. For the instrumentalist, ascriptions of unobservables are legitimated by their role in theories that play an 'instrumental role' in prediction, manipulation and problem solving grounded in observable phenomena. As such, for the instrumentalist, we are not compelled to believe in the existence of a theoretical entity just because that entity is of explanatory value in our best theory of some phenomenon.

Cognitive representation is an unobservable in so far as it is a type of theoretical entity posited by scientific explanations that cannot be directly observed. If representation does feature in our best theory then, like quarks and DNA, one's view on the ontological status of representation will be informed by broader assumptions pertaining to the ontological commitments of a theory. In this way, a representationalist can adopt either some form of realism or anti-realism toward cognitive representation. A realist about representation takes cognitive representation to feature in our best theory of some phenomenon and holds that this supports belief in the existence of such representation (in a strong, metaphysical sense). An anti-realist about representation takes cognitive representation to feature in our best theory of some phenomenon but does not hold that this supports belief in the existence of such representation (in a strong, metaphysical sense).

At this stage, one may point out a complication in our gloss of scientific realism and anti-realism. As presented so far, anti-realism reflects a universal attitude toward the ontological commitments of scientific theories. However, some hold a narrower anti-realist stance toward cognitive representation without necessarily subscribing to anti-realism more generally. Such a position results from scepticism about the realism of

subpersonal representation in particular, not from scepticism about the existence of unobservables in general. Proponents of this view maintain that cognitive representation makes a contribution to our best scientific theory of some phenomenon, thus avoiding eliminativism, whilst resisting belief in the existence of such representation—though they do not necessarily extend that resistance to other theoretical posits (such as quarks and DNA).[12] In fact, the possibility of this narrower anti-realism further highlights my general point: belief in the explanatory value of cognitive representation is consistent with multiple, often nuanced, ontological stances toward cognitive representation.

I also raise the possibility of this narrower anti-realism because it helps to frame attacks on anti-realism about cognitive representation. To see this, take Rescorla's (2016) objection to 'instrumentalism' about representation. Rescorla's principle concern is Bayesian psychology which he says posits representations in the form of 'probabilistic state estimates': subpersonal hypotheses with probability assignments carrying correctness conditions fixed by a system's environment (*ibid.*, p. 20). Rescorla criticises anti-realists about representation—under the bracket of which he includes McDowell, (1994), Dennett (1987), and Hornsby (2000)—for failing to appreciate the seriousness of representation-talk within domains like Bayesian psychology. According to these anti-realists, representation plays a merely 'instrumental role' in explanation and prediction. However, for Rescorla, Bayesian psychology 'does not advance these intentional

[12] This position is especially evident in recent discussions of 'fictionalism' about cognitive representation (for example, see Sprevak, 2013; and in particular, Downey, 2018). Fictionalism is a variety of narrower anti-realism that stresses the fictional status of (otherwise useful) representation ascriptions at the subpersonal level. The fictionalist agrees with the realist that representation ascriptions are theoretically valuable, but unlike the realist, denies that cognitive representation exists in a strong, metaphysical sense. In this context, fictionalism concerns the ontology of cognitive representation in particular, and not the scientific realism vs. anti-realism debate more generally.

attributions in a metaphorical or "as if" fashion' in the way that anti-realists suggest (*ibid.*, p. 28). He goes on to say,

> Talk about forward models, priors, state estimates, and cost assignments is not just a useful predictive device. Sensorimotor psychologists postulate these theoretical entities for their explanatory power, just as physicists postulate gravitational forces and biologists postulate genes for their explanatory power. (*ibid.*, p. 29)

Of particular note, Rescorla is a scientific realist (*ibid.*, p. 23), and his attack on anti-realism about representation appears to assume that his opponent is also a scientific realist about things like gravitational forces and genes. As such, his analogy between representation in Bayesian psychology and gravitational forces in physics targets a version of anti-realism that is only concerned with the instrumental nature of cognitive representation. In other words, the version of instrumentalism that Rescorla has in mind assigns an exclusively instrumental role to cognitive representation whilst maintaining realism towards other unobservables. Rescorla thus appears to assume that his opponent is an anti-realist in the narrower sense. However, we should recognise that if one is an anti-realist in general, then the explanatory power of gravitational forces or genes, let alone cognitive representation, bears no special ontological weight.

In closing this section, it is worth visiting Chemero's (2009) discussion of dynamical systems theory. Doing so will help us to further appreciate the importance of separating representationalism from realism about representation. In brief, Chemero believes that our best explanations in cognitive science are acquired by adopting a particular stance: the 'dynamical stance'. The dynamical stance offers explanations in terms of an agent's trajectory through state space described via differential equations, where cognition is characterised as complex, non-linear, self-organising and emergent (for a useful

introduction to dynamical system approaches, see Clark, 2014, chapter 7). This stands in stark contrast with a neuro-centric, symbol-crunching 'cognitivist' approach. According to Chemero, the dynamical stance has no need for cognitive representation.[13] At the same time, Chemero says, the dynamical stance avoids unnecessary metaphysical baggage: 'The dynamical stance, like Dennett's stances on which it is based, is blissfully metaphysics-free' (2009, p. xi). This suggests, quite rightly I think, that the predictive and explanatory effectiveness of dynamical systems theory is independent of any deep metaphysical contention. However, Chemero adds,

> Computationalism and representationalism, though, are not: they are tied to the posit that the mind (or brain) is a computer and full of representations being acted upon by algorithms. (2009, p. xi)

Chemero's presentation indicates a contrast between the metaphysical neutrality of the dynamical systems stance with the staunch realism of computationalism and representationalism.[14] It is true that advocates of representational explanations have historically identified as realists about representation. For instance, many self-identifying representationalists have been proponents of the 'language of thought', which has been taken to imply 'intentional realism'—a position that affirms the reality of certain representational states (Fodor, 1975). However, 'intentional realism' is not necessary for representationalism more broadly construed. A representational theory of cognition is no

[13] In fact, the idea that dynamical systems theory is anti-representational remains controversial. See Bechtel, 1998, for an early attempt at reconciliation.

[14] Chemero's (2009) discussion notably takes place within the context of responding to the 'problem of discovery': the problem of how a theory generates new hypotheses for testing. Chemero believes that the instrumentality of dynamical system theory engenders a lack of background assumptions that are helpful for producing new hypotheses—a problem not faced by computationalism and representationalism because of their staunch (but ultimately misguided) commitment to the existence of computations/representations within the brain. I suspect that Chemero is conflating two commitments here. One is the commitment of computationalism and representationalism to the explanatory power of positing certain internal mechanisms and processes within brain. The other is a commitment to the metaphysical reality of those entities. The former is consistent with anti-realism. The wider issue surrounding the problem of discovery need not concern us here.

more inherently committed to realism about its theoretical posits (in a strong, metaphysical sense) than dynamical systems theory. In short, representationalism is as metaphysically neutral as Chemero's dynamical stance because any ontological stance toward cognitive representation requires further suppositions about the relationship between representational theories and their ontological commitments. Once more, there is no straightforward mapping from the explanatory role of cognitive representation to how we should think about the final ontological status of cognitive representation.[15]

4.2 The metaphysical vs. the scientific path

Section 4.1 above suggested that the debate between representationalists and eliminativists is not, first and foremost, a debate over the final ontological status of cognitive representation. From this starting point, we can see that divergent metaphysical commitments do not directly impinge on whether and where representation plays an explanatory role in cognitive science. If cognitive representation plays a role in our best theory of some phenomenon, then it does not matter (for the purposes of that theory) if one is a realist, anti-realist or whatever-else-ist about cognitive representation; just as one's view on the ontological status of quarks does not directly impinge on whether quarks feature in our best theory in physics. The priority then becomes assessing how cognitive representation could play a part in our best theory, and in turn, whether it does.

If what I have said so far is correct, then there is potential common ground across the realist/anti-realist divide so long as all parties agree on the conditions for the justifiable

[15] The unique metaphysical neutrality of dynamical explanations becomes even harder to defend if one thinks that dynamical explanations ultimately service 'mechanistic explanations' (Zednik, 2011)—or that dynamical explanations must service mechanistic explanations if they are to truly explain and not merely describe (Kaplan & Craver, 2011)—*and* that computational and/or representational explanations are a kind of mechanistic explanation. In fairness, Chemero is reluctant to accept the continuity of dynamical and mechanistic approaches (Chemero & Silberstein, 2008; Chemero 2009).

attribution of representation in explanations of cognition. In this way, we can imagine two paths from which to approach the puzzle of cognitive representation, borrowing from Dennett (1991): the 'metaphysical path' and the 'scientific path'.[16]

The metaphysical path and the scientific path form two approaches to theoretical entities in general. The metaphysical path concerns itself with what ultimately exists, and so addresses the final ontological status of theoretical entities (thus overlapping with the scientific realism vs anti-realism debate sketched above). It is the job of the metaphysical path to guide us toward a true understanding of the ultimately real things, non-real things and any things left in between. The scientific path concerns itself with the utility of theoretical entities within a scientific explanation. It is the job of the scientific path to guide our understanding of whether something plays a useful part in scientific practice, and what that part is.

The idea that we can separate assessing the scientific role of a theoretical entity from evaluating its metaphysical status is evident in work by Dennett (1987, 1991). Dennett is concerned with the reality of what he dubs 'abstract objects', a particular consequence of his notion that representation ascriptions pick up on certain abstract patterns in a system's holistic behaviour from the point of view of an observer's 'intentional stance'. Of note, Dennett principally discusses the mental states of folk psychology, such as beliefs and desires, and is not necessarily concerned with all entities that might fall under the umbrella of cognitive representation as defined above. Furthermore, interpreting whether Dennett is committed to some form of realism or anti-realism about representation has long been debated, by Dennett himself, and by others (for a recent discussion, see Kukla,

[16]The two paths are also akin to Chemero's distinction between 'metaphysical' and 'epistemic' claims about representation (2009, pp. 67-68).

2018). In any case, Dennett's own theory of representation need not concern us for now (see **chapter 2** for further discussion). The present point is to draw on the helpful distinction between 'metaphysical' and 'scientific' concerns about representation. The primary question of interest to Dennett is the one that interests those following the scientific path: 'is it good or bad for our science?' Once again, this question can be asked independently from asking about the final ontological status of cognitive representation.[17]

At first glance, the scientific path seems easily traversed: we need only look and see whether representations feature in our scientific theories of cognition. Yet this is harder than it seems. We cannot discover whether representations feature in our scientific theories of cognition solely by examining whether theorists *talk* about unobservables as representations. The issue is not purely verbal. As **section 3** above highlighted, we must also evaluate purported cases of representation and determine whether 'representation' really captures the role of the entity in question. After all, eliminativists might embrace a scientific theory whose unobservables have been historically labelled as 'representations', insisting that such entities are not really playing a recognisably representational role and therefore that the scientific theory is not really representational.

To judge *whether* cognitive representation plays an explanatory role within a theory in cognitive science we must first secure reasonable criteria that show *how* an entity could play an explanatory role (*qua* representation) in cognitive science. Our initial

[17] Those of a pragmatic persuasion might question whether there are any meaningful metaphysical questions about cognitive representation beyond asking whether and where it is good or bad for our predictive and explanatory purposes. Therefore, anything that can be said about the ontological status of representation will be settled by discoveries made along the scientific path. I sympathise with this view. However, this position still requires going beyond an exploration of representation's explanatory role in scientific practice because it demands further controversial assumptions about how to interpret the ontological implications of prediction and explanation. To this extent, it still requires treading the metaphysical path.

understanding of representation as a functional kind characterised by the role of standing-in for something on behalf of a consumer provides a starting point. Building on this, **chapter 2** affirms the importance of comparing the causally-relevant features of a purported cognitive representation to the core features of familiar representations such as models, maps and portraits. It also identifies more precisely what sort of theoretical entity might play the explanatory role of representation within a theory of cognition: namely, a type of functional mechanism.[18]

4.3 Realism and mechanism

There is an emerging consensus that contemporary cognitive science is in the business of discovering, reconstructing and manipulating mechanisms (Machamer, Darden, & Craver, 2000; Glennan, 2002; Bechtel & Abrahamsen, 2005; Craver, 2007; Bechtel, 2008). The 'mechanistic framework', 'new mechanism' or just 'mechanism', will be further explored in the next chapter. However, given our present discussion, it is prudent to point out now that mechanists often hold that successful mechanistic explanations track truth and uncover the real existing constituents responsible for a phenomenon (for related discussion, see Craver, 2007; Bechtel, 2008; and in particular, Gładziejewski, 2015). Taken at face value, this position commits mechanists to a form of scientific realism: orthodox mechanists believe that we ought to believe in the truth of our best mechanistic

[18]The importance of settling on 'rules' to judge cases of purported representation is reflected in Haugeland's conception of 'objective truth' in science (1998, chapter 13). As I interpret Haugeland, there is a fact of the matter as to whether a theoretical posit counts as a quark, gene, representation or whatever. Nonetheless, that fact is not independent of the greater instrumentally justified norms pertaining to scientific practice. For example, determining whether Pluto is a planet required empirical observation but also involved settling on the appropriate extension of the term 'planet', and was constrained by factors such as a desire for consistent and useful categorisation practices. In general, determining whether *x* counts as *y* is both empirical and normative. Such a perspective also bears an interesting resemblance to themes in the conceptual engineering literature, though space prohibits further analysis at present (I intend to explore these connections in future research). For some relevant discussion, see; Dupré (1995, 2002); Haslanger (2012); Brun (2016); Sawyer (2018); see also Lee (2017), for expanded discussion on Haugeland (1998).

explanations. Thus, if our best explanation of a phenomenon says mechanism *x* performs operation *y*, then we ought to believe in the existence of *x* and the existence of *y* (in a strong, metaphysical sense).

Regardless of any intuitive connection between mechanism and scientific realism, I want to suggest that any substantive conclusions about the existence of mechanisms and their operations must result from a further debate over how best to interpret the ontological commitments of mechanistic explanation. This debate is non-trivial. In fact, there is no straightforward consensus on the relationship between mechanism and scientific realism in the literature (for example, see Kaiser, 2017).

This point matters for our purposes because there are reasons to believe that our most promising understanding of representation will be a mechanistic understanding (for example, see Gładziejewski, 2015; Miłkowski, 2015b). Gładziejewski (2015) suggests a way of situating the popular theory of 'S-representation'—a theory that provides cognitive representation with a clear set of functional criteria—within a mechanistic framework, increasingly taken to capture the dominant form of *explananda* in cognitive science. This is an approach that subsequent chapters will concur with and develop. According to this mechanistic perspective, cognitive representations are mechanisms or mechanism components (components are often mechanisms in their own right), with the function to stand-in for something on behalf of a cognitive system. At least one way a mechanism achieves this role is by realising four functional criteria: structural correspondence, action guidance, decouplability and system-detectable error.

As an orthodox mechanist, Gładziejewski claims that the effectiveness of a representational explanation depends on the existence of representational mechanisms. In doing so, he explicitly rejects Chemero's (2009) distinction, similar to my own, between 'metaphysical' and 'epistemic' questions concerning representation (Gładziejewski, 2015, p. 70). Here I think Gładziejewski oversteps what is required to establish the value of a mechanistic approach. The explanatory success of positing 'representational mechanisms' may ultimately favour a kind of realism about representation. More generally, conceptualising cognitive representation as a kind of mechanism may help to frame discussion over the existence of cognitive representation: if cognitive representation is a kind of mechanism, then perhaps we can better tackle the issue of its ontological status by assessing the ontological implications of mechanistic explanation more generally. All the same, the mechanistic framework does not straightforwardly imply any particular ontological stance toward representation, that is, in a manner free from further argumentation. By emphasising realism about representation and collapsing the distinction between the metaphysical and scientific path, the mechanist about cognitive representation thereby carries the baggage of an ontological commitment. This baggage is not necessary to appreciate the value of a mechanistic approach to representation (as we shall soon see). As this approach continues to evolve, it would be prudent for those marketing mechanistic accounts to distinguish their beliefs about the greater ontological implications of mechanistic explanation from narrower claims about the epistemic benefits of treating cognitive representations as a variety of mechanism.

To summarise **section 4**, evaluating the ultimate ontological status of cognitive representation is not the same thing as evaluating its explanatory role. We can postpone our final metaphysical judgement of cognitive representation whilst we assess if and how

it contributes to our scientific theories of cognition. Because of this, it is important to remember that ‘realism about representation’ and ‘representationalism’ are not equivalent.

5. Conclusion

This chapter discussed representationalism and eliminativism about subpersonal cognitive representation and presented two important distinctions that help to clarify the conceptual landscape surrounding these broad positions. The first distinction is between two tendencies within eliminativism: a posteriori eliminativism and a priori eliminativism. I advanced two objections that undermined a priori eliminativism thereby leaving open the question of whether appealing to subpersonal representation is explanatory in cognitive science. The second distinction is between representation’s role in scientific explanations and its ultimate ontological status. Along the way, I suggested that following the ‘scientific path’ involves more than settling on whether theorists talk about representation. It also requires formulating reasonable standards for genuine cognitive representation in order to evaluate the commitments of our best theories.

Moving forward, mechanistic approaches would do well to separate the epistemic virtues of characterising cognitive representation as a type of mechanism from any strong commitment to its final ontological status. Building on this discussion, the next chapter offers minimum criteria for an entity to play a distinctively representational-like role within an explanation of a cognitive capacity, drawing on the characteristics of familiar representations, and situating these within a mechanistic framework of explanation.

Chapter 2

Cognitive Representations as Mechanisms

1. Introduction

To discover whether and how representation plays a role in explanations of cognition we require an understanding of what sort of explanations cognitive science offers, and hence what sort of theoretical entity a subpersonal cognitive representation might be. This chapter builds on the idea that representation is a functional kind and discusses what sort of subpersonal entity might function as a representation. To this end, I situate the notion of function within the emerging consensus that contemporary cognitive science offers mechanistic explanations. This lends itself to a 'mechanistic account of representation'.

According to the mechanistic account of representation, to be a cognitive representation is to be a mechanism with the function (i.e., causal role) to stand-in for something relative to the realisation of a cognitive capacity; spelling out how a mechanism might fulfil this function then becomes the key challenge for demonstrating the explanatory value of cognitive representation in practice. Equipped with rough criteria for the legitimate ascription of cognitive representation, we can assess existing notions in the literature. It turns out that several orthodox notions of representation are inadequate. This is because they fail to demonstrate that there are cognitive mechanisms (or analogous causal entities) that function in a distinctly representation-like way.

The chapter proceeds as follows. Building on the discussion from the previous chapter, **section 2** spells out in greater detail what it means to function as a representation, and in doing so, raises Ramsey's (2007) 'job description challenge' (JDC). This test demands that for any entity to qualify as a cognitive representation, its role within a cognitive process or architecture must be recognisably representational. I take this opportunity to distinguish between two kinds of information in cognitive science: 'natural information' and 'non-natural information'. Representational content ought to be identified with the latter. **Section 3** suggests that we understand 'function' in terms of activities performed by cognitive mechanisms: organised sets of causally-related spatiotemporal parts that are, in part, responsible for a cognitive capacity. I will discuss what it means for a mechanism to have a function, which in turn grounds what it means to have the function to represent. This lays the foundations for the mechanistic account of representation. **Section 4** tests out the JDC using the mechanistic account of representation as a reference point. I examine three popular notions of representation in the scientific and philosophical literature. I dub these 'receptor', 'action-oriented', and 'intentional stance' representation. I show that none of these notions affords sufficient justification for subpersonal cognitive representation. Therefore, we will have to search elsewhere if we wish to defend the value of representation in explanations of cognition.

2.1 The job description challenge: Representation as a functional kind

To assess whether cognitive representation plays a role in explaining cognition, and if so, how, we must secure a grip on the approximate conditions required for legitimate ascriptions of representation (see **chapter 1**). To my mind, the clearest expression of what is required to meet this task is offered by Ramsey (2007) and his 'job description challenge' (JDC). Ramsey is concerned that the term 'representation' is often used in an

overly liberal manner, with little clarity over what it contributes to a scientific theory of cognition. Ramsey suggests that many representation ascriptions are eliminable without any loss of understanding because supposedly representational entities often fail to resemble anything recognisably representational. As a result, more parsimonious, non-representational terminology would better describe many theoretical entities (even if theorists continue to talk about representation). In a similar vein, Gładziejewski writes, 'the representational terminology too often serves as an empty and misleading ornament, devoid of any real explanatory value—a mere representational gloss on what is at its core a non-representational story about cognition' (2016b, p. 560). Put differently, though the representation label may frequently appear in theories of cognition, the concept is often stretched to the point where it no longer signifies anything of interest, and so, contributes nothing to a theory of cognition *qua* representation (more on this in **section 4** below).[1]

For cognitive representation to contribute towards a theory, *qua* representation, it must pass the job description challenge. To pass the JDC, the entity in question must play a distinctly representational-like role. To borrow again from Ramsey, we must ask, 'Is there an element of a proposed process or architecture that is functioning as a representation in a sufficiently robust or recognizable manner, and if so, how does it do this?' (2007, p. 34). If a theory of cognition posits an entity at the subpersonal level that is distinctly representation-like, then talk of cognitive representation accurately reflects a theoretical commitment. Otherwise, representational descriptions are inappropriate or, at best, a mere gloss. The JDC thus guards against the over-extension of representation.

[1] Cummins likewise warns against accounts of cognitive representation that trivialise representational explanations in science. He advises that 'A good theory of mental representation […] ought to make us understand how appeals to the capacity to represent could explain cognitive capacities' (1996, p. 93).

The most straightforward way to show that a purported cognitive representation possesses a distinctly representation-like role is to show that it functions in a way that resembles prototypical representation (for related discussion, see Ramsey, 2007, pp. 8-14). In other words, a purported cognitive representation is distinctly representation-like when we would reasonably count its functional role as analogous to the functional role of more familiar representations.

A puzzle remains: what would it take for a purported cognitive representation to function in a way that is analogous to more familiar representations? Following **chapter 1**, familiar representations, such as models, maps and portraits, are notable for the fact that they function to stand-in for other entities (objects, processes, states, activities etc.) on behalf of a consumer (Haugeland, 1991; Ramsey, 2007). By functioning in this way, familiar representations are of causal relevance to the behaviour of a consumer in virtue of their semantic properties. Taking this as my starting point, I suggest that for a theoretical entity *x* to function as a cognitive representation, for it to pass the JDC, *x* must be a subpersonal entity (at first pass, a neural or computational structure) which possesses a functional role that meets three broad conditions:

(i) *x* possesses representational content;

(ii) *x* is of causal relevance to a cognitive system S (i.e., *x* affects subsequent cognitive processing and/or motor output of S);

(iii) *x*'s causal relevance to S is due, at least in part, to *x*'s representational content.[2]

These conditions are intended to reflect the minimum standards by which we would judge something to possess a distinctly representation-like function but avoid assuming details about what form cognitive representation might take (Sprevak, 2011, p. 670, articulates similar criteria in reviewing Ramsey's position; see also Strasser, 2010). In this way, we remain open to evaluating types of purported cognitive representations on a case-by-case basis. Let's examine these conditions further.

First, condition (i). As **chapter 1** highlighted, 'content' is key to the identity of any representation. Content concerns a representation's semantic properties. Representations are said to be about those things they stand-in for such that they can be correct or incorrect. For instance, a map might stand-in for features of a mountain range on behalf of mountaineers; Tudor portraits might stand-in for physical features of long-dead royalty on behalf of historians; CCTV footage might stand-in for events at a crime scene on behalf of detectives—in each case, the representation may succeed or fail to correspond to that which it stands-in for. For a theoretical entity to count as a representation in any substantive sense, it must possess content in a manner that similarly implies correctness conditions.

[2] Following much of the mechanism literature I subscribe to an interventionist account of causation. Roughly, according to an interventionist account, *C* is a cause of *E* just in case manipulating *C* results in a manipulation of *E* (for example, see Woodward, 2003). However, I take it that nothing major hinges on the specifics of this account. In particular, I assume that the causal relevance of content, as explicated in **chapter 4** and **chapter 5**, could be couched in the terms of any plausible account of causation (for related discussion, see Gładziejewski & Miłkowski, 2017).

Next, condition (ii). Familiar representations are of causal relevance to their consumers. Representations are exploited by agents, affecting the agents' inferences concerning and actions towards the things represented. An ordinary map of the Himalayas stands-in for the Himalayas, at least in part, because it is consumed (i.e. used) or intended to be consumed, by some past, current or future agent in a way that informs their behaviour towards the Himalayas. By contrast, a slab of rock on a distant lifeless planet that just so happens to resemble the Himalayas does not represent the Himalayas, at least in part because it is causally irrelevant to the behaviour of any agent. For something to play a role *qua* representation in explaining the behaviour of a cognitive system, it must be causally relevant.

Note that in familiar cases where we explain an agent's behaviour with reference to a representation, it is that agent who counts as the consumer. This is because it is that agent for whom the representation is causally relevant. For instance, when we explain the capacity of a hiker to ascend a mountain using a map, it is the hiker who consumes the map. Though obvious, this point is worth emphasising because it highlights the analogous part to be played by a cognitive system in the case of cognitive representation: when we explain the capacity of a cognitive system by ascribing subpersonal representation, we imply that it is the cognitive system itself (and not us) that counts as the consumer (Millikan, 1984). This is because it is the cognitive system itself for whom the representation is causally relevant. At first pass, for a representation to be of causal relevance to a cognitive system, the representation must affect the cognitive processing and/or motor outputs of the cognitive system in question. To do this, the representation must interact with other states, activities and so on, or must interact directly with the system's motor effectors. **Section 3.4** below and **chapter 4** will elaborate on this idea,

suggesting that a cognitive system counts as a consumer when some capacity of that system depends on a mechanism playing an appropriate causal role.

Finally, condition (iii). Familiar representations not only earn their representational status by bearing content and by being causally relevant to their consumer, but by being causally relevant to their consumer *in virtue of* bearing content. We ordinarily think that a map being about the Himalayas is relevant to how it affects a mountaineer—its semantic properties are of causal relevance. Following Dretske (1988), Ramsey writes,

> [T]o be a representation, a state or structure must not only have content, but it must also be the case that this content is in some way pertinent to how it is used. We need, in other words, an account of how it actually *serves as* a representation in a physical system; of how it functions as a representation. (2007, p. 27. Original emphasis.)

In articulating a version of anti-representationalism, Garzón makes a similar point,

> When we claim that representational states have causal powers we mean to say that a system contains physical states that stand for other states and that can play a causal role in the behavior of the system because of the content of the state in question. (2008, pp. 261-262)

Condition (iii) bridges conditions (i) and (ii), ensuring that semantic properties are causally relevant in cases of cognitive representation, as they are in cases of familiar representation.

The remainder of the thesis will be devoted to assessing whether any subpersonal entity posited as part of a theory of cognition does or could satisfy these three conditions. To make progress, it will first prove useful to point out a mapping between the terms

'representational content' and 'non-natural information'. Along the way, we can distinguish between two potentially conflated notions of information in cognitive science.

2.2 Representational content and two kinds of information

Chapter 1 revealed that some eliminativists have assailed representationalism by undermining the plausibility of naturalising representational content at the subpersonal level. For instance, Hutto & Myin (2013, 2017) claim that content is an 'informational notion', but that information in cognitive science amounts to information in the sense of 'information-as-covariance'. In turn, information-as-covariance cannot accommodate correctness conditions. If radical enactivists are correct, nothing at the subpersonal level could pass the JDC. At this stage, it is worth recognising that there are at least two notions of information utilised by cognitive science. Distinguishing these will help us better appreciate what it means to establish or reject the claim that content is of causal relevance. This distinction will be drawn on again below and in subsequent chapters.

Hutto & Myin's separation of mere information from full-blown (content implying) representation corresponds to a common distinction in philosophy between two kinds of 'information' or 'meaning'. Most famously, Grice (1957) distinguishes 'natural meaning' from 'non-natural meaning'. Natural meaning is exemplified in cases such as 'smoke means fire' or 'spots mean measles'. In cases of natural meaning, if *x* means that *y*, then the presence of *x* entails the presence of *y*. By contrast, in cases of non-natural (or 'conventional') meaning, as Grice says, '*x* means that *p* and *x* meant that *p* do not entail *p*' (1957, p. 378). For example, suppose that three rings on the bus mean that the bus is full. This is consistent with the bell's ringing and the bus not in fact being full. Whilst Grice thinks both these kinds of information are kinds of meaning—notably mirroring

the conclusion below that false non-natural information is a kind of information—only the latter implies conditions of semantic evaluation.

Grice's two kinds of meaning help unpack two kinds of information implicit in cognitive science. These are 'natural information' and 'non-natural information' (Piccinini & Scarantino, 2010, 2011). At its most basic, *x* bears natural information about *y*, iff *x* nomically (non-accidentally) correlates with *y*. In this case, *x* bears natural information about *y* just in case changes in the value of *y* are accompanied by systematic changes in the values of *x*. I here assume what has elsewhere been labelled a 'deflationary view' of natural information, where there is nothing more to information-carrying than nomic dependency. This contrasts with a 'realist view' where information is something that emerges from nomic dependency but is ontologically distinct from it (Ramsey, 2007, pp. 132-140). In keeping with Grice's understanding of natural meaning, it was once common to understand natural information as implying veridicality, whereby if *x* bears information about *y*, *y* must obtain (Dretske, 1981; Floridi, 2005, 2010). However, it is increasingly common to understand natural information probabilistically, whereby if *x* bears information about *y*, *x* increases the likelihood of *y* being the case (Piccinini & Scarantino, 2010, 2011; Scarantino, 2015; Millikan, 2017). Intuitive illustrations of ordinary natural information include dark clouds bearing information about imminent rain, or the number of growth rings in a tree bearing information about the organism's age.

In contrast to natural information, *x* bears non-natural information about *y* iff *x* stands-in for *y*, where *x*'s bearing information about *y* does not depend on a nomic correlation with *y*. As such, a tokening of *x* does not entail the truth or increased probability of *y*. In this case, *x* may bear non-natural information about *y* without changes in the value *y*

accompanying systematic (non-accidental) changes in the values of *x*. Intuitive illustrations of ordinary non-natural information include a portrait bearing information about its subject's appearance, or a map bearing information about a mountain's topology. This should remind us of representational content (more on this momentarily). Piccinini & Scarantino summarise by writing that non-natural information bearers,

> […] need not be physically connected to what they are about in any direct way. Thus, there must be an alternative process by which bearers of nonnatural information come to bear nonnatural information about things they may not reliably correlate with. (2011, p 24)

Everyday conventions provide at least one way in which artefacts come to bear nonnatural information about things they may not correlate with, such as when a community decides that a set of marks on paper will stand-in for features of a mountain (features which may or may not obtain). As illustrated by ordinary conventions, non-natural information requires a consumer of some variety to connect *x* and *y* as, by definition, *x* and *y* need not bear any independent causal relationship for *x* to stand-in for *y.* Marks on paper do not bear non-natural information about the topology of a mountain range unless an agent or community use those marks to bear non-natural information about the topology of a mountain range. Non-natural information presumes a consumer of that information.

As already implied, both kinds of information constitute an exploitable relation: if *x* bears (natural or non-natural) information about *y*, then a consumer may exploit *x* to make inferences about or guide behaviour towards *y.* In other words, *x* 'informs' its consumer about *y*. Following Piccinini & Scarantino (2010, 2011), I take these two kinds of information to be evident in the practices of cognitive science. Both kinds of information refer to relations that a cognitive system may use to behave within a noisy, complex world,

acting and communicating with other systems in a manner that is ‘informed’. In other words, both kinds of information concern the reduction of uncertainty within a cognitive system (*ibid.*, p. 21).

I indicated above that non-natural information bearers are distinct from natural information bearers in that they are both causally decouplable from the conditions they bear information about, and the information they bear may be incorrect (false/inaccurate etc.). Yet both intuitively and implicitly within cognitive science, false/inaccurate non-natural information remains ‘informative’. This is reflected in the way we treat ordinary false beliefs. For instance, if one holds a false belief (e.g., that Pokhara is the capital of Nepal) then one holds information about a state of affairs (e.g., the capital city of Nepal), albeit the information is incorrect. This indicates that false non-natural information should still be considered as a genuine kind of information. Piccinini & Scarantino write,

> An important implication of our account is that semantic information of the nonnatural variety does not entail truth. On our account, false nonnatural information is a genuine kind of information, even though it is epistemically inferior to true information. (2011, p. 24)

The authors observe that the orthodox assumption in philosophy has been to regard all so-called false information as not really information at all, principally because philosophy has equated information with natural information. This has played an important part in theorising about representation, where a common assumption has been that representation reduces to a form of natural information-carrying.

It is worth pausing here to elaborate on this idea, examining the role played by the philosophical conceptualisation of information in traditional ‘causal-historical theories’

of representational content (for example, Dretske, 1981, 1988; Millikan, 1989b, 1990; Neander, 1991).[3] Many causal-historical theories of content understand representation in terms of information but distinguish between what an entity merely carries information about from what an entity represents—in our terminology, they distinguish mere natural information from non-natural information—by defining the latter as a subset of the former, typically thought to be determined by the 'proper function' of the representing entity. In their simplest form, causal-historical theories say that R is about C if C reliably causes a tokening of R. For example, a population of neurons firing might represent 'fly' because the presence of flies causes the neurons to fire. This reliable causation establishes a type of reliable correlation that defines natural information. However, as it stands, this delivers an exceptionally weak notion of representation because it fails to yield any semblance of misrepresentation. After all, if R represents the causes of its tokening, R will always (trivially) succeed in representing. Therefore, most sophisticated causal-historical theories add that R only represents when tokened by the subset C that R has the proper function to represent. In turn, R's proper function is given by some feature of R's evolutionary, developmental or learning history. For instance, according to Dretske, R must not only bear a causal relationship with C1 to represent C1, but it must also be the case that C1 caused R to become part of its system's cognitive processing fixed under the evolutionary history of ancestor organisms or under an organism's 'period of learning'. This secures R's proper function. Therefore, the reason why a token representation R is about C1 and not C2 even if R was caused by C2, is because R is a 'symbol token' that belongs to a 'symbol type' whose functional role is to respond to C1 and not C2 (Dretske,

[3] I avoid categorising these theories as 'teleological', as is commonplace given the appeal to functions. This is because I regard my preferred theory of content determination, defended in **chapter 4** and **chapter 5**, to be teleological but not 'causal-historical'. As will soon become apparent, I understand mechanism functions in cognitive science principally in terms of what a mechanism causes—its 'causal role' relative to some cognitive capacity—and not in terms of what causes a mechanism.

1981, p. 198). In summary, for at least some traditional causal-historical theories, a representation cannot carry false information, for there is no such thing, but it can misrepresent if it fails to carry the appropriate information fixed by its proper function.[4] Using our terminology, non-natural information reduces to the carrying of the natural information, plus proper function. Representation succeeds when natural information and proper function coincide and fails when they diverge.

Each causal-historical theory faces its own individual criticisms. However, there are at least two generic objections that are worth surveying at this stage. These will be explored further later. The first objection is that causal-historical theories, like Dretske's, are insufficient for establishing that an entity functions as a representation. The problem is that even if an entity is required to be caused by appropriate conditions to fulfil its proper function, this does not mean that the entity functions as a representation. For example, hearts contract in the presence of excess blood, but we would not describe the heart as a representation, even if we can say that hearts bear natural information (following our deflationary definition above). In other words, lots of non-representational mechanisms have functional roles that involve reliably correlating with certain conditions (I will discuss this issue further in **section 4.1** below). Nonetheless, one might think that the constraints introduced by a theory like Dretske's forms one part of a complete story of cognitive representation; that evolutionary history, learning periods or other historical factors tell us something about how the content of a representation is determined, even if

[4] Miłkowski has responded to Hutto & Myin (2013) by noting that causal accounts of content typically acknowledge the inadequacy of mere covariance as a foundation for representation, observing (with approval) that some notion of proper biological function is typically taken to be the key additional ingredient for establishing correctness conditions (2015a, pp. 82-84). For the reasons outlined in this section and subsequently, I regard this as an unsatisfactory response. In short, Hutto & Myin can reasonably claim that the conditions introduced by theories like Drestke's are insufficient for transforming a functional role that relies on mere covariation into full-blown representation.

they fail to identify in virtue of what something functions as a representation.[5] This brings us to the second objection. Under the mechanistic approach developed throughout the remainder of the thesis, functions are attributed to cognitive mechanisms in virtue of their causal roles within some cognitive capacity, irrespective of their own causal history. This undermines appeals to historical factors in determining content because these factors are not strictly relevant to whether a cognitive mechanism has the causal role to represent. According to the 'mechanistic account of content' that I will develop, it does not matter how a mechanism got there, it matters what a mechanism does. I address this objection further in **section 3.3** below and in **chapter 4** and **chapter 5**.

With worries about causal-historical theories of content in the background, let's now return to our distinction between natural and non-natural information. Again, in contrast to the idea that genuine information cannot be false and following observations of scientific practice, Piccinini & Scarantino (2010, 2011) advise us to understand non-natural information, whether it's correct or incorrect, as a genuine kind of information. Following Piccinini & Scrantino's observations of the potential dual use of information in cognitive science, I take their conceptual mapping to be instructive and useful for our present purposes. Therefore, non-natural information is a genuine kind of information, in keeping with the nuanced notion of information in cognitive science.

We can turn to the example of the increasingly popular 'predictive processing' framework to gesture toward how cognitive science may utilise both kinds of information. In brief, the predictive processing framework conceives of the brain as a hierarchically-structured

[5] Ramsey (2016) suggests such a view. Though a theory like that offered by Dretske is insufficient for addressing the 'functional role dimension' of representation, it might address the 'content grounding dimension'. Ramsey takes addressing these two independent dimensions to be critical for any complete theory of cognitive representation. I discuss this idea further in **chapter 4**.

hypothesis-testing device. The brain is essentially tasked with minimising the error in its own internally generated (top-down) predictions of the incoming (bottom-up) sensory input (for an introduction, see Friston, 2009; Hohwy, 2013; Clark, 2013, 2016). This constitutes an energetically efficient strategy, the thought goes, whereby the brain processes only the difference between the incoming signal and its own predictions. It does this by encoding prior expectations (or 'priors') about sensory input, pitched at multiple spatial and temporal scales spread across its processing hierarchy. Priors consist of a 'hypothesis' that reflects an expectation of the hidden (worldly) causes of stimuli, and an assignment of probability to that hypothesis. According to some versions, hypotheses are selected (are assigned a 'posterior value') as a function of their prior probability plus their 'likelihood'—the probability that the state of affairs captured in the hypothesis would cause the received sensory input, were it true—in approximate accordance with Bayes' theorem:[6]

$$p(A|B) = \frac{p(B|A)\,p(A)}{p(B)}$$

A cognitive system has two strategies for minimising the error that results from the discrepancy between its own predictions and the incoming sensory signal: update its hypothesis or update the world. For some proponents, these two strategies underlie the difference between perception and action respectively. Perception and action form two strategies aimed at the same error-minimising task, but with different directions of fit (world-to-mind and mind-to-world).

[6]Bayes' theorem describes how probable an event is given prior knowledge. A and B are events. $p(A)$ is the marginal probability of A occurring. $p(B)$ is the marginal probability of B occurring. The marginal probability of an event is the probability of the event occurring without reference to any other event. $p(A|B)$ and $p(B|A)$ are conditional probabilities: respectively, the likelihood of event A occurring given that event B occurred, and the likelihood of event B occurring given that event A occurred.

This brief sketch of the predictive processing framework is sufficient to appreciate the *prima facie* plausibility that both kinds of information are at work. Incoming signals carry probabilistic natural information about things in the world in so far as the occurrence of these signals reliably correlates with the worldly causes of stimuli. At the same time, the predictive processing framework relies on non-natural information in so far as it posits a mechanism for endogenously generating top-down predictions of the sensory signal ahead of the sensory signal itself. This is possible, the thought goes, because the brain deploys a kind of internal generative 'model' of the world in the form of stored priors (Hohwy, 2013; Clark; 2016). Very roughly, this internal model can be thought of as a causal-probabilistic structure that maps onto the hidden worldly causes of stimuli. One might reasonably treat such internal models as possessing non-natural information: the generative model produces predictions which inform the cognitive system about states of affairs that may or may not obtain, with the resulting error between (top-down) predictions and the (bottom-up) sensory signal fed back into future predictions. Importantly, the notion of information invoked through the positing of these models goes beyond causal dependency. Indeed, these models are plausibly interpreted as embodying a case of 'S-representation', a purported type of map or model-like representation that **chapter 4** and **chapter 5** argue satisfies the JDC (for further discussion on the relationship between predictive processing and S-representation, see Gładziejewski, 2016b; for some anti-representationalist pushback, see Hutto & Myin, 2017, pp. 57-74).

In summary, there are at least two notions of information that must be distinguished: natural and non-natural. We have been creeping up on the idea that representational content is equivalent to the latter. If an entity bears non-natural information, then it is

semantically evaluable—it has some 'correctness value'. For an entity to bear non-natural information, the thought goes, it must serve as a stand-in for something on behalf of a consumer. When Hutto & Myin talk of content as an informational notion, and information at the subpersonal level as information as covariation, they imply that there is no such thing as non-natural information at the subpersonal level. Through a brief sketch of the predictive processing framework, we have seen how some theories plausibly appeal to both natural and non-natural information. Moving beyond this cursory sketch, more is required to expatiate the value of positing cognitive representation, and with it, non-natural-information/content at the subpersonal level. To understand what ascriptions of cognitive representation might successfully refer to, it will prove instructive to outline a framework that characterises the form of explanation that cognitive science offers.

3.1 The mechanistic framework of explanation[7]

This section lays the foundation for the idea that cognitive representation is a type of cognitive mechanism, and that representational explanations of cognition are a type of mechanistic explanation. Mechanistic explanation is a common feature of the biological and engineering sciences, but a consensus has begun to emerge that says contemporary cognitive science is concerned with offering mechanistic explanations too. According to mechanists, cognitive science explains 'cognitive capacities' by identifying and decomposing underlying causal mechanisms, typically located in the brain, and

[7]Parts of **section 3.1** appear in a manuscript co-written with Joe Dewhurst. This is a work in progress intended for future publication.

sometimes beyond.[8] Paradigmatic cognitive capacities include event recollection, object categorisation, self-relative position tracking, route planning, object-distance estimation and so on. I will largely take the mechanistic framework for granted, focusing my efforts on elucidating what mechanistic explanation is, and how the notion of mechanistic explanation informs our understanding of what it means for a subpersonal entity to possess a function, such as representing.

The past two decades have seen a move towards the adoption of a mechanistic framework of explanation in philosophy of biology and cognitive science. There are several distinct (and somewhat heterogeneous) accounts of mechanistic explanation, including relatively canonical presentations by Machamer, Darden, & Craver (2000), Glennan (2002), Bechtel & Abrahamsen (2005), Craver (2007) and Bechtel & Richardson (2010). In this section, I focus on the essential elements of mechanistic explanation, capturing generally agreed upon features, but focusing especially on Craver's (2007) presentation.

The mechanistic framework developed partly in response to the classical deductive-nomological accounts of explanation (Hempel, 1942; Hempel & Oppenheim, 1948; Popper, 1959), which was thought to be ill-suited to explanation in the special sciences (Bechtel & Abrahamsen, 2005). According to the standard deductive-nomological

[8] As previously noted, according to some 4E approaches, cognitive vehicles sometimes extend into the world beyond the biological boundary of both brain *and* body. For example, an artefact, such as a notebook, may form a constitutive part of the extended cognitive system, assuming a role that might otherwise be performed entirely by a mechanism within the brain (Clark & Chalmers, 1998). I take it that there is nothing incompatible between a mechanistic approach and the 'extended mind hypothesis'. At first pass, whether some external object counts as a cognitive vehicle depends on whether the object is integrated (as a constitutive causal component) into the mechanism responsible for some cognitive capacity. Mechanists introduce notions like mutual manipulability to differentiate mechanism components from environmental factors, where not only should changes in a component affect the mechanism but changes in the mechanism should affect the component (see Craver, 2007). Future research should further unpack the relationship between the mechanistic framework and the extended mind hypothesis (for some initial discussion, see Miłkowski et al., 2018).

approaches, explanations must be expressible as logical arguments, with the *explanandum* forming the conclusion and *explanans* forming the premise set, where the latter includes at least one 'covering law' (or 'general law'). In this way, scientific explanations possess the structure of a deductive argument (for influential criticism, see Eberle, Kaplan, & Montague, 1961; Forge, 1980; Salmon, 1984). By contrast, mechanistic explanation typically focuses on explaining the production or constitution of particular phenomena under particular circumstances by identifying underlying causal structures.[9] A phenomenon, such as the pumping of blood around the cardiovascular system, is to be explained by positing a mechanism, such as the heart, that produces the phenomenon.

A mechanism is an organised set of causally-related physical entities (components), whose activities (or 'operations') and interactions with one another 'produce' or 'constitute', or more generally, 'are responsible for', a phenomenon.[10] For example, the heart is a mechanism for blood circulation because the causal properties and organisation of its parts (valves, atrium, aorta etc.) produce the pumping of blood around the body. The production or constitution of the phenomenon by the mechanism is said to explain the phenomenon, insofar as we can tell a satisfactory story about how the phenomenon is

[9] Following Glennan (2002) and others, I take mechanistic explanations to still offer generalisations of at least two varieties. Firstly, they offer generalisations in so far as they describe a mechanism's regular behaviour over time; what Glennan calls the 'stable dispositions' of 'stable arrangements' (2002, p. 345). Secondly, they offer 'general models' of mechanism types (e.g. a model of the neuron that 'subsumes countless neural events'; Glennan, 2002, p. 345). Following Bechtel & Abrahamsen (2005), I take general models of mechanism to form 'prototypes' that frequently guide research into a set of mechanisms that share a family resemblance. Part of mechanistic explanation involves assessing how a mechanism under study diverges from a model system (*ibid*., pp. 438-439). See Craver, 2001, 2009, for related discussion on individuating mechanisms.

[10] Mechanisms are variably said to produce or constitute a phenomenon. A mechanism produces a phenomenon if its activities result in the production of the phenomenon to be explained. A mechanism constitutes a phenomenon if the activities and interactions between components are themselves the phenomenon that is to be explained. For example, we might say that the stimulated tendons and muscles in the Biceps brachii constitute the phenomenon of elbow flexion. We will mostly be concerned with production mechanisms, in so far as representational mechanisms are part of a wider 'sensory motor mechanism' that produces behaviour; for example, cognitive maps, *qua* representational mechanisms, are in part responsible for producing navigation behaviour in rats (see **chapter 4**). Mechanists sometimes replace 'produce' and 'constitute' with the more neutral 'is responsible for'.

produced or constituted, by reference to the causal and organisational properties of component parts. Mechanisms are hierarchically organised: they are invoked across varying levels of complexities in nature—for instance, ecosystems, kinship groups, organisms, organs, cells etc. Mechanisms are also nested: the components of one mechanism are often mechanisms in their own right. For example, the circulatory system is a mechanism comprised of other mechanisms, such as the lungs, kidneys and heart, which are comprised of other mechanisms such as valves, chambers and so on.

Mechanistic explanations involve identifying and analysing a mechanism (the *explanans*) that is causally responsible for a target phenomenon (the *explanandum*). A mechanism's causal powers are explained by the activities, organisation and structure of its underlying component parts. The causal powers of a component are in turn explained by the activities, organisation and structure of its component parts and so on.[11] At some level, components will bottom out in purely physical descriptions, namely when their activity is no longer explained by the composition of organised entities. Even though such a 'bottoming-out' is in theory always possible, it is not always necessary for a satisfactory grip on a target phenomenon. For example, one need not understand the chemical components of heart cells to possess a working understanding of the heart's role in the circulatory system. The extent to which one is required to descend the mechanistic hierarchy will be relative to the *explanandum* and the relative explanatory demands given the context of the scientific investigation. An ethologist studying spatial navigation in rats need not necessarily understand the cellular mechanisms that comprise the hippocampus

[11]It is not always the case that descending the hierarchy is required for increased explanatory power. For many phenomena, the explanatory burden demands that we focus up the hierarchy. For example, to explain why an organism's insulin levels have increased, one might 'look up' to the organism's recent carbohydrate consumption, rather than the depolarization of Beta cells in the Islets of Langerhans. See Craver & Darden (2013), for discussion.

to have sufficient understanding of the hippocampus as a mechanism that underwrites the spatial memory system.

A target phenomenon (*explanandum*) is whatever it is that we are explaining. In the case of a mechanism such as the heart, the phenomenon to be explained might be the pumping of blood around the body. Accordingly, mechanists observe that a mechanism is always a mechanism *for* something, meaning that there is no such thing as a mechanism in the absence of a target phenomenon (for instance, see Craver, 2007, p. 122). Before a mechanistic explanation can be given, a target phenomenon must be identified, often through a process that Piccinini & Craver (2011) describe as 'functional analysis' (following Cummins, 1975). Functional analysis produces a 'mechanism sketch' consisting of a description of the phenomenon and a parenthetical (or otherwise incomplete) description of a mechanism or mechanisms responsible. For example, we might conduct an initial analysis of blood circulation and conclude that the heart is likely to be functioning as a pump, without yet describing in any detail how it performs this function. A complete explanation requires us to specify the underlying structures that realise this pumping role. In this way, mechanists hold that there is no sharp distinction between functional explanations and mechanistic explanations in cognitive science (Piccinini & Craver, 2011; for some pushback, see Barrett, 2014).

As our discussion so far indicates, talk of functions is ubiquitous in mechanistic explanation. However, philosophers continue to dispute exactly how to understand mechanism functions (for an overview of the debate, see McLaughlin, 2000; Garson, 2016). I take most theories about function to fall within two major families: 'causal-role' accounts and 'selectionist' (or 'etiological') accounts. Following Schwartz (2004) and

Garson (2018), I think it is a mistake to conceive of these accounts as competing. Rather, these accounts reflect different sets of questions. Causal-role functions explain how a trait causes a phenomenon; in other words, they explain the contribution of a trait to a 'system-level effect' (Cummins, 1975). Selectionist functions explain why a trait is there. Kitcher similarly argues that there are two kinds of questions that biology may answer: historical questions that concern the evolution of a system, and structural questions that concern how traits work (1984, p. 320).

Given that different accounts of function sometimes address different sets of questions, I adopt a form of 'pluralism' about function that makes space for both causal-role and selectionist accounts (for related discussion, see Godfrey-Smith 1993; Garson, 2016, 2018). Nonetheless, I take selectionist functions to address the wrong kind of question when considering the role of mechanism function in explanations of cognitive capacities. Cognitive scientists typically wish to explain how an organism or artificial system achieves some cognitive capacity such as facial recognition. Relative to this *explanandum*, the function of a cognitive mechanism is its causal contribution, which can come apart from its selection history. It will prove fruitful to further sketch these two approaches to function and raise the limitations facing selectionist accounts.

3.2 The causal-role account of function

The causal-role account says that the function of a mechanism is the causal contribution of that mechanism to a target phenomenon. In turn, the phenomenon is fixed by the agent or community investigating the phenomenon (Cummins, 1975; Hardcastle, 1999, 2002; Craver, 2013). Craver (2013) offers the most focused presentation of the approach and I will take this as my lead (there are hints of the causal-role account in Craver's earlier

work; see Craver, 2001, 2007, 2009). The attribution of a function to a mechanism, or component of a mechanism, is often an essential step in giving a mechanistic explanation. This is because the physical system that composes the mechanism will be involved in all sorts of activities, behaviours, and causal interactions, not all of which will be relevant to the *explanandum*.[12] In the context of explaining how blood is pumped around the cardiovascular system, the function of the heart is to pump blood, because it is pumping blood that contributes to the production of the target phenomenon. Each individual component of the heart will have its own function, which in turn contributes to the production of the target phenomenon—one component might function as a valve, whilst another might function as a tube, allowing blood to flow at a certain speed and pressure.

Before we can explain how a mechanism causes a phenomenon we must ascertain which aspects of the physical system are relevant to our explanation. By first giving a plausible functional description of the system we can often sketch out, in rough terms, which parts of the system are relevant to the phenomenon at hand, and what kind of contribution we think those parts make. Further analysis of the system, in terms of the activities and organisation of physical components, can then determine how (or indeed if) those parts carry out those functions. If it transpires that there is no conceivable way a system's part could perform the attributed function, then we go back to the drawing board, but if we can identify a plausible physical structure that implements or performs the function, then we have something that is beginning to look like a complete mechanistic explanation. Therefore, we can begin putting together a full mechanistic explanation, based upon an initial functional attribution.

[12]It is for this reason that Craver is not entirely satisfied with earlier causal-role accounts, such as that given by Cummins (1975), although he does take inspiration from these.

Craver elaborates the causal-role account by usefully outlining three kinds of functional description that are involved in mechanistic explanations: 'etiological', 'constitutive', and 'contextual'. These functions serve 'as a way of tersely indicating an etiological explanation, as a way of framing constitutive explanations, and as a way of explaining the item by situating it within higher-level mechanisms' (2013, p. 133). Let's examine these in turn.

Etiological functions are typically attributed in the context of giving an 'etiological explanation', that is, an explanation in terms of the history of a system (see Craver 2013, pp. 145-146). One common type of etiological explanation is selectionist explanation, particularly evolutionary explanation: explaining the existence of a trait by appealing to its evolutionary history. What all etiological explanations have in common is that they are 'backwards-looking', explaining how something is now in terms of how something was in the past. Etiological functions are also backwards-looking. If we say that the function of the heart is to pump blood because that is the role it was selected for in the evolutionary history of ancestor organisms, then we are appealing to the past to say something about the present. Craver raises some familiar concerns about etiological explanations, including that they are speculative and typically indeterminate, and say nothing about how a mechanism produces or constitutes a phenomenon in the here-and-now (*ibid.*, pp. 146-148). However, Craver concedes that attributions of etiological functions 'can be heuristically useful as a guide to creative thinking about what an organism or organ is doing' (*ibid*., p. 148). For the causal-role theorist, I take it, attributing an etiological function to a mechanism helps explain why a mechanism exists. It may also capture one causal role that a mechanism presently serves. However, that etiological function is not privileged. For example, the function of past hearts to pump blood is the causal role that

led to the reproduction of hearts and explains why hearts now exist. Pumping blood is also one causal role that hearts continue to serve. Nonetheless, hearts possess other causal roles relative to phenomena other than blood circulation, for instance, making *thump-thump* noises relative to their role in diagnosing cardiovascular illness (for related discussion, see Glennan, 2002; Craver, 2013). Pumping blood may indeed be the heart's 'proper function', in so far as pumping blood is the causal role which caused hearts to be reproduced—but so-called proper functions are not the only functions a mechanism may serve. We will return to this issue momentarily.

Constitutive explanations play a key role in mechanistic explanations, focusing on the synchronic causal structure of a system (Craver, 2013, pp. 149-151). In contrast with purely etiological explanations, constitutive explanations capture how, in the here-and-now, a mechanism is causally capable of causing a phenomenon. A description of the heart in terms of the opening and closing of valves might serve as a constitutive explanation of how blood is pumped around the body. A function, in this sense, is a description of how relevant physical structures produce the phenomenon that we are interested in: to say that the heart (constitutively) functions as a pump is simply to say that it has the correct kind of physical structure to perform the pumping function. For Craver, such attributions are 'perspectival' insofar as there are many possible ways in which we could describe the physical structure of a system, only some of which will be relevant to our current explanation. The choice of which description to give thus depends on our explanatory perspective: if we are interested in explaining the circulation of blood, then it makes sense to attribute the function of pumping to the heart, whereas if we are interested in explaining the synchrony of a child's heartbeat to its mother's, then it might be more appropriate to attribute a different function, and thus to emphasise different

aspects of its physical structure. The term 'perspectival' is somewhat misleading in so far as the function of a mechanism is perfectly objective relative to a given *explanandum*. For instance, relative to the circulatory system, the function of the heart is, as a matter of fact, to pump blood. **Chapter 5** will raise a related defence of a mechanistic approach to content, against the accusation that it makes content unpalatably observer-dependent.

Finally, a contextual explanation is one that locates the constitutive function within the broader environment of a system and attributions of contextual functions must, therefore, consider a mechanism's environmental context (Craver 2013, pp. 151-154). As Craver summarises,

> [W]e should add that functional characterizations often describe those capacities in a manner that includes wider and wider regions of the causal structure of the system under consideration [...] There is a difference, after all, between knowing that plugs produce sparks and knowing how that sparking is situated in the mechanisms of an engine [...] in the latter we describe its role contextually. (2013 p. 152)

Livers in general function (constitutively) as filters. A token liver functions (contextually) to filter alcohol out of the blood of an individual who has drunk heavily the night before; however, it would not make sense to attribute this function to the liver of a teetotaller. Therefore, two mechanisms may share the same constitutive function whilst differing in the details of their contextual function. Contextual functions are also important when it comes to situating a component within a higher-level mechanism that it is part of. It only makes sense to talk of the heart as a mechanism for pumping blood, as opposed to just being a pump *simpliciter*, within the broader context of its role within the cardiovascular system, which is itself a mechanism for circulating oxygenated blood around the body.

Building on the observation that mechanistic explanations often consider a mechanism's broader context, it is worth stressing that cognitive mechanisms are typically taken to be responsible for capacities that facilitate coordination between a cognitive system and its environment. Because of this, explanations of cognitive capacities often make ineliminable reference to task environments. Explaining an organism's capacity to estimate the distance between two objects, for example, implies the existence of two objects in its surroundings—the capacity relates the organism to its environment. In **chapter 4** and **chapter 5**, I will suggest that we can understand the 'target' of a given representational mechanism—roughly, what that mechanism functions to stand-in for—in terms of its embeddedness within a higher-level mechanism and task environment.

3.3 Selectionist accounts of function

Selectionist accounts state that the function (or 'proper function') of a biological mechanism is fixed by what that type of mechanism was selected for in the past (Williams 1966; Wright, 1973, 1976; Millikan, 1989a, 1989b; Neander, 1991). At their most basic, selectionist accounts say that a mechanism type M (for example, hearts) has the function to *x* (for example, pump blood) just in case M exists (was selected for) in virtue of doing *x*. Selection history is traditionally understood in terms of Darwinian natural selection: the change in heritable traits of a population across generations due to differential reproduction. In recent years, more nuanced versions of the selectionist account have developed that take into consideration other natural selection processes, such as trial-and-error learning and antibody selection (for example, see Godfrey-Smith, 1992; Garson, 2015, 2016, 2018; Millikan, 2017). My focus in this section will be on orthodox

selectionist accounts that take Darwinian evolution as their basis. Nonetheless, some of the problems raised below apply to any selectionist account.[13]

The selectionist approach goes back at least as far as Williams (1966), who writes that 'The designation of something as a *means* or *mechanism* for a certain *goal* or *function* or *purpose* will imply that the machinery involved was fashioned by selection for the goal attributed to it' (p. 9. Quoted in Garson, 2013, p. 322. Original emphasis). For example, the function of the heart is to pump blood because pumping blood contributed to the survival of ancestor organisms (presumably, the same does not apply to the heart's making *thump-thump* noises). Likewise, a mechanism has the function to represent, the thought goes, just in case it was representing that caused the mechanism to be reproduced.

Selectionist accounts face long-standing challenges. For brevity, I will limit myself to highlighting one general issue and two problems that are particularly pertinent to the topic of cognitive representation. To begin, selectionist accounts concern a mechanism's function relative to the selection history of that mechanism. A *prima facie* limitation arises when we observe the common practice of assigning functions to produce or constitute phenomena that mechanisms were not selected for. Most strikingly, biologists and cognitive scientists talk of mechanisms for pathologies. For example, one might talk of the mechanism for metastasis or the mechanism for psychosis (see Craver, 2013, for discussion). And yet, in most cases, the mechanism responsible for causing a pathological

[13] Selectionist accounts are not the only alternative to orthodox causal-role accounts, though they have been the most popular. Working within a mechanistic paradigm, Maley & Piccinini (2017) reject selectionist accounts as too narrow. However, they also claim that causal-role accounts are too permissive. They opt for a view that emphasises a mechanism's contribution toward the 'survival' and 'inclusive fitness' of an organism. I will return to this 'objective goal account' in **chapter 5**, where I will suggest that it is not really a rival to the causal-role account. Rather, it is a nuanced version of the causal-role account fit for the purposes of biology and cognitive science. As we shall see, the objective goal account helps to assuage worries over 'perspectivalism' about content, which is the view that the content of a given representation is determined, at least in part, by an observer.

phenomenon was not selected for producing that phenomenon. Such considerations work in the favour of the causal-role account: selectionist accounts are limited to explaining how a trait came to be, but scientists often wish to explain how a trait causes a phenomenon, regardless of its history.

In response to these observations, Garson suggests that talk of mechanisms for pathologies in biology is 'elliptical' (2013, p. 329). What scientists really mean when they talk about a mechanism for, say, metastasis is that there is a mechanism for cell elasticity that has been disrupted (where cell elasticity is the function that the mechanism was selected for). In fact, following Krickel, the causes of metastasis cannot be explained solely with reference to the disruption of 'normal healthy mechanisms', such as RNA repair mechanisms (2018, p. 46), but requires positing unique entities and processes (though Krickel does not give an example, such mechanisms and processes plausibly include protein degradation by the tumorous cells in the extracellular matrix). Garson does acknowledge that there may be genuine counterexamples in which scientists do talk of mechanisms functioning (non-elliptically) for pathologies, yet he dismisses these cases as insignificant for selectionist accounts 'as long as they are infrequent' (*ibid.*, p. 329). However, based on Garson's discussion, the following remains unclear: (i) how much non-elliptical talk of mechanisms for pathologies would count as significant, and (ii) whether biologists (non-elliptically) talk of mechanisms for pathologies to this degree. Notably, Garson does subscribe to a selectionist account that considers broader selection processes beyond evolution (2017) and has subsequently developed a form of pluralism that allows a place for function ascriptions based on causal roles (2018).

There are two additional obstacles facing selectionist accounts as they apply to function attributions in cognitive science. The first issue picks up on a general theme in contemporary cognitive neuroscience: the plasticity of neural circuitry (for example, see Wexler, 2006). More specifically, this issue concerns the phenomenon of 'neural reuse' and 'neural recycling'—that is, the recruitment of evolved circuitry to serve novel cognitive capacities. Theorists who study neural reuse point out the disparity between the evolutionary era under which at least some human neural structures evolved, and the recentness of many cognitive capacities (such as reading) that those structures appear responsible for causing. This suggests that many cognitive mechanisms are a special kind of 'exaptation' or 'spandrel': traits that serve one or more role which they were not selected for.

Feathers are a well-known example of non-cognitive exaptation. It has been suggested that feathers were initially selected for their role in thermoregulation (feathers are good insulators), only later serving a role in flight. Schwartz (2004) offers another example, following Millikan (1993), of front flippers in sea turtles. These limbs were likely selected for their role in swimming, but later became used by female turtles to dig holes in which to lay their eggs. Barve & Wagner (2013) go so far as to suggest that common metabolic traits may be exaptations, concluding that, 'Metabolic systems thus contain a latent potential for evolutionary innovations with non-adaptive origins' (2013, p. 203). If these examples are correct, then exaptations are common across nature.

It should perhaps be especially unsurprising that exaptations play a prominent role in the brain, given the plasticity and sensitivity of neural circuitry to specific environmental features during development. Gould presents an early and powerful expression of this

idea, claiming that we should expect most of the brain's traits to be a kind of exaptation (Gould & Lewontin import the term 'spandrel' from architecture in their 1979 essay). Gould colourfully summarises his view as follows:

> The human brain is the most complicated device for reasoning and calculating, and for expressing emotion, ever evolved on earth. Natural selection made the human brain big, but most of our mental properties and potentials may be spandrels—that is, nonadaptive side consequences of building a device with such structural complexity. If I put a small computer (no match for a brain) in my factory, my adaptive reasons for so doing (to keep accounts and issue paychecks) represent a tiny subset of what the computer, by virtue of inherent structure, can do (factor-analyze my data on land snails, beat or tie anyone perpetually in tic-tac-toe). In pure numbers, the spandrels overwhelm the adaptations. (1997. Online)

Recent studies into numerical cognition and reading have produced a neuroscientifically informed version of the same general idea: the 'neural recycling hypothesis' (Dehaene & Cohen, 2007; Dehaene, 2009). This hypothesis focusses on the plasticity of the brain, whereby the brain 'hijacks' the neural circuitry evolved for one purpose to serve capacities that have only recently emerged from cultural development, such as numerical cognition and reading. From a somewhat different angle, Anderson (2010) argues that evolved networks of neurons are constantly recruited and linked to serve a multitude of novel tasks. Essentially, the brain's extreme plasticity allows the same neural circuits to be flexibly combined and redeployed, preventing the need to evolve unique neural mechanisms for more recent cognitive capacities.

The precise details of the brain's exaptive nature remain controversial. For example, there is some debate between the frameworks just mentioned over the degree to which recent cognitive capacities either directly hijack existing circuits (Dehaene & Cohen, 2007; Dehaene, 2009) or are required to recombine older circuits into more complex novel

circuits (Anderson, 2010). Nonetheless, the broader point remains consistent: the evolution of a trait often comes apart from at least some of the capacities that trait now serves. The causal-role account can accommodate this result, noting that functions are often attributed to traits solely in virtue of the capacities they cause, not their evolutionary role. Admittedly, more nuanced selectionist accounts may be able to address neural reuse too, for instance, drawing on an organism's learning history to explain how evolved circuits are selected to serve newer capacities. At the very least, the phenomenon of neural reuse puts pressure on narrower selectionist accounts. The onus remains on broader selectionist accounts to demonstrate how other selection processes adequately account for function attributions to neural circuits for recently developed cognitive capacities.

The second problem facing selectionist accounts in cognitive science targets even the most sophisticated version. This problem concerns the opaqueness of a mechanism's selection history and the apparent irrelevance of that opaqueness to mechanistic explanation. The unknowability of a trait's history is a general problem for any inherited biological mechanism, but it is particularly salient in the case of cognitive mechanisms. The selection history of almost all traits cannot be observed. This is less of a problem when the selection history of a trait appears straightforward. For instance, there is little doubt that the activity of blood circulation was key to the propagation of hearts, and so we assume that hearts have the (selected) function to pump blood. However, it is more challenging to provide straightforward historical interpretations of cognitive mechanisms (of which cognitive representations, let's suppose, are a kind). This is because cognitive mechanisms are exceedingly complex and difficult to study. If functions are understood in terms of selection histories, then the opaqueness of selection histories would seem to undermine the explanatory power of function attributions. At the same time, ignorance

about the history of a mechanism does not prohibit scientists attributing functions to cognitive mechanisms in practice. For instance, we need not know the history of the visual cortex to attribute the role of detecting edges in the primary visual cortex (V1). Again, the causal-role account makes sense of this: scientists are interested in the causal contribution of mechanisms in V1 to visual processing. In such cases, a mechanism's function is attributed because of its causal role and not its selection history.

I embrace a pluralism towards theories of function, as noted above. Sometimes scientists appeal to selection histories, and sometimes they appeal to causal roles. The modest conclusion I wish to draw here is that selectionist accounts are inadequate for capturing the explanatory role of function attributions in at least many scientific explanations, specifically explanations involving cognitive mechanisms. If what I have said is correct, then cognitive scientists (at least often) attribute functions to mechanisms based on their causal contribution to an *explanandum* capacity and not based on their etiology. A major consequence follows: if cognitive representation is conceived of as a kind of mechanism, then there is pressure on theories of representational content that rely on a mechanism's etiology (such as Dretske's 1981, 1988 account; or Millikan's, 1989b, 1990 account). This is consonant with the broader worry raised above that the historical causes of a representation's tokening (of whatever kind) are not constitutive of its content-determining relations. I will return to this point in **chapter 4** and **chapter 5** when I discuss how to think about content in representational mechanisms. For now, when discussing the function of a mechanism or its component, I will assume the causal-role account.

3.4 The mechanistic account of representation

If cognitive science is in the business of studying the mechanisms responsible for cognitive capacities, then a promising strategy for showing that cognitive science does or could posit theoretical entities with a distinctly representation-like function begins by conceiving of cognitive representation as a type of mechanism with an appropriate causal role. This brings us to the 'mechanistic account of representation'. This account depicts the 'vehicle' of cognitive representation as a mechanism contained within a cognitive system—most straightforwardly, a mechanism realised by neural structures in the brain. To count as a representational vehicle, a mechanism must play the causal role of a stand-in for something relative to some capacity of a cognitive system. The 'consumer' is the cognitive system whose capacity depends on that mechanism. For example, a cognitive map is a candidate representational mechanism; a hypothesised mechanism located in the mammalian hippocampus, whose role as a stand-in for features of an organism's environment explains an organism's capacity to navigate (see **chapter 4** for more on cognitive maps).

To my knowledge, the closest position in the existing literature to the mechanistic account of representation is offered by Gładziejewski (2015). For Gładziejewski, representational explanations are also a kind of mechanistic explanation. He summarises the commitments of his approach as follows:

> (1) A mechanistic explanation M of a cognitive capacity C is representational iff M explains C by a representational mechanism.

> (2) A mechanism M is representational iff M has at least one component part whose function within the mechanism consists in representing. (2015, p. 67)

Note that in (2), Gładziejewski defines a representational mechanism in terms of a mechanism that possesses a *component* that functions to represent. Again, this representational component will itself be a mechanism, that is, a system whose capacity to represent is explained by the organisation and properties of its component parts (and so on).[14] We can elaborate on Gładziejewski's explication of representational explanation by adding a third condition that combines (i) what it means for something to function as a representation (i.e., to serve as stand-in), with (ii) what it means for a mechanism to possess a function at all (i.e., for it to possess a causal role in some capacity):

> (3) A mechanism has at least one component part whose function within the mechanism consists in representing iff that component has the causal role to stand-in for something relative to the production or constitution of a cognitive capacity.

We can further distil the essentials of the mechanistic account of representation in the following way:

> **Mechanistic Account of Representation**: M is a cognitive representation iff M is a mechanism with the causal role to stand-in for something relative to a cognitive capacity.

[14] Gładziejewski's phrasing draws attention to the fact that a representational mechanism will likely explain some capacity as part of a higher-level mechanism of which it is component. For example, cognitive maps are not solely responsible for causing successful navigation in rats—various other neural and physiological components are required. See **chapter 4** for related discussion.

Observe that the mechanistic account only commits a proponent to what sort of entity a cognitive representation might be (a mechanism), and what constitutes a representational explanation (a mechanistic explanation). In this sense, it is a partial account. It does not commit one to a view of what component properties could fulfil the required criteria, and whether such entities play a part in our best scientific theories. As such, one could agree that the mechanistic account provides a promising way to conceive of representation's role in explanations of cognition but conclude that there are no such mechanisms in our best scientific theories. To determine this, however, we must answer the following question: how might a cognitive mechanism fulfil the function to represent? Following Cummins (1989), Ramsey (2007), Gładziejewski (2015) and others, **chapter 4** will defend a version of the 'S-representation' account, arguing that it provides satisfactory conditions under which a cognitive mechanism may be said to function as a representation. Before we turn to this account, it will prove illustrative to examine some of the most popular notions of representation in cognitive science and philosophy and test them against the mechanistic account of representation.

4.1 Receptor representation

Armed with the JDC, and a better understanding of what it could mean for something to count as a cognitive representation, we are ready for a field test. I will review three notions of representation in the literature. Each of these fails to supply a notion of representation that passes the JDC because they do not provide the basis for subpersonal mechanisms, or analogous causal entities for that matter, with the function to represent on behalf of a cognitive system. 'Receptor' and 'action-oriented' representation concern subpersonal mechanisms, but do not, by themselves anyway, supply criteria for properly representational mechanisms. 'Intentional stance' representation abstracts away from the

mechanistic composition of a system and is indifferent to whether there are any subpersonal mechanisms that function as representations. I do not intend this discussion to serve as a comprehensive presentation of the notions visited, each of which could fill their own thesis, but to provide a flavour of the mechanistic account in action.

An orthodox use of 'representation' in cognitive neuroscience refers to the activity of a set of neurons or computational structures reliably tokening in response to a stimulus, where that activation is relevant to the role the entity plays in subsequent cognitive processing and/or motor output. For instance, cells in the primary visual cortex reliably activate in the presence of edges, hence the common label 'edge detectors' (Hubel & Wiesel, 1962). In turn, these cells are often referred to as representations of edges. Underlying this notion of representation is the idea of 'nomic dependency', where *x* represents *y* just in case *x* is nomically dependent on *y*. In Ramsey's words, representation is supposed to occur where 'some sort of internal state reliably responds to, is caused by, or in some way nomically depends upon some external condition' (2007, p. 123). Here we find our first common notion of cognitive representation: 'receptor representation'.

Much has already been written about the receptor notion and how it fails to provide representation with a robust explanatory role (for discussion see Cummins, 1996; O'Brien & Opie, 2004; and especially Ramsey, 2007; Gładziejewski, 2015; Gładziejewski & Miłkowski, 2017). Nonetheless, receptors remain an important part of the representation debate. Therefore, I will briefly summarise why receptors fail the JDC, before contributing an additional comment that situates the temptation to classify receptors as representations in the context of the distinction made in **section 2.2** above between the two different kinds of information relevant to cognitive science.

By themselves, receptors fail to pass the JDC. The essential problem is that describing receptors as representations adds nothing of explanatory value on top of their role as 'causal mediators' or 'relays'. In turn, describing them as representations encourages an overly permissive notion of cognitive representation. This is chiefly because representational content plays no role in the explanation of how receptors affect a cognitive system's behaviour. Receptors rely on their relationship of dependency with external states of affairs, but *x*'s reliable tokening in the presence of *y* is insufficient for *x* to represent *y* (compare the reliable tokening of steam following boiling water). In general, covariation is insufficient for representation.

Of course, receptors do more than covary. Receptors affect the cognitive system *in virtue of* covarying. In other words, *x*'s activation following *y* affects *x*'s containing system. The activation of edge detectors, for example, is crucial for visual processing in many animals. In this way, receptors function as causal mediators or relays in so far as they mediate or relay between a stimulus and subsequent cognitive processing and motor behaviour (Gładziejewski, 2015, p. 68). However, there are plenty of mechanisms that have parts whose function involves similar causal mediation, but whose function we would not classify as representation. For example, the firing pin in a gun 'bridges a causal gap between the pulling of the trigger and the discharge of the round […] However, no one thinks the firing pin serves as some sort of representational device' (Ramsey, 2007, p. 136). To classify receptors as representations is to obscure an explanatorily relevant difference between distinct functional roles, and thus weaken the efficacy of representation ascriptions. Borrowing from Ramsey again, to classify receptors as representations is to turn the 'substantive idea that cognition is a process that involves

representational states', into the 'remarkably boring thesis that cognition is a process that involves states that are triggered by specific conditions' (*ibid*., p. 125).

A natural response to dismissing receptors as insufficient for providing representation with a substantive role in cognitive science involves bolstering the notion with an additional constraint of the sort offered by a causal-historical theory of content. As Ramsey (2007) discusses, a proponent of Dretske's theory acknowledges that *x*'s causal dependency on *y* is insufficient for *x* to represent *y*; roughly, it must also be the case that it was *x*'s dependency on *y* that caused *x* to become part of the system's processing, in other words, *x*'s proper function must be to reliably respond to *y*. The problem, as anticipated earlier, is that Dretske's additional constraint fails to transform *x*'s functional role into one of representing. In assessing Ramsey's (2007) position, Sprevak usefully summarises the inadequacy of the receptor notion by observing that even in the face of a view like Dretske's, nomic dependency is sufficient to capture the role of receptors:

> The problem with the receptor notion is that the nomic dependency relations do all the explanatory work. The *content* does not have an explanatory role over and above the effects involved in the nomic correlation. This goes for firing pins and spark plugs and just as much for 'edge detector' cells in the V1 cortex. (2011, p. 673. Original emphasis.)

As introduced in **section 2.2** above, Dretske's view is supposed to allow for misrepresentation, because *x* only successfully represents when it is tokened by certain conditions. However, all an account like Dretske's can tell us when applied to receptors (if we accept it) are those conditions under which *x* successfully plays the role of causal mediator; not those conditions under which *x* misrepresents. In other words, a receptor may *malfunction* when it fails to respond to the appropriate causes, but it does not

misrepresent. Let's grant that the proper function of edge detectors is to respond to edges and not non-edges, in which case, when edge detectors are caused by non-edges (as when directly stimulated by experimental instruments), they are responding to the 'wrong' stimuli and so fail to fulfil their proper function.[15] This is analogous to the heart failing to contract in response to excess blood. Both cases of malfunction are disanalogous to, say, the case where a cartographic map fails to stand-in for some geographical region. In general, not all malfunction is misrepresentation.

We have seen that information and representation are closely connected. In his introduction to cognitive science, Bermúdez writes that 'Hand in hand with the concept of information goes the concept of representation. Information is everywhere, but in order to use it, organisms need to represent it' (2014, p. 24). Bermúdez goes on to define representation as a 'structure carrying information about the environment' (*ibid.*, p. 493). However, this invites an overly-liberal understanding of cognitive representation that fails to distinguish between mere receptors and representations. The tendency to think of receptors as representations is appreciable given that receptors bear interesting correlations to select states of affairs and given the common conflation between the two kinds of information presented above. If we understand content informationally and fail to distinguish between two kinds of information, then we can easily slip into labelling anything whose role might be classed as information bearing as a representation. Indeed, we can grant that receptors bear a class of information because they inform a consuming system about that which they reliably respond to—for example, edge detectors inform

[15] The causal-role account can acknowledge this function's importance without needing to privilege learning periods, selection history or the like—though, again, consideration of such factors may explain how a mechanism came to have a causal role and may serve as a useful heuristic for identifying probable causal roles in the here-and-now. Ultimately, cognitive scientists wish to explain a prominent cognitive capacity like the visual processing of a 3D environment, which causally depends on a mechanism reliably responding to edges, regardless of how the mechanism got there.

their containing system about edges—but, by these lights alone, receptors only bear natural information.[16] Once again, not all information is non-natural information, i.e., representational content. Not all information-bearing entities are representations. [17]

One may fail to shake the feeling that receptors do bear a kind of content in so far as their causal role concerns external entities (that which they reliably respond to). In this way, receptors are 'directed to' external entities, and to this extent, are still 'about things' in the world. For example, edge detectors reliably respond to edges, and that responsiveness to distal properties plays a crucial role in visual perception. As such, one may be tempted to describe edge detectors as being about edges and, therefore, as possessing a kind of content. However, nothing major hangs on this way of talking. This is a weaker sort of content than representational content; a 'natural content', analogous to natural information, that any eliminativist should concede. In fact, this concession is somewhat reflected in the literature when theorists give a semantic characterisation of otherwise non-representational entities. For instance, Piccinini & Scarantino classify natural (non-representational) information as a kind of 'semantic information' (2010). Under their

[16] One may worry that the functional role of receptors no more involves the 'bearing of information' than it does for hearts, firing pins and other causal mediators; all involve reliable correlation with certain states of affairs (for related concerns, see Ramsey, 2007, pp. 132-140). In so far as I adopt a deflationary account of natural information, I think hearts bear natural information about volumes of blood and firing pins bear information about triggers being pulled. But I also think there is reason why we might describe receptors, and not hearts and firing pins, informationally. The fact is that a heart's correlation with blood volume and a firing pin's correlation with a trigger being pulled is uninteresting because the correlation is relatively uniform, inflexible and proximal. By contrast, much of cognitive science involves uncovering the complex web of correlations that occur between neural activity and distal states of affairs. As Piccinini & Scarantino put it, 'Large portions of neurophysiology are devoted to the detection, description, and explanation of the detailed causal correlations that exist between neural responses and variables in the external (distal) environment of organisms' (2011, p. 30). These correlations are of interest because they are often exploited by cognitive systems to direct inferences and behaviour toward states of affairs in interesting ways.

[17] The difference between 'receptors' and 'representations' bears some correspondence to the distinction between two fundamental forms of evolved interaction between an organism and its environment recently proposed by Schulz (2018): those purely reflexive interactions where a stimulus triggers a response, and those representational interactions that require further downstream processing and informational integration. A full analysis of the relationship between the current ideas under scrutiny and Schulz's theory will need to wait for another day.

classification, something can bear semantic information without bearing representational content. From a different perspective, Hutto & Myin also allow semantic notions to come apart from representational content, separating the primitive 'intentionality' of basic cognition from the representational content of complex cognition (2013, 2017).

4.2 Action-Oriented representation

Briefly introduced in **chapter 1**, the broad church of embodied, embedded, extended and enactive ('4E') cognition was, and continues to be, a response to 'cognitivist' and related 'classical sandwich' models of cognition (Hurley, 2011). Amongst other things, cognitivist approaches are characterised by their tendency to depict cognition as the sole purview of the brain, taking cognition and action to be wholly independent processes, and holding the brain to construct detailed, descriptive internal models of the world. To varying degrees and in diverse ways, 4E approaches replace cognitivism and the sandwich model with theories and models that emphasise the role of the agent's body and active engagement with the environment in constituting or enabling cognition. For proponents, perception, cognition and action are closely entangled, and for some, wholly inseparable (for example, Hurley, 2001, 2002).

Many in the 4E movement have been suspicious of cognitive representation in its traditional guise because it seems to go together with the passive, brain-bound view of cognition that they resist. The thought goes that the cognitivist approach erroneously depicts the brain as constructing action-independent, computationally-costly representations of the world with purely 'declarative', or 'factive' contents. Wheeler elaborates, writing that,

> According to the generic orthodox cognitive-scientific model, representations are conceived as essentially objective, context-independent, action-neutral, stored descriptions of the environment. (2005, p. 196)

Action is then conceived of as a kind of mere output, the product of the brain manipulating what Mandick labels 'output representations' (2005, p. 293). Yet in the process of rejecting such ideas, many 4E proponents have held onto the notion of representation, infusing it with their own action-oriented focus. This has given birth to what Clark (1997) dubs 'action-oriented representation' (AO-representation).

Different proponents offer somewhat diverging versions of AO-representation, but there are two typical traits associated with the notion. The first is that AO-representations are sensitive to the biological needs of the representing system. According to proponents, ecologically plausible representations will not involve the construction of action-independent and computationally-costly models of the world but will reflect only those features of the world that are relevant to the system's needs in a time-sensitive manner. The second trait is that AO-representations concern 'commands', 'imperatives' or 'instructions' for action, often framed in terms of what their contents refer to. AO-representations not only represent action-independent states of affairs but how the system can and should act. As Clark puts it, 'These are internal states which […] are simultaneously encodings of how the world is and specifications for appropriate classes of actions' (1997, p. 151). Mandik summarises AO-representations as those, 'that have, in whole or in part, imperative content' (2005, p. 293).[18]

[18]According to Mandik (2005), Clark's (1997) view contrasts with his own in so far as Clark holds that AO-representations are both descriptive and imperative, whereas for Mandik, some AO-representations possess only imperative content. Clark's notion resembles Millikan's (1995) 'pushmi-pullyu' representations, which are defined as simultaneously 'descriptive' and 'directive'.

A guiding principle behind the idea of AO-representation is that traditional treatments of representation associated with cognitivism and the sandwich model are mistaken in thinking that cognitive systems construct descriptively rich representations of the world through passive perception, where action is a mere product of manipulating those representations internally. Rather, it is more biologically plausible that cognitive systems evolve and develop representations that are geared toward organism-specific behaviour, and which efficiently utilise the system's own bodily resources. Such representations 'involve the capacity to support the computationally cheap guidance of appropriate actions in ecologically normal circumstances' (Clark, 1997, p. 152). To illustrate the idea implemented in artificial systems, Wheeler (2005, p. 196) considers a robot, built by Franceschini, Pichon & Blanes (1991), tasked with navigating to a light source whilst avoiding obstacles. The robot's navigation mechanism works by using a layer of motion-detectors that are responsive only to movement and blind at rest. The robot generates a 'snap map' using the previous movements in its own motor sequence to detect objects around it. This is combined with information concerning the angular bearing of light sources to form a 'motor map' which guides the robot's next movement. The content and format of the robot's representation of its environment, it would seem, are deeply dependent on the robot's need for action. These supposed representations also eschew the need to build and store environmental models in favour of using the robot's own bodily resources to help inform the system only of what's relevant to its ecologically situated behaviour. These mechanisms are 'ego-centric control structures for situation-specific actions' (*ibid.,* p. 197).

The central limitation of the AO-representation account is that it does not, by itself, demonstrate that there are theoretical entities with distinctly representation-like functions.

Proponents of AO-representation rightly draw attention to the importance of considering cognitive representation within the context of evolved, resource-bound systems, which in turn informs the architecture and engineering of embodied artificial systems. However, action-oriented approaches have been less concerned with establishing that theoretical posits really do serve as representations. Indeed, many purported cases of AO-representation are not good candidates for representational mechanisms. For instance, Millikan (1995) considers cells located in the inferior premotor cortex of Macaque monkeys, whose firing reliably correlates both with the execution of an action and the perception of the same action in a conspecific, to count as a 'pushmi-pullyu' representation—a version of what others call AO-representation. However, though this correlation with the execution and perception of motor activity may signify the duel importance of these cells for both the execution and perception of action, this does not establish that these cells function in a distinctly representation-like way. Such purported AO-representations may actually function as receptors, or otherwise, possess functional roles that diverge from anything distinctly representation-like.

Gładziejewski (2016a) similarly worries that action-oriented approaches have tended to underemphasise the need to show that a purported representation plays the right functional role in favour of focussing on the importance of action guidance for plausible cognitive mechanisms. He relates this worry back to the JDC, ultimately claiming that such approaches require a supplementary account that stresses 'structural similarity' to ensure that the theoretical entities in question function as representations (see **chapter 4** for a related account). He writes,

> According to this diagnosis, by putting so much emphasis on the role that representations play in controlling actions, proponents of ACToRs [action-

> oriented theories of representation] have lost sight of what is equally important for making representations what they are, namely the fact that using representations consists in exploiting a relation that holds between the representational vehicle and what is represented. (2016a, p. 24. My parenthesis.)

In short, more is needed to demonstrate the explanatory value of representation than an emphasis on ecological plausibility, imperative contents and the like. We also need to demonstrate that a theoretical entity is functioning as a representation in the first place. I do not mean to dismiss the importance of 4E cognition for theorising about cognitive representation—only to acknowledge the limitations of action-oriented approaches when it comes to addressing whether appeals to representation are of value in explanations of cognition. The hope is that action-oriented considerations can be accommodated within a fuller account that properly considers whether and how a cognitive mechanism functions in a distinctly representation-like way. **Chapter 5** will present a view of content that I take to be broadly harmonious with the spirit of 4E cognition.

At this juncture, it is worth mentioning the related 'guidance theory of representation' (Anderson & Rosenberg, 2008). According to guidance theory,

> [R]epresentational content is derived from the role a representational vehicle plays in guiding a subject's actions with respect to other things. (*ibid.*, p. 68)

Guidance theory also emphasises the importance of characterising representational content in terms of ecologically plausible features of a system's real-world need for action in its environmental niche, and a rejection of accounts of content that stress 'objective environmental conditions' (*ibid.*, p. 64). As it stands, guidance theory alone, like many action-oriented approaches, does not provide us with an adequate defence of how theoretical entities in cognitive science serve as representations in the first place. At times,

Anderson & Rosenberg imply that representational content is found wherever there is an internal state that guides actions toward objects in the world. This permits the same weak conception of representation cautioned against already. For instance, Anderson & Rosenberg classify simple fly-detection mechanisms in frogs as representations. In such cases, retinal ganglion neurons reliably activate in the presence of fly-like stimuli, causing activation in the optic tectum which subsequently causes fly-catching behaviour (Lettvin Maturana, McCulloch & Pitts, 1959). Yet such mechanisms seem to function as receptors and are not distinctly representation-like (Gładziejewski, 2016a). Therefore, such examples do not demonstrate the value of representation in explanations of cognition.

Anderson & Rosenberg write, 'Our contention is essentially that representations are what representations do' (2008, p. 56). However, when we recognise that representing is a non-trivial functional role for a potential subset of cognitive mechanism, we appreciate that representation is not just about *what* a cognitive mechanism does for a consumer (for example, causes spatial navigation), but about *how* a mechanism does it (i.e., causes spatial navigation by standing-in for the environment). When it comes to evaluating whether a mechanism is a representation, the means of action guidance is as important as the action guidance itself.

4.3 Intentional stance representation

Section 2 explicated what it means for a subpersonal entity to have the function to represent in terms of a mechanism with a certain causal role in producing or constituting a cognitive capacity. One might resist this picture by arguing that it wrongfully assumes representational explanations are only justified when they identify 'concrete' theoretical entities, akin to other scientific posits, like DNA, quarks or proteins, but that this is not

the role of representational explanations. Such might be the thinking of a proponent of Dennett's 'intentional stance theory', briefly introduced in **chapter 1**. This brings us to our final variety of representation to be explored in this chapter: 'intentional stance representation' (IS-representation).[19]

According to Dennett's intentional stance theory, we can (and regularly do) usefully predict the behaviour of many systems, including humans, non-human animals and some artefacts, in terms of representation. By 'representation', Dennett principally has in mind the 'propositional attitudes' of folk psychology, such as beliefs and desires (see **chapter 3** for more on folk psychology). This strategy operates from a certain interpretive stance called the 'intentional stance'. When we adopt the intentional stance towards a system, we treat that system as rational. By treating a system as rational, we treat its behaviour as interpretable in terms of reasons. Beliefs and desires (and other propositional attitudes) are just those reasons we offer to make sense of a system from the intentional stance. The beliefs and desires we attribute to a system are those that best make sense of the system as rational given its capacities, biological needs and biography.

The intentional stance is most transparently at play when we interpret the behaviour of other persons. However, the intentional stance can also sometimes predict behaviour below the level of whole persons too. A system can sometimes be decomposed into parts that can be interpreted using the intentional stance, so long as treating those parts as rational remains predictively effective. To this extent, whole agents can be taken to be composed of mini-agents, or 'sub-persons' (for a recent discussion, see Drayson, 2012;

[19] Ramsey (2007, chapter 5) also discusses Dennett's work in relation to the JDC but his presentation and discussion differs somewhat from that given here, taking place within a larger critique of 'tacit representation'—the idea that a system represents knowledge implicitly through the dispositions of its cognitive machinery rather than explicitly through discrete, identifiable states.

Huebner, 2018). Again, what systems count as possessing beliefs and desires, and what those beliefs and desires are, is a matter of whether and what propositional attitudes predict a system's behaviour. Dennett summarises his own position best when writes,

> Here is how it works: first you decide to treat the object whose behaviour is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in most instances yield a decision about what the agent ought to do; that is what you predict the agent will do. (1987, p. 17)

In short, by interpreting the behaviour of a system through the intentional stance we assume that target system is (i) rational, (ii) possesses reasons, and (iii) is predictable given the conjunction of (i) and (ii).

The intentional stance is not the only stance that we can adopt when interpreting the behaviour of a system. Two others stand out for Dennett: the 'physical stance' and the 'design stance'. The physical stance allows us to interpret a system through the lens of physical properties and causal laws (think physics and chemistry). The design stance allows us to interpret a system through the lens of design principles, function and teleology (think biology and engineering). The shift from the physical stance to the design stance to the intentional stance is a shift in the level of increasing abstraction with which one interprets a system. As one goes up a level, one loses accuracy but gains computational traction, essentially by dropping irrelevant and computationally costly details. What matters, from the perspective of the intentional stance, is not the physical or design details of a system, but whether a system is usefully interpretable in terms of intentional idioms.

An important result follows: the validity of IS-representation, that is, propositional attitude ascriptions from the perspective of the intentional stance, does not hinge on empirical details about the inner causes of cognition. After all, the intentional stance can operate without any knowledge of how a system's machinery is structured and organised; what matters is that the gross behaviour of the system in question is predictable by treating it a rational with appropriate beliefs and desires. In this way, the success of the intentional stance is indifferent to how a system works *mechanistically* (Dennett, 1969, 1987). By contrast, the kind of representation that we are searching for reflects important details about the machinery of subpersonal cognition: the properties and relations within and between its spatiotemporal parts that are causally responsible for behaviour. In this way, the intentional stance does not identify entities that would pass the JDC.

The preceding conclusion might be thought to have one of two implications: either the propositional attitude ascriptions of the intentional stance do not imply genuine representation; or the JDC enforces overly restrictive conditions (it simply turns out that legitimate representational explanations do not require positing the sorts of functional mechanism that we have been gesturing toward). To think either of these is to miss the point. The intentional stance is playing a different game from the mechanistic explanations of contemporary cognitive science. The intentional stance treats the system it is interpreting as a black box, abstracting away from mechanistic details to predict behaviour at a less demanding level. In other words, IS-representations are not *cognitive* representations. Shea similarly distinguishes between the propositional attitude representations that Dennett's intentional stance targets from the 'neural representations' of cognitive science (2018, p. 14). Pöyhönen also notes that 'Dennett's disregard for the mechanistic constraints in explanation implies that his theory should not be interpreted as

a full-blown theory of causal explanation' (2014, p. 104). Rather, so far as it goes, 'the explanatory strategy of the intentional stance consists in analysing rational behavior into sub-tasks that themselves can be described as if they were rational.' (*ibid.*, p. 104). This also reminds us that the intentional stance is concerned with representation within a rational framework. And yet, rationality is not a necessary constraint on the justifiable attribution of subpersonal mechanisms that much of cognitive science is concerned with, for instance, cognitive maps that underly spatial navigation (see **chapter 4**).

In summary, Dennett's intentional stance theory attempts to capture the power of representation ascriptions from the viewpoint of a particular interpretive strategy that is independent of the representation ascriptions of mature cognitive science. It is because of this difference in explanatory goals that the intentional stance is unconcerned with subpersonal representational mechanisms. As a result, one could accept Dennett's intentional stance theory whilst accepting either representationalism or eliminativism about subpersonal representational mechanisms.

In this section, we have seen that the receptor representation, action-oriented representation and intentional stance representation notions all fail to guarantee entities that pass the JDC. The receptor and AO-representation account do not, by themselves, identify mechanisms or analogous entities that play a distinctly representation-like role on behalf of a cognitive system and the IS-representation account is not concerned with identifying subpersonal mechanisms at all. An important consequence follows: one could consistently champion some version of global eliminativism—in the sense of rejecting the explanatory value of subpersonal representational mechanisms—whilst simultaneously advocating the explanatory value of the entities posited by these accounts.

This is because receptor-representation, AO-representation and IS-representation do not, by themselves, imply full-blooded representation in the sense that should matter to the eliminativist. As a result, we must look elsewhere if we are to find a convincing defence of representation's role in explanations of cognition.

5. Conclusion

This chapter has shown that any successful account of cognitive representation must make sense of how a subpersonal entity can function as a representation on behalf of a cognitive system, such that content/non-natural information is causally relevant to that cognitive system. I have argued that the mechanistic account of representation provides the broad outlines of how this might be possible. Drawing on the view that cognitive science offers mechanistic explanations, it says that a cognitive representation is a mechanism that has the function to stand-in for something on behalf of a containing system, such that its causal role (*qua* stand-in) contributes to the realisation of a cognitive capacity.

I have also argued that several popular notions of cognitive representation do not show that representational mechanisms (or analogous entities) do serve or could serve a role in theories of cognition. The next two chapters will explore additional species of so-called representation common to the scientific and philosophical literature. The first suggests that our best theories of cognition posit representations because our best theories of cognition are computational (see **chapter 3**). Unfortunately, support for the explanatory significance of cognitive representation is found lacking here too. It will take a new notion, that of 'S-representation', to show how cognitive representation could serve a robust explanatory role, by detailing how a mechanism could function in a distinctly representation-like way (see **chapter 4**).

Chapter 3

Computation and Content[1]

1. Introduction

This chapter explores the relationship between computation and cognitive representation. There is a common assumption that our best theories of cognition imply cognitive representation because our best theories are computational, and computation presupposes representation. Fodor writes,

> My point will be that not only considered action, but also learning and perception, must surely be viewed as based upon computational processes; and, once again, no computation without representation. (1975, p. 34)[2]

Anti-representationalists also speak of computation and representation as mutually supporting pillars of received cognitive science (for example, see Garzón, 2008), and often take the elimination of one to entail the elimination of the other (for example, see Varela, Thompson & Rosch, 2016).

And yet, on closer inspection, the relationship between computation and representation is less clear-cut than many imagine. This chapter argues that computational theories of cognition are not necessarily representational theories of cognition. This is because computation and representation are distinct functional kinds. Some computing systems,

[1] Portions of this chapter appear in Lee (forthcoming a).

[2] Sometimes Fodor seems to suggest that the inference is the other way around: cognition involves the transformation of representations; therefore, cognition is computational (for example, see Fodor & Pylyshyn, 1981, pp. 139-140). Read at face value, one may fear an encroaching circularity.

including cognitive systems, may represent, but computation does not entail representation.

The primary argument addressed by this chapter can be summarised as follows:

P1. There is no computation without representation.

P2. Our best theories of cognition are computational.

C. Our best theories of cognition are representational.

Our focus will be P1. In contrast to the traditional idea that 'semantic properties'—typically equated with representational properties—are necessary for individuating computation, I contend that all that is required for individuating computation is some mechanism that transforms medium-independent digits in accordance with formally specifiable rules. As such, there is computation without representation. I will also explore and reject the idea that there is something special about computation in the context of cognition that entails representation; in particular, I reject both the claim that computational explanations of cognition imply representation because they vindicate the propositional attitudes of folk psychology and the claim that computational explanations of cognition imply representation because of their task-decompositional structure.

The chapter proceeds as follows. **Section 2** begins by examining the 'semantic view' (SV). Semanticists think that computation presupposes semantic properties because semantic properties are required to individuate computation. I take this to be the 'received view' of computation (Sprevak, 2010), and the source of many claims about computation's representational nature. I then discuss two problems with the SV. **Section**

3 sketches a promising alternative, the 'mechanistic view' (MV), before discussing a counter-attack on behalf of the semanticist. I close this section with a further challenge to the representationalist who relies on the SV, questioning whether the semantic properties invoked by the SV are really representational properties. **Section 4** looks at the 'classical computational theory of cognition' (CCTC), and with it, the 'representational theory of mind' (RTM). I claim that the argument for representation in the orthodox RTM is not satisfactory. This is because the RTM either (a) adopts the semantic view, or (b) incorrectly assumes that if the CCTC provides a naturalised reduction base for propositional attitudes, then the CCTC implies cognitive representation. I close by critiquing the notion of 'input/output representation' in cognitive science.

2.1 What is computation?

Debates over the nature of computation typically begin by considering the Turing Machine. The Turing Machine (originally, 'the automatic machine') was the first formalisation of computation as a concrete concept (Turing, 1936/2004). It is a mathematical model of an abstract machine designed to perform rote operations of the sort historically performed by humans (the term 'computer' originally referred to a person who performed calculations).[3] The Turing Machine offers a specification for a machine that automates the kinds of rote mathematical operations human computers would perform using fixed, formal methods, on paper or in their heads. A Turing Machine consists of an infinite tape (avoiding storage limitations) divided into cells that contain

[3]Turing later wrote that electronic computers are 'intended to carry out any definite rule of thumb process which could have been done by a human operator working in a disciplined but unintelligent manner' (1950, p. 436). If the mind is a kind of physical Turing Machine (as Turing thought), it does not follow that humans compute in the same way that an electronic computer computes—nor that Turing thought so. For discussion on Turing's views on the relationship between computation and mind, see Boden (2006, chapter 4).

'symbols' or 'digits' taken from a finite alphabet,[4] a read/write head, a state register, and a look-up table containing a finite set of 'instructions' or 'rules'. The Turing Machine 'reads' the symbols on the tape, then 'manipulates' the symbols by moving the head, and erasing, writing or ignoring the symbols. The instructions in the machine's table entirely determine what manipulations it will perform given (i) the current state of the machine's state register, and (ii) the symbol the machine reads. In turn, this captures the process of the human computer, whose behaviour 'at any moment is determined by the symbols which he is observing, and his "state of mind" at that moment' (1936/2004, pp. 75-76). Note that this description does not make any explicit reference to representation.

The Turing Machine is an abstract model that shows how, in principle, an automated device could compute any sufficiently well-specified problem. Though computation can be studied solely from the perspective of mathematical formalisms, this chapter is concerned with 'physical computation'. This is because we are interested in computation as it relates to cognition. By physical computation, I mean physical processes performed by physical systems, where the fact that computation is performed explains some behaviour of those systems.[5] Here I limit myself to consideration of physical 'digital computation'—computation performed over discrete variables (as implied by the Turing Machine). The other notion of physical computation relevant to cognitive science is 'analogue computation'—computation performed over continuous variables.[6] I limit the scope of the discussion for brevity and because historically most computational-

[4]In computer science, an alphabet is a finite, non-empty set of elements commonly referred to as 'symbols' or 'digits', over which operations are defined. The most common set in artificial computing systems is the 'binary alphabet' {0,1}. Elements of an alphabet combine to form 'strings'.

[5]From an anti-realist perspective, a physical computer is a physical system whose behaviour is best explained by treating that system as computing.

[6] This way of defining the digital-analog distinction is not entirely without controversy. However, weighing in on this debate is not necessary for our purposes. For some discussion, see Maley (2018).

representational approaches to cognition have assumed digital computation. Nevertheless, much of the discussion below will apply to analogue computation too. In summary, 'computation' hereafter refers to physical digital computation.

A final clarification: I will sometimes be required to use the term 'symbol' and 'digit' interchangeably given their use in the literature. I will use the term 'digit' where possible. This is because the term 'symbol' connotes representational content—the very thing at stake. The more neutral term 'digit' is suited to capture what is generally agreed upon by semanticists and non-semanticists alike: that computation involves physical input states whose functional properties, alone or alongside the present state of the system and a set of instructions, determine the production of a physical output state.

Accounts of physical computation draw from their theoretical roots. As such, almost everyone acknowledges that physical computation involves the transformation of input states to output states in accordance with formally-specifiable rules. What they disagree on are the details and implications of this gloss. In the remainder of this section, I will sketch the semantic view (SV) of computation. If the SV is correct, physical computation presupposes semantic properties because semantic properties individuate computation. Under an orthodox interpretation, these semantic properties are equivalent to representational content (though see **section 2.4** below).

2.2 The semantic view of computation

Most would agree that not all representations involve computation. The Mona Lisa, Michelangelo's David and the Arc de Triomphe are unlikely candidates for computing

systems. If correct, then representation is not sufficient for computation.[7] However, the semanticist claims that semantic properties are necessary for computation. A brief detour will show that there are in fact two ways of interpreting this claim.

The first interpretation concerns the conditions that make a physical system a computing system as opposed to something else. This raises the following question: under what conditions does a system implement computation? Here, the semanticist thinks that semantic properties form part of the necessary criteria for a system to count as performing computation *simpliciter*. Typically, these semantic properties are cashed out in terms of representation: for a system to count as performing computation, it must represent something. This criterion is then used to show how computing systems are demarcated from other systems by demonstrating that only a subset of systems—of an appropriate size and kind that reflect our intuitions about paradigmatic computing systems—meet the representation requirement.

The second interpretation concerns the question of what individuates computations themselves. When we ask what individuates members of *x* we ask by what criteria different members of *x* are determined, or equivalently, what makes two or more members of *x* of the same or different kind. This second interpretation raises the following question: what conditions determine the identity of a computation? Here, the semanticist thinks that semantic properties are necessary for distinguishing between the realisation of different kinds of computation, where equivalence and difference of computation are understood by reference to the contents represented by input and output states.

[7]Even if pancomputationalism is correct, and paintings, marble statues and triumphal arches do compute, they do so because just about *everything* computes, not because they are representations. According to pancomputationalism, collar bones, pogo sticks and asparagus compute too (see below, this section, for further discussion).

These two interpretations of the semanticist's claim are asymmetrically related. If representation is necessary for individuating computation, then it follows that for any system to perform computation, it must represent something (i.e., the second interpretation implies the first). However, one need not justify the claim that computation presupposes representation by holding that computation is individuated by representation. In principle, one could hold that all computation involves representation, without thinking that the contents of the representations involved individuate the computation being performed (i.e., the first interpretation does not imply the second)—though, to my knowledge, no one holds this view.

Distinguishing between these two interpretations aids our appreciation of the conceptual landscape. In practice, I take the orthodox semanticist to accept both claims. Accordingly, computation is individuated by representational content, and in turn, computing systems are set apart from other systems, in part, by representation. Shagrir captures the scope of the orthodox SV when he writes,

> The semantic view claims that the individuation of these systems (etc.) takes into account their semantic properties. This means that semantic properties play a role in determining whether a certain system (etc.) is computing or not, whether two systems (etc.) are computationally similar or computationally different, whether or not changes in semantic properties alter computational identity, and so on. (Forthcoming. Manuscript, p. 3)

Variations on the SV can be found throughout the literature (for instance: Crane, 1990, 2016; Churchland & Sejnowski, 1990; Shagrir, 2001, 2006, forthcoming; Ladyman, 2009; Sprevak, 2010; Rescorla, 2012). Crane captures the shared spirit succinctly when he writes,

> So what is essential to a computer? The rough definition I will eventually arrive at is: *a computer is a device which processes representations in a systematic way*. (2016, p. 59. Original emphasis.)

According to the SV, for instance, to understand a spell-checker as performing computation we must reference the manipulation of states that represent letters, and to understand a calculator as performing computation we must reference the manipulation of states that represent numbers. Physical computers manipulate physical vehicles, but those vehicles are individuated by what they represent. In this way, the symbols of computation have a 'dual character', at once both semantic and physical (Egan, 2010, p. 253). The argument for the orthodox SV can be summarised as follows:

> Pi. If *y* is individuated by *x*, then there is no *y* without *x*.
> Pii. Computation is necessarily individuated by representation.
> C. There is no computation without representation.

Notice that the conclusion here is identical to P1, the premise that is needed to establish the earlier, primary argument for deriving cognitive representation from computation. The present argument for the SV is valid. I take Pi to be self-evident. The argument then rests on Pii. Putting this together, the argument for the necessity of representation from computation depends on whether computation really is individuated by representation.

A prime source of support for Pii is that representation provides constraints with which to minimize the number and kind of systems that count as computing in a way that is approximately consistent with our intuitions about which systems fall under the umbrella of computation. For example, computation is often framed as the transformation of an

input state into an output state in accordance with a formally specifiable (mathematical/logical) function—this is chiefly how a computing system's instructions or rules for manipulating digits are defined. However, very many physical systems can be described in terms of instantiating such functions. The orbit of the planets, for example, can be described as instantiating Newton's second law of motion (F = ma), taking two values as inputs (corresponding to the planet's mass and acceleration), performing a multiplication function on them, and producing another value as output (corresponding to the force exerted on the planet). At its most extreme, the 'mapping view' of computation states that a system implements a computation iff the system contains a sequence of physical states that can be mapped onto the sequence of states specified in a formal description of a computation (Putnam, 1988; Searle, 1992; Godfrey-Smith, 2009). Proponents of the mapping view often see the multiple-realizability of computation as giving way to trivial-realizability because the formal states of a computation can be mapped onto any large enough open system (i.e., a system with input/output relations of sufficient cardinality). As such, any arbitrary computation will be implemented by any large enough open system, giving way to a kind of 'pancomputationalism'. This is an undesirable outcome, both because it goes against our intuitions that only a limited number of systems are computing systems and because it undermines the significance of computational explanation: if everything computes, then explanatory appeals to computation are uninformative.

The SV offers an alternative to the mapping account. It evades pancomputationalism by adding the constraint that a computing system must represent. Most physical systems do not represent, after all—planets do not represent their mass or their acceleration. In turn, representation marks a distinction between physical systems that may be described

computationally from those that actually perform computation. Crane (2016) embodies this perspective when he writes,

> What the adding machine really does is take *numerals* – that is, representations of numbers – as input, and gives you numerals as output. This is the difference between the adding machine and the planets: although they instantiate a function, the planets do not employ representations of their gravitational and other input to form representations of their output. (p. 71. Original emphasis.)

Therefore, representation constitutes a vital ingredient for determining whether a system is truly performing computation. Representation also provides plausible grounds for 'computational equivalence'. Sometimes, two systems that take distinct physical vehicles as inputs and outputs appear input/output equivalent in a manner that matches our intuitive categorisation practices. Let's borrow an example from Sprevak (2010, p. 268). Imagine two different devices. The first device reads ink-marks shaped like Roman numerals (I, II, III etc.) as input, and produces the same as output. The second device reads ink-marks shaped like Arabic numerals (1, 2, 3 etc.) as input, and produces the same as output. Despite the differences in the inputs and outputs, these two systems could perform the same numerical calculation, such as the addition function. These systems appear to have equivalent inputs and outputs. What does this equivalence consist of? One plausible answer, the one that Sprevak (2010) offers, is that the inputs and outputs represent the same thing. Representation thus accounts for computational equivalence.

2.3 Two problems with the semantic view

The SV is an intuitively appealing account that provides a way to rescue computation from the threat of triviality. Nonetheless, there are two related worries that call into question the truth of Pii. As these worries have been discussed at length elsewhere (for

example, see Piccinini, 2008, 2015; Miłkowski, 2013; Dewhurst, 2018), and will be further drawn out in the rest of the chapter, I will limit myself here to a brief sketch of each. There is also a third, less well-explored problem concerning the assumption that the semantic properties required by the SV are full-blown representational properties in the sense that matters to cognitive representation. I will turn to this in **section 3.3** below.

The first worry is that the SV suffers counterexamples. As Pii is a necessity claim, only one counterexample is needed to prove that the claim is false. Proponents of the SV defend Pii by claiming that paradigmatic computing systems are intuitively thought to manipulate representations, and that manipulation of representations is key to the character of the computation in question. For example, an ordinary syntax checker manipulates electric signals that represent strings of letters (Sprevak, 2010, p. 268). However, even if this correctly characterises our understanding of a syntax checker as a computing system, there appear to be cases where states are not required to represent anything for us to recognise and identify computation. For example, one might construct a machine that takes some strings of entirely arbitrary digits as input, performs an arbitrary (but formally specifiable) operation on those inputs, then outputs further arbitrary digits. Such a machine looks very much like a paradigmatic computing device—but not because it represents anything. Imagine, for instance, a physical device that is fed marks on a page as input, taken from the following complete alphabet: {~, <, *}. The machine takes one of these marks as input and produces another mark as output. The machine has one rule which can be captured as follows: ‘if the input is * write < otherwise do nothing’. We do not need to characterise this machine as representing anything for us to recognise it as a computing system. This recognition is possible because there are physically distinct digit types that are transformed in a nomic fashion (mirroring the

character of a Turing Machine). One might insist that the system *does* represent something: the marks it reads. However, though we must acknowledge that the system is sensitive to the marks, it seems superfluous, for the purposes of identifying computation, to say that the system represents the marks. We recognise inputs, outputs, and transformation rules, even in the absence of characterising the marks in terms of representation. As Orlandi puts it,

> [D]igits in a computer do not need to stand for something in order to count as digits. As long as a state is distinguishable from others in different contexts, and as long as it enters into lawful transitions with other states, it is a digit even if it has no semantic interpretation. (2014, pp. 206-207)

In turn, these digits and their transformations provide a basis for individuating computation utilising only the number of distinct digit types and the transformation rules that process them. In so far as this is true, appeals to representational content are not necessary for recognising computation or characterising the computation being performed. This anticipates the mechanistic view to be sketched momentarily.[8]

The second worry about the SV is that any semantic individuation is parasitic on prior non-semantic individuation and that this undermines the necessity of the former. To begin, notice that it is unclear in what way, if any, the semantic properties of digits are causally relevant to computing systems. After all, calculators do not manipulate electrical states in virtue of what their states represent (for some pushback see Rescorla, 2014). As the useless computer example helps to show, it is often assumed that computing systems are causally sensitive only to the 'form' of their digits (for an early formulation of this idea as it pertains to cognition, see Fodor, 1980). The form of a digit refers to the abstract

[8] For a different take on a similar point see Piccinini (2008); Miłkowski (2013).

property of the physical vehicle that comprises the digit and which the computing system's transformation rules respond to. Types of digits and internal states of the system are mutually defined by the number and kinds of form which the system is sensitive to and the lawful ways in which the system responds to those digits.

The fact that computing systems are causally sensitive to the formal features of their digits implies that for any given semantic individuation, a non-semantic 'notation' is required to pick out the relevant inputs and outputs (Piccinini, 2015, p. 37). For example, in the case of a logic gate, semantic individuation presupposes a way of individuating the inputs and outputs of the device in terms of the difference between digits instantiated in the difference of voltage levels to which the device is sensitive to (see **section 3.2** below). In traditional parlance, any 'semantic' individuation supervenes on a prior 'syntactic' individuation (though computing systems need not have a syntax in the sense of a grammar-like structure).

We can better appreciate the point being made by observing that traditional theories of content in computation cannot, by themselves, individuate computation. If one claims that computing systems necessarily represent, as the semanticist does, then one must ultimately deliver a supplementary account of how states acquire their content. The SV itself is neutral as to which account of content determination is correct. However, the semanticist is committed to there being some correct account of content determination down the line. Thus, the truth of the SV depends on the possibility of showing how, say, a calculator represents numbers, by what means a syntax checker represents strings of letters, and for the representationalist who justifies cognitive representation on the basis of the SV, in virtue of what cognitive systems represent the states of affairs that they do.

Several theories of content determination are popular amongst semanticists. Each of these theories falls under the umbrella of either 'naturalistic accounts', conceiving of content as determined by a naturalistic relation, or 'non-naturalist accounts', conceiving of content as determined by an agent's attribution of content. For illustrative purposes, I will defer to Piccinini's (2015) taxonomy of the main players associated with the SV and will briefly outline each. Firstly, there are theories which fall under the umbrella of 'functional role semantics' (for example, Block, 1987). FRS theories are naturalistic. FRS theories claim that states have contents in virtue of the functional relations that obtain between other states, inputs, and outputs that exist internally within a system. In this context, those functional relations are understood to be computational relations. Secondly, there are 'causal-informational' accounts, equivalent to what I called causal-historical theories in **chapter 1**, which includes Dretske's (1981, 1988) causal-informational theory and teleosemantics (Millikan, 1989b, 1990). Causal-informational accounts are naturalistic. Causal-informational accounts claim that states acquire their content in virtue of some privileged causal or informational relation that exists between the state and what it represents. The precise nature of this relation differs depending on the theory. Lastly, there is 'interpretivism' or 'interpretational semantics' (Dennett, 1969, 1987; Cummins, 1983, 1989). Interpretivism is a non-naturalistic theory. [9] It claims that states have their content fixed by an observer's semantic interpretation of the underlying syntactic or other functionally defined processes. Interpretivists deny that computation involves manipulating 'natural representations' and claim that there is no intrinsic relation which fixes a digit's content. Rather, the sort of representation involved is 'conventional

[9] The SV is not wed to a naturalised theory of content. As Shagrir writes, 'the debate about the semantic view is not about naturalism […] the semantic view is consistent with, but not committed to, the view that all computational contents can be naturalized' (Forthcoming. Manuscript, p. 26).

representation'. The reason why a calculator represents numbers rather than something else (or nothing at all), is because an agent decides so. Note that 'hybrid theories' of content are also popular amongst semanticists (Shagrir, forthcoming). One common view holds that artificial computing systems, like calculators, have their contents in virtue of an observer's interpretation, but that natural computing systems, like brains, have their contents in virtue of some privileged causal relation.

Each of these accounts of content depends on the prior non-semantic individuation of computing systems and their states. As Piccinini points out, FRS relies on a prior notion of there being computationally relevant states that acquire their semantic content in virtue of the relations that obtain (2015, p. 33). A similar story can be given for causal-informational accounts. These require a way of individuating the states in which the relevant semantic-bestowing relation holds. Finally, the interpretivist relies on there being a principled way of picking out the relevant processes and their states which are subject to semantic interpretation. Every theory of content determination requires some pre-semantic individuation criteria.

The semanticist may respond that this second worry only demonstrates what everybody already agreed: that representation is not sufficient for computation. This does not preclude the possibility that representational content is jointly necessary for computation. However, when coupled with examples like the useless computer above, we begin to see the possibility that non-semantic features are not only necessary but sufficient; non-semantic features of (at least some) systems are exhaustive of those systems'

computational properties.[10] To fully appreciate this, however, we need to turn to an alternative view that takes as its starting point the need for a physical parallel to the (non-semantic) mathematical notion of computation, and places at its heart the lawful transformation of medium-independent digits. Enter the 'mechanistic view'.

3.1 The mechanistic view of computation

This section begins by charting the mechanistic view (MV). This will help to anchor the remainder of our discussion on the relation between computation and representation. The MV constitutes an instructive and promising alternative to the SV. It encompasses the intuitions raised above by the criticisms of the SV and resonates with the broader mechanistic approach to explanation outlined in **chapter 2**.[11] Though proponents of a mechanistic approach to computation differ in the details they espouse, I will outline those characteristics that I take to be most common and relevant for our present discussion.

The MV approaches physical computation as a species of mechanistic activity resulting from the properties of a physical system (Piccinini, 2008, 2015; Fresco, 2014; Miłkowski, 2013; Orlandi, 2014). Computation is a functional kind in the teleological sense. A computing system is, in turn, a kind of functional mechanism, where a mechanism is defined as a system comprised of organised spatiotemporal components (see **chapter 2**). Paradigmatically, such components might include strings of digits, processors and memory stores and the relevant relations between such components would include signal transmission (Piccinini, 2015).

[10] Some semanticists may think that I have under-emphasised a crucial step in their argument: semantic properties are necessary for individuating computation in line with the logical/mathematical operations we take computations to perform. I confront this point in **section 3.2** below.

[11] In accordance with the broader mechanistic framework outlined in **chapter 2**, we can understand the brain as a computing system just in case it contains a mechanism that contains at least one component with the causal role of computing relative to the production or constitution of some cognitive capacity.

Mirroring the mathematical conception of computation, mechanists typically take the function to compute as the function to transform strings of medium-independent digits according to a rule that is sensitive only to a subset of the vehicles' physical properties (for example, see Piccinini, 2015, pp. 125-126). A rule is a lawful mapping of input states to output states. Digits are 'medium-independent' in so far as rules by which they are manipulated may be implemented in different physical media (Garson, 2013; Piccinini, 2015). This is possible because an input/output (I/O) mapping is not characterised by the way particular types of physical states (electric, magnetic, mechanical etc.) are transformed, but by the variation in the degrees of freedom of the vehicles that the mapping is sensitive to.[12] Therefore, the same transformation can be implemented in different physical systems so long as the appropriate degrees of freedom are preserved and manipulated correctly by the mechanism (Piccinini, 2015, p. 122). Digits are understood to interconnect to form strings. As Piccinini says, 'A *string* of digits is a concatenation of digits, namely, a structure that is individuated by the types of digits that compose it, their number, and their ordering (i.e., which digit token is first, which is its successor, and so on)' (2007, p. 107. Original emphasis). Digits and the rules that describe their manipulation are mutually specified: rules are defined by the kinds of nomic manipulations the systems perform on types of digits, and types of digits are individuated by the portion of the vehicles (degrees of freedom) that the system is sensitive to.

[12]In his more recent work Piccinini presents a 'generic account' of computation that encompasses both digital and analogue computation (2015). The main difference between that account and the one presented here is that 'digits' are replaced with the more encompassing notion of a 'vehicle'. This latter allows for computation to be defined over the manipulation of both discrete and continuous variables. I present the orthodox 'digital' version of the mechanistic account for the reasons noted above and because it resonates most straightforwardly with other authors working within the burgeoning mechanistic paradigm.

For the MV, any ‘symbols’ necessary for individuating computation are not defined by their semantic character. Again, they are equivalent to digits, which are physical vehicles that fall under different functional types, defined by number and kinds of transformation a system performs on them. Symbol strings are ‘to a first approximation, a physical realization of the mathematical notion of string’ (Piccinini, 2007, p. 108). Miłkowski puts it as follows,

> From the semiotic point of view, a formal symbol is a sign whose function is determined merely by its form (which is not to be identified with its shape, as not all vehicles have visual forms; a form is an abstract property of a vehicle). (2013, p. 36)

This characterisation of computing symbols does not preclude the possibility that some computing systems represent, or that semantic interpretations of digits form part of our conventional relationship with computing systems. However, the MV does insist that semantic properties are not essential for digits to play a role in computation. In short, the manipulation of ‘symbols’ is the manipulation of functionally-typed physical states. This strengthens the idea that computation and representation are functional kinds that come apart.[13]

3.2 The semanticist strikes back: Individuating computation by task

So far, we have examined the claim that computation is individuated by semantic properties. We looked at two problems facing that claim and outlined the mechanistic

[13]It was noted above that one appeal of the SV is that it evades pancomputationalism. The MV supplies a non-semantic solution to this same problem. In brief, some physical systems count as computing systems because they are mechanisms with spatiotemporal components that function to manipulate strings according to rules defined over medium-independent digits (relative to explaining some capacity of the system). Like other mechanisms and their functions (hearts, thermostats and engines etc.) it is generally supposed that there will exist a limited class of entities whose capacities are explained by the appropriate causal structures required to be legitimately typed as performing computation (Piccinini, 2015, chapter 7).

view as a non-semantic alternative. In this section, I will assess what I take to be the strongest response on behalf of the semanticist. Doing so will ultimately strengthen the mechanistic view.

The semanticist counterattack I wish to consider was anticipated in the above example from Sprevak (2010) of two seemingly equivalent computations. It appeared that two devices performed the same computation because they performed the same mathematical operation. Reflecting on such considerations, it has been claimed that computation must be individuated with respect to the operations or tasks commonly assigned to computing systems (typically logical/mathematical functions), and that semantic properties are required to achieve this. In general form, the argument is as follows:

> Pa. Computing systems perform tasks/operations (e.g., mathematical/logical functions).
>
> Pb. Tasks/operations require semantic individuation.
>
> C. Therefore, computation requires semantic individuation.

Take the example of a logic gate. Logic gates are paradigmatic computing devices that take one or more input, usually of a binary form, perform some function, and produce some output, usually of a binary form. These inputs/outputs are typically instantiated by voltage levels. Nothing thus far indicates the need for semantic individuation. Plausibly, a logic gate's activity is individuated (*qua* computation) by the number of physically instantiated digits the device is sensitive to and the transformations they enter into. Two logic gates perform the same computation when the number of digits and their nomic manipulations (regardless of their physical basis) remain the same.

At this point, the semanticist will point out that logic gates are conventionally individuated in accordance with the logical operation they perform (AND, OR, NAND, NOR etc.) For example, a logic gate would implement the AND operation iff its output was 1 (or 'True') in case both inputs are 1, and 0 (or 'False') in all other cases. A logic gate would implement the OR operation iff its output was 0 just in case both inputs were 0, and 1 in all other cases. But logic gates do not operate on numbers or truth values. Logic gates operate on physical states. The problem is that which states of a logic gate correspond to 0 and which states correspond to 1 is underdetermined by mere consideration of the device's transformation of, say, electrically implemented digits. Sprevak (2010) employs a version of this idea to argue in favour of the SV. Imagine a logic gate with the following functional profile:

Table 1. AND or OR gate?

Input a	Input b	Output
0V	0V	0V
0V	5V	0V
5V	0V	0V
5V	5V	5V

Intuitions tend toward the conclusion that the gate implements the AND operation [0V = 0; 5V = 1] because we tend to associate higher numbers together and lower numbers together (Dewhurst, 2018). However, the profile shown in **table 1** is consistent with the OR operation too! We simply need to switch our labelling around [0V = 1; 5V = 0]. In the absence of any prior reason to suppose that 5V = 1, our labelling is arbitrary.

Therefore, the mere consideration of the I/O table above underdetermines the Boolean operation performed. Moreover, because the above description is consistent with more than one computation being implemented, the device appears to perform multiple computations simultaneously. This is the so-called 'problem of multiplicity' (Coelho Mollo, 2018).

The mechanist claims that the transformation of (non-semantically defined) digits is sufficient for individuating computation. But such intrinsic properties of a computing mechanism underdetermine which logical operation the above system performs. The semanticist insists that individuation should make sense of such paradigmatic operations and promises a way out of this problem: we can determine which operation is performed if we suppose that different vehicles represent different truth values. In other words, a logic gate performs AND or OR because the states of the logic gate represent one or other truth value.[14]

There are two main strategies available to the mechanist in response. I will label these the 'narrow-functional' and 'wide-functional' strategies. The narrow-functional strategy concedes that computational individuation underdetermines which logical/mathematical operation a computing system performs according to the MV. However, it maintains that this does not matter for the purposes of individuating computation. The wide-functional strategy maintains that the MV can account for which logical/mathematical operation a

[14]It could be argued that the SV does not have the resources to deal with multiplicity because it does not tell us how contents are determined (Miłkowski, 2017, pp. 4-5). Therefore, the worry goes, the SV cannot tell us whether the device performs AND or OR because it cannot tell us which truth values are represented by what states. I think this is correct, but what it really highlights is that the SV is an incomplete theory of individuation. As noted above, the SV only commits one to the idea that semantic properties are required to individuate computation. A complete account of individuation would require an additional theory of content determination that explains how computational states come to have the contents that they do. This division of the problem space is useful in so far as two semanticists may disagree over how to think about content determination.

computing system performs after all. In what follows, I will examine these two strategies before suggesting that they are compatible and strengthen each other when combined.

The narrow-functional strategy sees the mechanist bite-the-bullet and accept that the functional transitions captured by the I/O table above are unable to determine which logical operation a logic gate performs. The mechanist can do this whilst maintaining that computation remains individuated by the non-semantic transformation of digits in the way outlined in the previous section. Dewhurst (2018) offers a version of this response (see also Piccinini, 2015, p. 128). Dewhurst does agree with Sprevak (2010) that an answer to the question of which logical operation is being performed requires semantic individuation but holds that this semantic individuation is not equivalent to computational individuation. The key move here is to claim that any individuation required of a computing system, on top of the sort captured in **table 1**, is not strictly computational individuation. Accordingly, the mechanist denies the premise implied by the objection: that an account of computational identity must settle which logical operation (if any) a computing device performs. As Dewhurst would have it, the inability of the MV to determine which logical operation is being performed,

> […] does not mean that computational processes cannot be individuated without representation—rather, it means that computational processes must be individuated in a way that remains neutral with regard to what logical function they carry out. (2018, p. 107)

The essence of Dewhurst's positive view is that all one needs to individuate computation, in a way that matters *qua* computation, is the 'physical description' of the system.

To help appreciate the force of the narrow-functional strategy, take the following table

for a second logic gate (**table 2**):

Table 2. AND or OR gate?

Input a	Input b	Output
5V	5V	5V
0V	5V	5V
5V	0V	5V
0V	0V	0V

Mirroring the case above (**table 1**), intuitions tend to suggest that this gate implements the OR operation because of our tendency to associate higher numbers together and lower numbers together (Dewhurst, 2018). But again, converse mappings of either [0V = 1; 5V = 0] or [0V = 0; 5V = 1] are perfectly coherent. This means that **table 2** can be interpreted as implementing the AND operation or the OR operation. Nonetheless, the MV can taxonomize the logic gate of **table 1** as different from that of **table 2** by reference to their functional properties alone. Therefore, the MV can distinguish between the two gates—albeit not by appealing to semantic content. Rather, computational individuation can be achieved 'with simply a physical description of how the various components of the mechanism function' (*ibid.*, p. 108).

As an aside, though Dewhurst (2018) refers to the 'physical description' of a computing mechanism as sufficient for individuation, I take the term 'functional description' to be preferable. Following Coelho Mollo (2018), it is important to stress that what matters for individuating computation is not any old physical properties of a system but those relevant to the number of digits the mechanism is sensitive to, together with the

transformations it performs on those digits. Hereafter, I will refer to these properties of a computing device as its 'narrow-functional properties'. The important point, for present purposes, is that the microphysical states of a computing device are grouped, in virtue of a set of their physical properties, into digits based on the sensitivities of the mechanism. In the case of the logic gate, electrical inputs are grouped by their voltage level because the system's nomic transformations are sensitive to electric potential difference; the difference that makes a (functional) difference. What matters as far as computation is concerned is not the different voltage levels of a logic gate *per se*, rather the fact that these provide a physical medium that instantiates the digits of an input-output mapping. In any case, a Dewhurst-style strategy avoids the problem of multiplicity by adjusting the level at which computational identity is determined. A logic gate does not perform multiple computations—*qua* narrow-functional individuation—it performs one computation, as defined by the number of digits and the transformations performed on them. A mechanist adopting the narrow-functional strategy can thus live with a computational description underdetermining which logical operation is being performed.

The narrow-functional strategy provides a strong rebuttal to the idea that the logic gate example demonstrates the necessity of semantic properties for individuating computation. A description of a mechanism's narrow-functional properties plausibly supplies a useful, well-specified, and non-semantic individuation of computation. However, two worries remain. First, one may feel that the narrow-functional strategy commits the mechanist to a troubling error theory. It is common practice for computer scientists to describe computations based on the logical operations they perform. One might feel uncomfortable with the resulting disconnect between individuating computation and logical descriptions, especially if one wishes their account of computation to do justice to the intuitions of the

scientific community that appears to individuate computation in line with the logical operations computing systems supposedly perform (Sprevak, 2010).

Second, Dewhurst concedes that semantic properties are required to pick out which logical operation a logic gate performs (maintaining that this does not affect its computational individuation). As such, the representationalist about cognition who draws on computation's supposedly semantic nature for support might grant that the narrow-functional strategy is correct so long as computation still invites some form of semantic individuation: representation does not individuate computation *per se,* the thought goes, but computational explanations of cognition still imply representation given that computing mechanisms perform mathematical/logical operations and representation is required to make sense of what mathematical/logical operation is performed.[15]

The wide-functional strategy is sensitive to both these worries. This second strategy in response to the semanticist's counterattack indicates that the MV does possess the resources required to individuate computation in line with paradigmatic logical/mathematical operations, but in a way that does not require representation. This is achieved by appealing to the wider functional context of a computing mechanism (the notion of a mechanism's 'wide function' resembles that of 'contextual function' introduced in **chapter 2**, and the overlap will be touched on again below). To appreciate the wide-functional strategy, it will prove useful to visit Shagrir's (2001) version of the claim that the paradigmatic tasks involved in computation require semantic individuation.

Shagrir (2001) shares similar concerns with Sprevak (2010), maintaining that semantic

[15] I thank Joe Dewhurst for discussing this point with me (personal communication).

individuation is necessary for capturing the explanatory power of computation. In fact, unlike Sprevak, Shagrir allows that computing systems implement multiple (non-semantically individuated) computations, provided under what he calls a 'syntactic interpretation' (2011, p. 374); this resembles Dewhurst's narrow-functional description. However, Shagrir says that there is only ever one computational description that explains a computing system's performance on some task—what he dubs 'the structure underlying the task in question' (*ibid.*, p. 375). Tasks are individuated semantically, Shagrir suggests. Therefore, in so far as computation explains a system's performance on some task, it must be individuated semantically (see also, Piccinini, 2015 p. 40). As Piccinini's (2015) interpretation of Shagrir's argument is more straightforward than the original presentation I will follow his reformulation of the problem.

Imagine a device that takes two input digits and produces one output digit. The device is sensitive to three digit types. Voltage levels physically realise these digits, corresponding to 0V, 1V and 2V. We can label these digits respectively with the following values: '0', '1/2', and '1'. The I/O table for the device then reads as follows (**table 3**):

Table 3. Averaging, AND or OR?

Input a	Input b	Output
0	0	0
0	½	½
½	0	½
0	1	½
1	0	½
½	½	½
½	1	½
1	½	½
1	1	1

On the face of it, this device performs a kind of averaging operation (albeit some of the averages are rounded up whilst others are rounded down). We might naturally imagine the designer of the device labelling the digits in this way for the very reason that it is intended to be used as an averaging device. Under this description (or 'labelling scheme'), the device takes two values as input and produces their average as output. However, other descriptions are possible (Piccinini, 2015, p. 41). For example, we can group '0' and '1/2' together and re-label them '0', or we can group '1/2' and '1' together and re-label them '1'. In choosing the former description, we have shown how the device could be used to implement the AND operation. In choosing the latter description, we have shown how the device could be used to implement the OR operation. These labelling schemes are contrived, but they are coherent (one really could use the device to perform the AND or OR operations). Nonetheless, Shagrir submits that in any given context, there is a fact of the matter as to which description captures the role of the computing device. For instance,

observing how the designer of our device uses her machine to calculate arithmetic problems, we see how describing the device as computing averages captures its role in completing the task at hand. Simply examining the number of digits and how they are manipulated would not tell us this. Shagrir rightly demonstrates that relations captured in the table above are insufficient for determining which logical/mathematical function (of the many that the device satisfies), is the appropriate description relative to its contribution to some task. We must appeal to some additional individuation criteria to make sense of this. Shagrir thinks this must be semantic individuation.

The crux of the problem is that the narrow-functional properties captured in **table 3** cannot privilege a description that captures the role of the system in completing some task because such a description captures the system *in context*. However, following Piccinini, I propose that semantic individuation is not required to solve this problem. Piccinini agrees with Shagrir that narrow-functional properties are 'insufficient to determine which task a mechanism is performing within a context, and thus which computation is explanatory in that context' (2015, p. 43.). However, whilst he accepts that multiple computations are being performed, Piccinini argues that non-semantic but 'wide-functional' factors constrain which of the many possible computational descriptions are explanatory. The key observation here is that the context of a mechanism's containing system, which itself can be understood as a higher-level mechanism, introduces additional constraints into our description of the computation. This context includes the broader causal nexus that exists between the computing mechanism and its containing system. For example, 'By looking at whether the containing mechanism responds differentially to a '0', '½' or '1' or responds identically to two of them, we can determine which computational description is explanatory without needing to invoke any semantic

properties of the computations' (*ibid.*, pp. 43-44). Therefore, we can determine whether the device is serving, say, an averaging role, by looking at whether the wider system is sensitive to all three digits, without needing to think that these digits represent numbers.

The semanticist may question just how far these wide-functional considerations get us: wide-functional properties supervene on the causal sensitivities of the containing system, but these sensitivities are not always sufficient to determine which task a system performs in a given context. In other words, wide-functional properties are not wide enough. For instance, imagine an explanation that posits a cognitive subsystem that computes distances between the containing system's hand and distal objects. One might question what causal sensitivities of the containing system could privilege this description. In turn, one might think the subsystem must represent features of the task.[16] I will offer three observations in response. These observations collectively act to further undermine the necessity of semantic properties for bridging computation and task performance.

First, wide-functionalism strongly indicates that semantic properties are not *always* necessary to select the task that a computing system performs in a given context because the causal sensitivities of the wider mechanism will sometimes suffice. Secondly, if semantic properties are only *sometimes* required for individuation, then it does not follow that the orthodox SV is true because the orthodox SV claims that semantic properties are *always* necessary for individuation. Admittedly, if semantic properties are sometimes required to individuate computation then this would also undermine the orthodox MV

[16] I share some of these concerns in Lee (forthcoming a). There I move away from conventional mechanistic views and outline a kind of pluralism about individuation. This pluralism permits semantic properties to play a limited role in individuating some cases of computation. However, I also remain neutral about the nature of these semantic properties, allowing much weaker notions than full-blooded representation to satisfy a semantic individuation of computation (see **section 3.3** below). The resulting position is remote from the orthodox SV.

because the orthodox MV claims that non-semantic properties are always sufficient for individuation. However, such a result would only imply a form of pluralism (not the SV). According to the pluralist view, semantic properties are sometimes required for individuating computation, and sometimes not (see Lee, forthcoming a). Thirdly, it remains unclear whether semantic properties are sometimes required for individuation—a clear counterexample is still owed. For instance, a wide-functional story about the above problem case could go something like this: to understand the sub-system as computing the distance between the hand and the distal object we need only observe the ways in which the subsystem is affected by the containing system's causal sensitivities (the containing system's processing and motor outputs). A cognitive system will be sensitive to stimuli corresponding to its hand and the distal object producing input for the computing subsystem, which subsequently (because of its output) generates motor behaviour consonant with judging the distance between two objects; for example, the system reaches for the object with its hand.

Dewhurst's 'narrow-functionalism' and Piccinini's 'wide-functionalism' have been recently framed as competing positions (Coelho Mollo, 2018). For the purposes of evaluating the SV, this hardly matters so long as semantic properties are not required. However, I think these views can be formulated in a mutually complementary way that is worth briefly exploring for clarificatory purposes. Coelho Mollo points out that Piccinini accepts 'multiplicity'. Piccinini allows that multiple computations are performed by a single mechanism; however, he holds that which explanatorily relevant computation is being performed is constrained by a wide understanding of functions (Coelho Mollo, 2018, p. 3492). For some, this is an unsatisfactory concession. I take the motivation for this dissatisfaction to be that multiplicity is counterintuitive, and therefore

accepting its existence weakens one's position.[17] It follows that the MV would be strengthened if it could avoid entailing multiplicity. Coelho Mollo thinks Dewhurst's account is stronger than Piccinini's because it rejects the premise that the MV leads to a multiplicity of computations. It does this by providing an account of individuation which references only the narrow-functional properties of the mechanism.

I propose that we combine the narrow-functional and wide-functional strategies into a strengthened mechanistic view of computation. The narrow-functional strategy guards against multiplicity—which otherwise besets wide-functionalism—whilst the wide-functional strategy guards against an error theory—which otherwise besets narrow-functionalism. The crux of this revised mechanistic view is as follows: computation is individuated *qua* computation at the level of narrow-functional properties and so multiplicity is avoided (only one computation is performed given the narrow-functional properties of the device). However, in addition to this individuation *qua* computation, computing mechanisms are also individuated by the logical/mathematical operations they perform (for example, the addition operation). There are thus two mutually compatible 'mechanistic levels' which inform how computing mechanisms are described: the narrow functional and the wide functional. One level concerns computation *per se*, the other level concerns computation in context. Representation is not required at either of these levels.

The distinction and compatibility between the narrow-functional and wide-functional properties of a computing mechanism—suggested by the revised MV—has a precedent in the mechanism literature in so far as they overlap with Craver's (2013) notions of

[17] I question whether multiplicity is really a problem for wide-functionalism if it can sufficiently constrain which of the multiple computations being performed is explanatory given the broader relations between a computing device and its containing mechanism. For now, let's grant that an account of computational individuation is strengthened if multiplicity is avoided.

'constitutive explanation/function' and 'contextual explanation/function' respectively (see **chapter 2**).[18] Recall that constitutive explanations concern whether and how the properties of a mechanism produce a phenomenon and focus on its synchronic causal structure (Craver, 2013, pp. 149-151), whereas contextual explanations concern the wider environmental context of a mechanism (Craver 2013, pp. 151-154). Two mechanisms may be functionally equivalent in so far as they perform the same constitutive function, but functionally dissimilar in so far as they perform different contextual functions. This means that two physically identical pumping devices could be used to pump different substances, for example, blood or air. Whether one wants to say these two mechanisms are functionally equivalent or different, therefore, depends on whether one is referring to narrow or wide functions. Likewise, whether two computing devices are functionally equivalent or different will be informed by whether one is taking into consideration the wider context of the computing devices.

3.3 Does the semantic view really imply representation?

We have so far been assuming that the semantic properties invoked by the SV imply representational properties, and hence that the SV implies representation. This allows the representationalist to support the explanatory value of cognitive representation by appealing to the idea that our best explanations of cognition are computational. And yet, one can question whether the SV really supports representationalism in this way. The problem is that the 'semantic properties' invoked by the SV are ambiguous and may imply something weaker than representation. This objection is not concerned with whether the

[18]One could possibly defend an even more liberal view that encompasses an additional 'semantic level' on top of the 'narrow-functional' and 'wide-functional' levels. Depending on how the relevent semantic properties are understood, this level might correspond to Craver's third kind of explanation/function: 'etiological explanation/function' (see Lee, forthcoming a, for related discussion).

SV is true, but instead questions whether the semantic properties invoked by the SV serve as a basis for full-blown cognitive representation.

Recall that to pass the job description challenge (JDC) an entity must function in a way that is distinctly representation-like, namely, as a stand-in for something on behalf of a cognitive system. To do this, I suggested that an entity must possess representational content, (a) in a sense that implies correctness conditions, and (b) in such a way that content is causally relevant to the cognitive system itself. However, it is unclear whether the semantically individuated states required of the SV necessarily possess content in this way. To begin, take Sprevak's understanding of the sort of representation required for the SV to be true,

> Roughly speaking, a *representation* need support no more than a basic notion of aboutness or reference. A representation should link an entity and a content, such that the entity represents its content. Nothing more is required. (2010, p. 261. Original emphasis.)

As it stands, this relatively non-committal characterisation appears to tolerate minimal notions of representation, such as receptor representation—the sort we might concede as being semantic in a weak sense but rejected in the previous chapter as unfit to support a distinctly representational theory of cognition. There are two problems mirroring the conditions of the JDC just mentioned: (a) it is unclear whether the semantically individuated states of the SV play a causal-role *in virtue of* the semantic properties they possess, and (b) it is unclear whether the semantic properties possessed by these states imply representational content. Let's examine these problems further.

The first problem is straightforward. We saw in **section 2.3** above that at least some computing systems appear to be sensitive only to the formal properties of their digits. This fuelled our suspicion that semantic properties are not required for individuation. However, one might insist that semantic properties are jointly necessary for computation, even if they are, at least sometimes, causally irrelevant. Sprevak (2010) makes this point explicit in his defence of the SV, claiming that the causal irrelevance of semantic properties does not entail their irrelevance for individuation. For Sprevak,

> [The semanticist] does not claim that the causal dynamics of computations depend on representational content. Her claim is that the individuation of computations depends on their representational content. The battleground for the received view are the facts about the individuation of computations, not the facts about their dynamics. (*ibid*., p. 261. My parenthesis.)

In other words, something does not need to be causally relevant to be necessary for individuation. One might question whether something causally irrelevant to a physical system could really determine the identity of processes in that system. Regardless, Sprevak demonstrates that the semanticist's priority is not to establish the causal relevance of semantic properties for a computing system. This differentiates the so-called representations required by the SV from the representations sought by the JDC.

The second problem concerns whether the semantic properties invoked by the SV are necessarily representational in nature, the suspicion being that weaker semantic relations suffice. Take the example of visual processing, typically taken to involve computation. Even if we grant some privileged semantic relation between state *x* in the visual cortex and some distal object *y*, where computations are performed over state *x* such that *x* 'is about' *y* in some sense, it does not follow that *x* stands-in for *y* in the robust sense required for cognitive representation. For instance, if the semantic relation consists solely of nomic

causal dependency, then *x* functions as a receptor and only carries natural information. Alternatively, if the relation consists of a conventional mapping fixed by an observer's interpretation, then the representing relation is non-naturalistic, and *x* is only a representation to the extent that an observer treats it as such. If either of these relations is sufficient for grounding the semantic properties of states under the SV, then those states do not function as representations on behalf of a cognitive system.

To see the point being made more clearly, recall the relations implicit in the causal-informational and interpretivist theories. First, consider causal-informational theories. The previous chapter distinguished between two notions of information, and by extension, content—a distinction that maps on to Grice's notions of natural and non-natural meaning. According to this taxonomy, representational content is equivalent to non-natural information/meaning and implies the existence of correctness conditions. I argued that, at best, receptor representation implies a weaker notion of content (equivalent to natural information) than representational content (equivalent to non-natural information). If this weaker notion of content is sufficient for the SV, then the SV does not necessitate representation. Orlandi makes a similar point:

> The digits in a computer may be states that co-vary with environmental elements—thereby carrying information about them—and they may enter in encapsulated transitions without being representations. (2014, p. 207)

They go on to add that,

> [I]n some cases, computations may be usefully individuated in terms of informational content while also not requiring appeal to representations […] It seems to follow from this that an information-processing system is not, *ipso facto*, a representation-processing system. (*ibid.*, p. 207)

In short, if computing systems and their states bear semantic properties only in virtue of

nomic causal relationships to distal objects, then those states possess semantic properties, but those states do not possess representational content.[19]

Second, consider interpretivist theories. In a sense, interpretivism side-steps the worry about semantic states not necessarily constituting representations.[20] The JDC raises the bar for the justifiable attribution of natural representation, but the interpretivist denies that computation involves natural representation. For the semanticist who embraces interpretivism, the representation that individuates computation is conventional representation (attributed by an observer).[21] However, interpretivism brings its own baggage. If one defends cognitive representation on the grounds of the SV, and one is an interpretivist about content even for cognitive systems, then one imports that observer dependency about content into one's conception of cognitive representation. This does not deliver the naturalistic notion of representation that representationalists typically subscribe to. At the very least, it undermines the idea that cognitive representation is as observer independent as other theoretical posits in cognitive science. This is in part because it reinforces the suspicion raised by the first problem—that the semantic properties invoked by the SV are causally irrelevant and, therefore, not the sort of properties that support the explanatory power of cognitive representation.

In summary, it does not follow from the fact that our best theory depicts the brain as a computing system, and the fact that the SV is true, that cognitive representation plays a robust role in explanations of cognition. If this is correct, then there are far-reaching

[19] A similar point applies to all naturalistic theories of content. For instance, even if FRS rightly identifies computational states as having a sort of content in virtue of their relationship with one another, this is insufficient for demonstrating that computational states function as representations.

[20] My thanks go to Mark Sprevak for drawing my attention to this point (personal communication).

[21] Egan (2010) offers an interesting position that arguably eschews the SV but maintains interpretivism about content, where content is assigned a purely heuristic role within a computational theory.

consequences for anti-representationalism. To close this section, I will touch upon one of these consequences.

Recent efforts have been made by sympathisers of 'autopoietic enactivism' (or simply 'enactivism') to free computation from the grips of representation (Villalobos & Dewhurst, 2017; Villalobos & Dewhurst 2018).[22] Autopoietic enactivism stands in opposition to representationalism. This is because autopoietic enactivism takes representationalism to imply an erroneous relationship between organism and world. Where an organism represents the world, the thought goes, an organism infers or re-constructs facts about the world. There are different ways of unpacking the enactivist's antagonism towards this idea, but one view has it that organisms and their worlds are better explained as 'co-arising' or 'co-constructed' whereas representation implies a fundamental separation between organism and world (Varela, Thompson & Rosch, 2016). For an organism to represent, the world must possess a set of pre-given properties to be represented. An organism that makes inferences over representations implies an organism that can get things 'right or wrong', measured against an objective fact about the world. In other words, representation implies correctness conditions.

That computation implies representation is typically taken for granted by enactivists (Varela, Thompson & Rosch, 2016; Thompson, 2010). For example, Thompson writes, 'A computer is supposed to be a symbol-manipulating machine. A symbol is an item that has a physical shape or form, and that stands for or represents something' (2010, p. 4).

[22] 'Enactivism' is ambiguous and often refers to one of two related but distinct traditions: autopoetic enactivism (discussed here) or 'sensorimotor enactivism'. See Ward, Silverman & Villalobos (2017) for recent discussion on the relationship between these traditions. 'Radical enactivism' (Hutto & Myin, 2013, 2018) is a third and more recent form of enactivism that draws on these earlier traditions (see **chapter 1**).

As such, received autopoietic wisdom holds that computational explanations of cognition must be rejected. However, Villalobos & Dewhurst argue that the SV is false, therefore, enactivism can draw on computational explanations whilst evading the toxic notion of representation (Villalobos & Dewhurst subscribe to the MV).

As we have seen, the SV is not the only game in town. If the MV is true, then Villalobos & Dewhurst are right to conclude that computational approaches are compatible with enactivism; or at least, they are not incompatible because of representation. I wish to emphasise a different point. If what I have said in this section is right, then the SV does not necessarily imply full-blown representation anyway. Thus, enactivists and other like-minded anti-representationalists need not reject computational explanations of cognition on the grounds that the SV is true. On the flipside, semanticists need not reject anti-representationalist approaches solely because our best theories are computational. As a general lesson, before we affirm or reject an approach which invokes 'representation', we ought to be clear on the strength of the notion implied.

4.1 The representational theory of mind

Preceding sections have undermined the idea that physical computation implies cognitive representation. In short, the semantic view is probably not correct, but even if it is, full-blown representation does not follow. We turn now to the 'representational theory of mind' (RTM). According to one interpretation of this theory, computational approaches to cognition imply representation because of the sorts of explanations they offer within the context of cognition, not because physical computation *simpliciter* presupposes representation. More specifically, the RTM claims that computational approaches to cognition imply representation because they depict the brain as instantiating a computing

system that vindicates the representational posits of folk psychology. The computational explanations that proponents of the RTM have in mind are part of the 'classical computational theory of cognition' (CCTC).

The CCTC explains some or all of cognition in terms of the serial processing (the execution of one task at a time within a well-specified time interval) of discrete symbols (inner 'syntactic states'), in accordance with formally specifiable rules. At heart, this view depicts cognition as caused by the deterministic or stochastic storing, combining, re-combining and erasure of discrete, atomic symbol tokens which join to make complex symbol tokens. Such symbol manipulation is typically seen as conducted in accordance with language-like rules within an algorithmic structure. This means that a given cognitive capacity is explained by breaking down a cognitive system into parts that perform increasingly simple symbol-crunching operations. At each level, the manipulation must conform to an appropriate 'input-output conversion', relative to the task the computation is supposed to explain (Ramsey, 2007, p. 42). This conformity can be understood in terms of an appropriate mirroring of the input-output transformations and processes between the task domain and the computation. [23]

So far, our gloss of the CCTC is consonant with the mechanistic view: nothing within the above characterisation entails representation. Yet according to the representational theory of mind, the computational architecture posited by the CCTC provides a way to think of

[23]The CCTC is characterised by a type of computational architecture and is not to be confused with the broader theory that cognition is computational in some sense. The broader theory is consistent with other frameworks—in particular, connectionism. In contrast to the CCTC, connectionism models cognition via the emergent behaviours resulting from the parallel processing of interconnected networks of single processing units. The differences and similarities between the CCTC and connectionism have been well documented (Franklin, 1995, provides a useful overview). Fortunately, we are not required to wade into the murky waters surrounding the debate between proponents of the frameworks concerning their relative compatibility, and their comparative explanatory value.

mental processes as causal processes that involve transitions between mental representations. The key move is to draw a parallel between the CCTC and folk psychology.

Folk psychology concerns the ordinary ways in which agents predict and explain behaviour. It is often assumed that this practice depends on the attribution of 'propositional attitudes', like beliefs and desires, as evidenced through daily mentalistic discourse—such as when I say, 'I believe blueberries are good for you' upon being asked why I have eaten a whole punnet. Indeed, the term 'folk psychology' is sometimes used synonymously with 'propositional attitude psychology'. For example, Hutto writes that folk psychology is 'stricto sensu […] the practice of making sense of a person's actions using belief/desire propositional attitude psychology' (2008, p. 3).[24]

Propositional attitudes are the paradigmatic representations familiar to the 'manifest image' of daily life (Stich, 1983, p. 5). Beliefs and desires are usually regarded as the most prominent, but they also include wishes, hopes, fears and so on. Propositional attitudes have three degrees of freedom, consisting of an *agent* who bears a *relation* to a *proposition.* The relation is the attitude (for example, belief, or desire). The proposition refers to a state of affairs specifiable by a 'that clause', for example, 'that blueberries are good for you'. In so far as propositional attitudes consist of an agent's relation toward a proposition, they stand-in for states of affairs in a way that implies correctness conditions:

[24] According to a growing body of research, primarily on animal and infant social cognition, the role of propositional attitudes in our daily prediction of behaviour is more restricted than has been traditionally assumed by philosophers. Many now claim that propositional attitudes play a predictive role only within a limited subset of relatively sophisticated 'mindreading', and/or propositional attitudes primarily play a role in justifying behaviour rather than predicting it. For some influential discussion, see Gergely, (2002); Gergely & Csibra, G (2003); Apperly & Butterfill (2009); Bermúdez, (2009); Sodian, (2011); Zawidzki, (2011). For some recent pushback, see Borg, (2018). I put this research aside due to limited space and because I wish to be maximally concessional to the RTM.

beliefs are true or false; desires are satisfied or unsatisfied (Fodor, 1987, p. 8). The belief that blueberries are good for you is either true or false. The desire for blueberries is either satisfied or unsatisfied.

The RTM offers a view on the relationship between folk psychology and cognitive science by drawing an intimate connection between the CCTC and folk psychology. According to the RTM, the causal relations between inner symbols in the CCTC mirror those causal relations between representations suggested by folk psychology. The symbols of the CCTC are syntactic states, in so far as their causal properties are purely formal, but they bear contents that correspond to the representations of daily mentalistic discourse. For proponents of the RTM, the explanation for why folk psychology is effective is that it directly tracks the real inner causes of behaviour.

The key move made by proponents of the RTM is to claim both that (a) the CCTC is our best scientific theory of cognition, and (b) the CCTC offers a scientific vindication of folk explanations. There is thus a convergence between the *prima facie* plausibility of folk explanations, supported by their apparent predictive and explanatory power, and our best scientific explanation of cognition. Most crucially, the propositional attitudes of folk psychology are to be understood as relations to a complex of inner syntactic states of the sort posited by the CCTC. According to the RTM, if one has a thought then one is standing in relation to a sentence, composed of atomic mental symbols in 'mentalese', or the 'language of thought' (Fodor, 1975). For example, if I think the proposition that Khan hates Kirk, I am standing in relation to the mentalese sentence composed of corresponding symbols organised in a corresponding structure {KHAN, HATES, KIRK}.

In constructing this parallel between folk psychology and computation, the RTM proponent claims that a type of attitude is individuated by its causal role within the matrix of rule-governed principles made possible by the computational architecture hypothesised by the CCTC. A belief is a belief, as opposed to a desire, because of the sorts of causal relations it enters into. Take a folk psychological explanation of behaviour, like going to the fridge to retrieve blueberries. This behaviour is to be explained in terms of the causal relations between attitudes and their contents, such as the desire that I want blueberries (P), the belief that there are blueberries in the fridge (Y), motoric outputs of going to the fridge (Q), and the generalisation that one tends to pursue one's desires in accordance with one's beliefs. If we understand the mind in terms of computation, and computation in terms of symbols processed in accordance with formally specifiable rules, then we discover a means by which the theoretical posits of folk explanation can be instantiated. What is required is the physical realisation of symbols that correspond to P, Y, the rule P $\wedge$ Y $\supset$ Q, and the capacity for generating Q.

In addition to the above, the RTM depicts the CCTC as naturalising supposedly elementary features of thought, such as they are implicit in folk psychology. Most notably, the CCTC explains the apparent properties of 'productivity'—the fact a thinker can generate novel and potentially unbounded numbers of thoughts—and 'systematicity'—the fact that mental states seem to enjoy systematic relations that partly determine the capacity for what other thoughts one can have. For example, if someone can entertain the thought that Khan hates Kirk, then they can also entertain the thought that Kirk hates Khan. These features can be understood by reference to an underlying 'representational system' that allows for the combining of atomic representations into compound representations via a combinatorial syntax, with no formal boundary on the

length of compound representations (for foundational discussion, see Fodor & Pylyshyn, 1988; for an argument that thought is not productive in the way RTM depicts, see Johnson, 2004).

According to the RTM, the CCTC uncovers a supremely explanatory parallel between folk and scientific explanation. This gives rise to a conception of the CCTC as mechanizing rationality, providing a bridge between the 'normative' (folk psychological) and the 'physical' (neurophysiological) (for critical discussion, see Dupuy, 2009). As Fodor puts it, computers 'show us how to connect semantically with causal properties for symbols' (1987, p. 20). If this picture is correct, the semantic relations amongst the contents of thought can be successfully captured via the mind as a 'syntax-driven machine' processing symbols with representational content (*ibid.*, p. 20). As Shores clarifies, Fodor's 'language of thought' is committed to a 'nomological' correspondence between the contents of folk psychological states and the contents of internal mental representations which agents bear relations to. This is because 'the same explanatory role is assigned to the belief as to the relation' (1985, p. 62). So, if an agent has the belief that blueberries are good for you, they are standing in a law-like relation to a representation with the content that blueberries are good for you. For the proponent of RTM, the CCTC thus provides a plausible explanation of how a physical system like the brain can capture the normative relations associated with the contents of thought through the processing of internal representations.

4.2 'Representation' in the representational theory of mind

There are two sorts of objections to the RTM: empirical objections and conceptual objections. Empirical objections attack the RTM's scientific basis. For instance, one

could deny that the CCTC offers the best theory of cognition, or alternatively, claim that the best version of the CCTC does not posit inner states that systematically correspond to folk psychology. One version of this objection is pursued by the eliminative materialist, for whom the correct account of cognition 'will bear about as much resemblance to FP [folk psychology] as modern chemistry bears to four-spirit alchemy' (Churchland, 1981, p. 82. Parenthesis added). Conceptual objections attack the way the RTM proponent infers representation from the CCTC. I think empirical objections to the RTM are powerful. However, I am presently interested in the extent to which the CCTC really implies representation, not whether or what version of the CCTC is correct. Therefore, the remainder of this section will focus on conceptual objections to the RTM.

According to the RTM, the CCTC implies computational states that correspond to the propositional attitudes of folk psychology. There are actually two possible ways of interpreting this claim. In evaluating the RTM, it is important that these interpretations are separated. The first interpretation says that the symbols of the CCTC are representations because any symbol within computation is a representation. This interpretation pertains to generic considerations about the nature of computation. The second interpretation says that the symbols of the CCTC are representations because they are analogues to propositional attitudes. This interpretation pertains to special considerations introduced by the naturalisation of folk psychology. I will briefly consider the first interpretation before focussing my discussion on the second interpretation.

The first interpretation essentially constitutes a commitment to the SV: computational approaches to cognition imply representation because all computation implies representation. If this is the appropriate interpretation of the RTM, then the Fodorian

slogan ‘no computation without representation’ is really a thesis about the nature of physical computation *simpliciter*. This interpretation is evident in parts of Fodor (1975) and is usefully summarised by Shores (1985). Shores writes that the notion of computation adopted by the RTM is one where computation ‘is a process that involves the manipulation of interpreted syntactic objects’ (*ibid.*, p. 13). He goes on to write that, ‘Interpreted syntactic objects are syntactic objects that have semantic properties such as meaning, truth, or reference’ (*ibid.*, pp. 13-14).[25] I have argued against this view of computation in preceding sections. It is enough to say that the RTM does not introduce any novel argument for the SV. Drawing from the lessons above, note that the SV does not follow merely from the fact that computations serve in cognitive tasks or capacities (see **section 3.2** above; we will return to this argument in **section 4.4** below). Moreover, there is nothing special about the RTM that would ease the concern, raised in **section 2.4** above, that a weaker notion of semantics suffices for the truth of the SV.

The RTM proponent may concede that physical computation itself does not presuppose representation but maintain that there is a stronger reason for thinking that the CCTC implies representation.[26] As sketched already in **section 4.1** above, the appeal to the necessity of representation in the CCTC then comes to rest on an observation about the special relationship between symbols in the CCTC and the representations of folk psychology. If our best theory explains the predictive power of propositional attitudes by conceiving of them as relations to the inner symbols of the CCTC, then the inner symbols

[25] Observe that if the SV is true, an appropriate version of the CCTC is true, and the SV does imply genuine representation (*pace* the worries raised in **section 2.4** above), then this is sufficient for justifying the explanatory significance of cognitive representation. In other words, cognition would be representational even if the CCTC failed to vindicate folk psychology.

[26] Shores (1985, chapter 4) acknowledges something like this in response to Stich’s (1983) ‘syntactic argument’ against the necessity of representation in the CCTC. See below, this section, for related discussion.

of the CCTC ought to be conceived of as a kind of cognitive representation.

The problem with this argument is that all the CCTC demonstrates (if correct) is that folk psychology is successful because there exists a level at which inner syntactic processes mirror those of folk psychology. The RTM assumes that if this parallel obtains, then computational processes inherit the representational properties of states in folk psychology. According to this logic, to show that propositional attitudes map onto the causally potent states of CCTC is to show that the CCTC must represent the contents of those propositions. To borrow from Cummins, for the RTM, 'thinking *that p* requires representing the proposition *that p*' (1996, p. 3. Original emphasis). The CCTC is committed to causally potent symbols whose processing is sensitive only to the 'syntactic properties' of those vehicles, and in turn, let's grant, those processes mirror the processes of folk psychology. But this mirroring relation does not entail that the symbols of CCTC retain the semantic properties of propositional attitudes. It does not follow from the fact that the success of folk psychology needs explaining, and that the CCTC explains this success by positing a level of corresponding syntactic states, that the properties of CCTC symbols inherit all the properties of states as they appear in folk psychology. In general, *explanans* need not feature identical properties of their *explananda*.

Dupuy (2009) articulates a version of this worry, distinguishing between the 'preservation' and 'creation' of meaning. Dupuy says that whilst 'syntactic rules' preserve the internal coherence of symbolic representations, this is not enough to show that they bear representational content. It is for this reason that 'one of the stumbling blocks the computer model faces is the problem of determining how symbols acquire meaning.' (*ibid.*, p. 39). I think the central insight implicit within Dupuy's observation is

that the traditional problem of content determination can be traced back to the opaque role of representational content in the CCTC, where attributions of said content are motivated by the fact that the syntactic relations of the CCTC are capable of mirroring the semantic relations of folk psychology. I will argue in **chapter 4** and **chapter 5** that representational content becomes less mysterious when we ground its attribution in the causal role played by a set of representation-like mechanisms.

The present objection can be traced back to Stich (1983), whose central idea can be summarised as follows: by itself, the CCTC only provides the grounds for positing processes over syntactically defined states, providing no additional reason to think that these states are properly representational. According to Stich's 'syntactic theory of mind', offering a computational description of the mind commits one to the idea that relations among neurological states map to syntactic relations among parts of 'mental sentences' that obtain at a higher level of description. Stich's point is that computational description occurs at this syntactic level and in the 'formal relations amongst the syntactic objects of mental sentences' (*ibid.*, p. 151).[27] It is these formal properties that are responsible for explaining the contribution of computational states to the behaviour of a system. According to Ramsey, following Stich, the 'standard interpretation' holds that the CCTC provides a reduction base for propositional attitudes; thus, the notion of representation ascribed to the CCTC is the very same notion of representation suggested by folk psychology. He writes,

[27] The MV has been framed as an evolution of Stich's syntactic view (Piccinini, 2015, pp. 46-47). The main difference is that the MV avoids committing to the characterisation of computation in terms of 'syntactic' relations. As far as the present discussion is concerned, the takeaway message of the syntactic and mechanistic view remains the same: for at least some cases of computation, the explanatory purchase earned by computational descriptions is carried by the transformation of functionally-defined digits, not by the transformation of representations.

> [T]he Standard Interpretation comes with the tacit assumption that we can show how symbols are representations by claiming that they realize or instantiate propositional attitudes [...] this doesn't work. You can't use the fact that A is the proposed reduction base for B to establish that A has all the relevant features of B. That is, you can't make computational symbols function as representational states by proposing that they serve as the things with which folk mental representations are identified. (2007, pp. 64-65)

Schweizer (2017) makes a similar point when he says,

> [T]he unique 'content' postulated by RTM is superfluous to the formal procedures of CTM [computational theory of mind]. And once these procedures are implemented in a physical mechanism, it is exclusively the *causal* properties of the physical mechanism that is responsible for all aspects of the system's behaviour. So once again, postulated content is rendered superfluous. (p.65. Parenthesis added. Original emphasis)

Each of these objections offers the same essential observation: even if the CCTC provides a reduction base for folk psychological attitudes, the role of representation for scientific explanations of cognition remains opaque.

4.3 Input-Output representation reconsidered

The criticisms of the RTM surveyed above accord with Ramsey's (2007) objections to supporting the explanatory significance of cognitive representation within the CCTC by appealing to its role in vindicating folk psychology. However, Ramsey insists that the CCTC does invoke representations of two sorts. These take the form of 'input/output representation' (IO-representation), and 'structural', 'simulation' or 'surrogate representation' (S-representation). I will defend a version of the latter in the next chapter. In closing this chapter, I suggest that the IO-representation notion provides less convincing support for the explanatory significance of cognitive representation.

For Ramsey (2007), following Cummins (1991) and others, cognitive science is generally in the business of offering explanations of how a system performs a task or achieves a capacity by demonstrating how 'inputs' are converted into 'outputs' via internal processes.[28] These inputs and outputs are not themselves features of the task or capacity to be explained. Rather, these inputs and outputs represent features of the task or capacity to be explained. In this way, cognitive science depicts cognitive processes as converting representational inputs into representational outputs. As Ramsey summarises,

> For now, the key point is that we are justified in treating a cognitive system's inputs and outputs as representations because, given what we know about cognitive systems, we are justified in characterizing many of their operations as having certain types of starts and finishes; namely, starts and finishes that stand for other things. (2007, p. 70)

Take the example of a cognitive system's capacity to recognise faces. The input that the cognitive system receives is not an actual face but 'some sort of visual or perhaps tactile representation presented by the sensory system' (*ibid.*, p. 69). The resulting output is also not an actual face but a representation: 'perhaps something like the recognition "That's so-and-so," or perhaps a representation of the person's name' (*ibid.*, p. 69). In short, Ramsey suggests that the states received and generated by a cognitive process can be thought to represent features of the cognitive capacity that they explain.

Not all cognitive theories offer explanations in terms of discrete starts and finishes. For example, the kinds of explanations offered by dynamical systems theory (Chemero, 2009)

[28]Such processes can be recontexualised in terms of the mechanistic framework of explanation. In so far as *explanans* possess an input-output structure, cognitive science posits what Glennan calls 'input-output mechanisms'. An input-output mechanism is 'a complex system that is situated in its environment in such a way that there are characteristic environmental events (inputs) that trigger a sequence—perhaps multi-stranded—of interactions between parts of the mechanism. This sequence concludes with some terminating event, the output' (2002, p. 347).

and some versions of enactivism (Hurley, 2002) eschew this neat input/output structure. Nonetheless, Ramsey's characterisation covers a common form of explanation including those offered by the CCTC. Before continuing, it is important to recognise the relative innocuousness of IO-representation as presented so far. Assume for a moment that Ramsey's analysis of IO-representation is correct. A theory that offers explanations in the form of I/O transformations is not necessarily a representational theory of cognition. To be a representational theory of cognition is to be a theory that posits internal representations as part of the cognitive apparatus that causes behaviour (for related discussion, see Ramsey, 2007, p. 71). As such, even global eliminativists may concede that IO-representations are a feature of cognitive science simply given the structure of *explanans* in cognitive science. This would not shake their faith that our best theories in cognitive science eschew all internal cognitive representation. And yet, Ramsey (2007) argues that the CCTC is indeed a representational theory of cognition. This is for two reasons: because the CCTC implies 'S-representations' (see next chapter), and because the CCTC implies *internal* IO-representations.

Under Ramsey's analysis, there are different possible 'layers' of IO-representation. As we have seen, the top layer characterises how explanations of cognition start by positing start-up and finishing conditions that reflect the capacity to be explained. In addition, Ramsey says that the CCTC depicts a series of sub-layers of I/O transformations as part of its explanatory apparatus. This is because the CCTC adopts a 'task-decompositional' approach, explaining the capacity of cognitive systems by positing a series of increasingly simpler computing sub-systems with their own I/O transformations. It follows that these computing sub-systems will themselves require IO-representations: 'Internal mini-computations demand their *own* inputs and outputs, and these representations that are

external to the mini-computation are, of course, *internal* to the overall system' (p. 72. Original emphasis). Hence, the CCTC implies internal IO-representations.

Ramsey is right that the sorts of sub-systems featured in the CCTC are treated as computing operations, like addition, in a way that accords with the system-level capacity—or higher-level mechanism capacity—that they explain. However, I am sceptical that this implies these states function as representations on behalf of the system.[29] The idea that a theory like the CCTC requires IO-representations depends on the assumption that if computational processes do not manipulate the features of the capacities that they explain—for example, if they don't interact directly with faces, but they do explain facial recognition—they must manipulate representations of those features—for example, they must manipulate representations of faces. However, following the conclusions reached above, I submit that representation is not required to capture the role that computational processes play in completing some task or capacity. What is required is some causal connection between the computing mechanism and features of the task being explained. A computing subsystem in the brain does not compute in a vacuum but is connected to myriad other perceptual channels and subsystems for motor output. A subsystem may be said to compute, say, the distance between the system's hand and some object to be reached (i.e., we can privilege *this* description of the many that are compatible with the narrow-functional properties of the subsystem), not because the subsystem's states represent hands and objects, but because of the way that subsystem is embedded in a system that is sensitive to hand and distal

[29] Ramsey does expresses doubt over whether the internal IO-representations of the CCTC '*actually are*' representations (2007, p. 76. Original emphasis). He (somewhat tentatively) concludes that IO-representations do function as representations within the CCTC in a robust sense: 'computational processes treat input and output symbolic structures a certain way, and that treatment amounts to a kind of job assignment – the job of standing for something else' (*ibid.*, pp. 76-77).

object stimuli, and which produces appropriate motor output. This corresponds to the kind of 'wide-functional' individuation sketched earlier.

Following Block (1990), Ramsey uses the example of explaining how an agent performs multiplication by positing an internal subsystem that repeatedly adds numbers. He writes,

> If there is an inner sub-system that is an adder, then its inputs must be representations of numbers and its outputs representations of sums. If these internal structures are not serving as representations in this way, then the sort of task-decompositional analysis provided by the CCTC doesn't work. (2007, p. 72)

However, it strikes me that we can understand a subsystem as performing addition in accordance with task-decompositional analysis without accepting that the digits of the underlying process represent features of addition, like numbers and sums, on behalf of the cognitive system. Adopting the revised mechanistic view, I propose that a task-decompositional analysis of a system is only committed to there being a parallel between the number of digit types and their transformations (its 'narrow-functional properties') and the addition operation—plus some fact about the context of the computation that privileges the addition operation over other descriptions (its 'wide-functional properties'). In general, a computing mechanism can perform an operation, like addition, relative to some capacity to be explained, without representing features of that operation. In the above example, the subsystem explains how an agent performs multiplication, so we can assume that the subsystem possesses causal connections with other subsystems within the agent that are sensitive to number-related stimuli and appropriate verbal and other motor outputs. These connections make sense of why the addition function is privileged. Another way of thinking about the same point is this: according to the MV, we should not begin by thinking of the subsystem as performing addition, which in turn

explains how the agent performs multiplication. Rather, we should begin by thinking of the subsystem as performing computation *qua* its narrow-functional (capacity neutral) digit transformations. When embedded in a wider mechanistic context, this process explains how an agent performs multiplication, at which point, the addition operation description captures the role that the subsystem plays for the agent (the same subsystem could perform a different operation relative to a different capacity).

In closing, note that a mechanistic interpretation of task-decompositional analysis does not preclude agents treating the states of a cognitive subsystem they are investigating as representations for their own heuristic purposes (for related discussion, see Egan, 1995). It may be that interpreting inner subsystems as transforming representations is useful for tracking the relevant transformations given their role in some cognitive capacity of a containing system. The present point is that a task-decompositional approach does not necessarily commit a theory to cognitive representation in any substantive sense. Recognising the potential heuristic value of representation-talk is different from claiming that a computational theory itself posits entities that function as stand-ins for a cognitive system (for related remarks, see Schweizer, 2017, pp. 74-75).

5. Conclusion

This chapter argued that there is no necessary connection between computation and representation. I claimed that the semantic view of computational individuation faces several potent objections. In turn, I showed that the alternative mechanistic view has the resources to individuate computation in accordance with its role in serving paradigmatic tasks without the need to appeal to semantic properties. In keeping with the mechanistic view, I contend that to compute and to represent are distinct functional kinds. However,

there is a good reason to suspect that even if the semantic view is correct, full-blown representation does not follow. Finally, I examined the relationship between the classical computational theory of cognition and the representational theory of mind. There are two principal reasons why a proponent of the representational theory of mind would think that the classical computational theory of cognition presupposes representation: because the semantic view is true, and because computation provides a naturalised reduction base for propositional attitudes. I have argued that both reasons are unconvincing.

If the above arguments are correct, then a system may perform computation without representation. Therefore, appealing to computational explanations of cognition alone will not serve to show that subpersonal cognitive representation plays a significant explanatory role in cognitive science. This carries positive consequences for explanatory acuity: computation and representation mark distinct functional kinds, and so determining whether cognition is representational as well as computational is a significant discovery. More generally, determining what systems compute, what systems represent, and the intersection of these sets identifies non-trivial facts about where and to what extent these different functional types overlap within mechanisms for different phenomena.

The following chapter defends a characterisation of cognitive representation in terms of internal structure-preserving map or model-like mechanisms. As it happens, these ‘S-representations’ are consonant with many computational explanations of cognition, including the classical computational theory of cognition. Though computation and representation are conceptually distinct, there are good reasons to suspect that our best explanations will depict cognition as both computational and representational.

Chapter 4

The S-Representation Account

1. Introduction

This chapter defends a version of the 'S-representation account'. I argue that when a cognitive mechanism meets the functional criteria for S-representation, that mechanism serves as a representation on behalf of a cognitive system, and therefore counts as a cognitive representation. In turn, when an empirical theory posits a mechanism that meets the functional criteria for S-representation, it counts as a representational theory.

As I intend it, the S-representation account depicts cognitive representation in terms of structure-preserving mechanisms that function analogously to a class of ordinary representation (the 'S' handily encompasses the terms 'simulation', 'surrogacy' and 'structural', all of which have been deployed in the literature to similar ends). This class of ordinary representation includes cartographic maps and at least some scientific models. Mechanisms that meet the criteria for S-representation determine, in part, the success or failure of a cognitive capacity in virtue of the extent to which they structurally correspond to some entity. In doing so, they imply correctness conditions that are causally relevant to a cognitive system. In addition to presenting the S-representation account, this chapter responds to some of the strongest objections in the literature. Along the way, I will suggest that far from being a relic of the classical computational theory of cognition, the S-representation account remains in touch with alternative computational approaches (cf. Ramsey, 2007). The S-representation account is an account of representation, not an empirical theory. However, frameworks like predictive processing, as well as specific,

cross-framework mechanisms such as ‘cognitive maps’ indicate the continued relevance of S-representation for cognitive science.

The chapter proceeds as follows. **Section 2** outlines the S-representation account. **Section 3** surveys how the S-representation account is vindicated in the empirical literature. **Section 4** examines two potent objections. The first objection says that S-representation is defined using a structure-preserving relation (isomorphism) that is too demanding. The second objection says that the S-representation account cannot naturalise the consumer required for genuine representation. I use this discussion as an opportunity to clarify the central commitments of the S-representation account as I present it. Finally, **section 5** presents Ramsey’s (2016) distinction between the ‘functional role dimension’ and ‘content grounding dimension’ of representation. Though Ramsey’s distinction is useful for highlighting what a complete account of representation must address, I raise two reasons to be cautious. The first reason is that the distinction, as presented, blurs the notions of ‘representational target’ and ‘representational content’ which are importantly separable under some accounts of content. The second reason is that Ramsey invites a traditional causal-historical theory of content to supplement the S-representation account, and yet such a theory conflicts with a mechanistic framework. I close by indicating that the S-representation account, when embedded within a mechanistic framework, can address both dimensions in one fell swoop. This discussion lays the groundwork for a complete treatment of how to think about content in **chapter 5**.

2.1 Ordinary S-representation

Previous chapters argued that for representation to play an explanatory role in cognitive science a purported cognitive representation must pass the job description challenge

(JDC) (Ramsey, 2007). I claimed that to pass the JDC, an entity must function in a distinctly representation-like manner in a way that implies representational content is of causal relevance to a cognitive system. The mechanistic account of representation further framed this perspective in terms of mechanistic explanation but did not demonstrate if or how a cognitive mechanism could fulfil the appropriate functional role. One strategy for demonstrating that a mechanism does function in a distinctly representation-like manner is to show that it functions analogously to a particular type of ordinary representation. The S-representation account deploys this strategy.

In broadest terms, the S-representation account concerns a type of representation characterised by the way a vehicle structurally corresponds to that which it functions to represent for a consumer (what I will come to label its 'target'). One set structurally corresponds to a second set when elements of the first set map to elements of the second set such that some relation between elements in the first set is preserved. Two sets are 'homomorphic' just in case there exists *some* structure-preserving mapping between elements of the two sets. Two sets are 'isomorphic' just in case there exists a *one-to-one* structure-preserving mapping between elements of the two sets. In this way, isomorphism is a type of homomorphism that admits an inverse. We will return to how best to unpack structural correspondence in cases of cognitive S-representation below.

Agents exploit 'ordinary S-representations' in everyday life to learn and reason about the entities that they stand-in for via structural correspondence. Cartographic maps are one example. Cartographic maps preserve various relations between geographical features in the relations between their own parts. Maps thus mirror the structure of a geographical region (to an appropriate degree of approximation). For instance, maps commonly

preserve the proportional spatial distance between points in a geographical region, where the proportional distance between, say, Lisbon, Berlin and Edinburgh might be preserved in the proportional distance between three marks on a map.[1] Similarly, contour lines mirror the elevation of points in real space by systematically joining points on the map that correspond to an equal height. As a result, the space between contour lines mirrors the difference in elevation between points in a geographical region. Agents exploit this structural correspondence, using the map to make inferences about features of a geographical region. For example, mountaineers use contour lines to learn about the topography of a mountain, which allows them to, say, judge the steepness of a slope. By playing this role, the map as a whole can be said to stand-in for the mountain, and elements of the map can be said to stand-in for particular features of the mountain.

Notice that maps often err by failing to structurally correspond to some geographical region. This affects the outcome of behaviours that depend on maps. When a mountaineer draws conclusions about the peaks, penitents and plateaus of a mountain by examining their map, the truth of those conclusions depends on how adequately the map structurally corresponds to the mountain. In turn, the mountaineer's failure to draw the right inferences about the mountain is sometimes explained by a failure of the map to structurally correspond to the mountain. This failure need not be all or nothing: parts of the map might fail to correspond whilst others succeed, and the failure of parts might be more-or-less severe—contour lines might stray from their real-world counterparts with greater or lesser severity. Error is often a matter of degree. In any case, the explanatory

[1] The preservation of proportional distance makes the correspondence abstract or formal. Two marks on a map representing Lisbon and Berlin are not required to be 1438 miles apart; the distance between the two marks need only preserve the proportional distance between the two cities, as determined by the relations between the two marks and the other marks on the map representing, say, other cities.

importance of error reveals that the correctness or incorrectness of the map is causally relevant (and crucially so) to the outcome of a mountaineer's behaviour.

In his influential paper, Swoyer (1991) offers another example of ordinary S-representation: a model plane in a wind tunnel that facilitates 'surrogative reasoning' about how a real plane will perform when flying. More generally, engineering and scientific models are good candidates for ordinary S-representation. Like cartographic maps, many ordinary models preserve spatial relations of a domain in the spatial relations of their own elements; for example, an orrery preserves the relative distances between solar bodies. More abstract models preserve various physical and temporal relations within equations; for example, a mathematical model of climate change might preserve, amongst other things, the relationship between values of CO_2 in the atmosphere, solar radiation, and average terrestrial temperature over a period of time. In each case, structural correspondence endows agents with the ability to conduct surrogative reasoning, allowing them to draw inferences about phenomena in the absence of direct observation. Models of planetary motion, DNA binding and climate change allow scientists to form predictions and explanations of diverse phenomena that challenge more direct forms of interaction. Mirroring cartographic maps, the correctness of conclusions drawn about *explananda* using such ordinary models can be explained by reference to the degree of structural correspondence that obtains.

Like ordinary maps or models, 'cognitive S-representations' are said to affect the behavioural outcomes of a consumer by structurally corresponding to some entity relevant to behaviour. For both ordinary and cognitive S-representation, understanding a vehicle as a stand-in for some entity on behalf of a consumer is critical for knowing how

a system fails or succeeds to achieve some capacity. By functioning analogously to ordinary S-representations, cognitive mechanisms play a causal role that is distinctly representation-like.

2.2 Cognitive S-representation

The S-representation account offers a well-articulated and empirically driven notion of cognitive representation in terms of a cognitive vehicle's ability to structurally correspond to some entity in a manner that is crucial to explaining the behaviour of a cognitive system (for influential discussion, see Cummins, 1989). This mirrors the characteristics of familiar S-representations like maps and models. Indeed, those frameworks within cognitive science that appeal to viable instances of S-representation, such as the classical computational theory of cognition, often conceive of representations as serving as internal maps or models of the distal world on behalf of the cognitive system (for an early expression of this idea, see Craik, 1943). If the S-representation account is correct, then explaining certain cognitive capacities will plausibly require reference to the exploitation of inner structures that map or model some behaviourally relevant target in a manner that parallels cases of ordinary S-representation.

Following Gładziejewski (2015), I take it that the S-representation account offers four functional criteria for cognitive S-representation (though my articulation differs somewhat). I take each of these conditions for cognitive S-representation to match those necessary for ordinary S-representation. The four criteria are as follows:

1. **Structural correspondence**: An S-representation *R* structurally corresponds to (or has the capacity to structurally correspond to) some entity *x*.

2. **Action guidance**: *R* affects the behaviour (cognitive processing and/or motor output) of some cognitive system *S* in relation to *x*, such that some capacity of *S* causally depends on *R*'s structural correspondence with *x*.

3. **Decouplability:** *R* need not be causally-coupled with *x* to perform action guidance for *S* in relation to *x*, i.e., *R* may be temporally and/or spatially distant from *x*.

4. **System-detectable error**: *S* is (at least sometimes) causally sensitive to some lack of structural correspondence between *R* and *x*, such that the lack of structural correspondence between *R* and *x* affects the behaviour of *S* (cognitive processing and/or motor output). Such behaviour may include updating *R*, i.e., changing *R*'s structure to strengthen structural correspondence between *R* and *x*.

The central conjecture of this chapter and the next is that these four criteria are sufficient for a mechanism to play a functional role that is distinctly representation-like. In turn, if a theory of cognition posits mechanisms that meet these four criteria, then that theory counts as a representational theory. We have not yet detailed precisely how to conceive of content within the S-representation account; however, the possibility of a functional parallel between ordinary artefacts and cognitive mechanisms already indicates the

plausibility of causally-relevant correctness conditions at the subpersonal level. To fully appreciate how these four criteria allow an entity to pass the JDC, we must unpack them further, paying special attention to structural correspondence and its interaction with the other criteria.

As with ordinary S-representation, the idea of structural correspondence (sometimes 'structural resemblance' or 'structural similarity') lies at the heart of cognitive S-representation. O'Brien & Opie clarify this condition when they write,

> [O]ne system *structurally resembles* another when the physical relations among the objects that comprise the first preserve some aspects of the relational organisation of the objects that comprise the second. (2004, p. 15. Original emphasis.)

For two entities to structurally correspond to one another they need not share the same 'first-order properties' (such as colour, size, density etc.), but only certain 'second-order relations' (O'Brien & Opie, 2004; Morgan, 2014, p. 223). Exploiting a structural correspondence between two superficially different systems is common to ordinary S-representations such as maps and models. For example, mathematicians might exploit a geometric diagram that represents an object moving through space to predict that object's velocity (Cummins, 1989). As Morgan puts it: 'we must look to the functional, dispositional properties of the mechanism, not to its static, categorical properties' (2014, p. 221). O'Brien & Opie further refine this idea when they write,

> Two systems can share a pattern of relations *without* sharing the physical properties upon which those relations depend. Second-order resemblance is actually a very abstract relationship. It is a mathematical or set-theoretic notion […] Essentially nothing about the physical form of the relation defined over a system S_V of representing vehicles is implied by the fact that S_V resembles S_O at second-order; second-order resemblance is a formal

> relationship, not a substantial or physical one. (2004, p. 13. Original emphasis.)

The formal nature of structural correspondence allows one entity to represent another even when their first-order order physical properties diverge. For example, neural activity can represent features of a geographical region through correspondence between firing rates of neurons and spatial distance between points in the environment (Bechtel, 2016; see **section 3.2** below).

The structural correspondence relation is often unpacked in terms of 'isomorphism'—that is, a one-to-one structure-preserving mapping. It is worth raising and setting aside an immediate worry that one may have about characterising cognitive representation in this way. The worry is that isomorphism possesses the wrong formal properties and so isomorphism cannot ground representation, hence structural correspondence cannot be a condition for cognitive representation. I take 'symmetry' and 'reflexivity' to be the most pertinent properties and so take these as my focus (for related discussion, see Goodman, 1968).

First, if isomorphism is a 1-1 correspondence, it is a bijective relation and admits an inverse; in other words, isomorphism is symmetrical. This means that if A is isomorphic to B, then B is isomorphic to A. However, symmetry is not characteristic of everyday representation (Sprevak, 2011, p. 671). A map may represent the Himalayas without the Himalayas representing the map. Isomorphism implies symmetry and so, the thought goes, implies the wrong formal relations for representation. Second, isomorphism is a reflexive relation because every structured object is 'automorphic': any structured entity will possess a 1-1 correspondence with itself. However, such reflexivity is not

characteristic of everyday representation. A map may represent the Himalayas without representing itself. Isomorphism implies reflexivity and so, the thought goes, implies the wrong formal relations for representation.

The problem with symmetry and reflexivity is that they seem to overproliferate the representing relation. At this stage, we should note that isomorphism is not necessary for defining structural correspondence. Isomorphism is one type of 'homomorphism'. As defined above, homomorphism refers to any structure-preserving mapping. Homomorphism encompasses isomorphism but also permits one-to-many and many-to-many mappings. Defining structural correspondence in terms of the more liberal notion of homomorphism is advantageous in avoiding other worries about the S-representation account, in particular, the worry that isomorphism is too restrictive (see **section 4** below). However, allowing for one-to-many mappings will not help overcome the present worry because all that does is permit additional kinds of mappings to fulfil the structural-correspondence condition alongside isomorphism. For instance, allowing for weaker forms of structural correspondence does nothing to prevent the fact that a structured entity is isomorphic to itself. The solution to the worry that isomorphism implies the wrong formal properties rather lies in stressing that structural correspondence is not sufficient for representation.

According to the S-representation account, one entity does not represent another just because some structural correspondence obtains. Equally important is the way a system's capacity depends on the exploitation of one entity's resemblance to another for action (for related discussion see Gallistel 1990; Cummins 1996). As Shea puts it, S-representations involve a form of 'exploitable isomorphism' (2014). The criteria for S-representation

ensure that the direction of action guidance between two entities plays a part in determining the 'representation' and the 'represented'. At least one reason why a map may represent a mountain range, but the mountain range does not (usually) represent the map is because the structure of the map is exploited to guide action towards the mountain range and not vice-versa. Similarly, the map does not (usually) represent itself because the structure of the map is not exploited to guide action towards itself. The necessity of action guidance ensures that an S-representation *R* may stand-in for an entity *x* without *x* standing-in for *R* (avoiding symmetry) and without *R* standing-in for *R* (avoiding reflexivity).

The worry that isomorphism implies the wrong formal properties is closely related to another objection to structural-correspondence based accounts of cognitive representation: that they trivialise content or entail 'massive indeterminacy in content' (Sprevak, 2010, p. 671). This objection is addressed in **chapter 5** where my solution is similar to the one just offered: the semantic properties of a given cognitive representation are determined (and thus limited) by the context in which the representation is exploited by a consuming cognitive system. Structural correspondence is only of semantic significance when it is used by a cognitive system to act in the world.

Building on the need for more than mere structural correspondence, the 'decouplability' criterion captures the fact that paradigmatic representations guide their consumer's actions without needing to directly interact with that which is represented (Haugeland, 1991).[2] After all, ordinary S-representations, like maps and models, are notable for their capacity to influence the behaviour of an agent in relation to a represented entity without

[2] Both decouplability and system-detectable error can be viewed as subcriteria of action guidance, in so far as they clarify and substantiate the conditions for action guidance that S-representation involves.

the representation being causally coupled to the represented entity. Decouplablity underlies some of ordinary representation's most impressive feats of action guidance, such as when mountaineers plan a trek of the Himalayas, or when scientists predict the behaviour of distant stars approaching supernova.

There are two kinds of decouplability: 'weak' and 'strong' (Gładziejewski, 2015, pp. 77-78). Weak decouplability requires only that there is no direct causal connection between some entity *x* and a representation *R*. Strong decouplability further requires that there is no causal connection between *x* and *R*'s consumer. Weak decouplability occurs in ordinary S-representation when a mountaineer uses a map as they trek the Himalayas: the map informs the agent about the Himalayas, and the map is not directly causally connected to the Himalayas—and yet the mountaineer is causally connected to the Himalayas. Strong decouplability occurs in ordinary S-representation when the mountaineer uses a map before they trek the Himalayas: the map informs the agent about the Himalayas, and the map is not directly causally connected to the Himalayas—and furthermore, the mountaineer is not causally connected to the Himalayas. To borrow from Gładziejewski, in cases of strong decouplability, S-representations 'enable the system to guide its action with respect to states of affairs that are not "reliably present" for the system because they are not present for the system *at all*' (2015, p. 78. Original emphasis).

'System-detectable error' further strengthens the action-guidance criterion—though it may strike some as the weakest ingredient in the above recipe for S-representation. Advocates of the system-detectable error criterion claim that it is only within systems capable of detecting error for which incorrect representation can be said to matter *for that system* (for this line of thought, see Miłkowski, 2013, 2015b; Gładziejewski, 2015;

following Bickhard, 1999). Nonetheless, one might contend that 'representational error' (*sans* system-detection) has a role to play for a mechanism that meets only the first three criteria for S-representation. This is because it seems that decouplable action guidance via structural correspondence is sufficient for something to stand-in for something else on behalf of a cognitive system. I leave this debate aside. For the purposes of this chapter and the next, I will adopt the system-detectable error criterion as necessary because I think it presents the strongest case for S-representation. As we shall see, the system-detectable error condition supports the action-guidance condition by strengthening the notion that it is the cognitive system itself which counts as a consumer. This is because the significance of structural correspondence between a mechanism and some entity for a cognitive system is strengthened if that system is sensitive to a lack of structural correspondence between the mechanism and entity in question. However, it should be noted that if the S-representationalist can establish that system-detectable error is not required for a cognitive mechanism to possess a function analogous to ordinary S-representation, then the empirical possibility of cognitive S-representation is strengthened, as the conditions for its realisation are weakened.

There are two *prima facie* worries one may have about the possibility of system-detectable error in a cognitive system. The first is how a cognitive system could detect a lack of structural correspondence between an S-representation and what it represents. The second is that the system-detectable error condition smuggles in representational content, which is the very thing at stake. This is because the ability to detect 'error' implies the ability to detect something false, inaccurate etc., the thought goes. Responding to both these worries requires us to observe how cognitive systems ordinarily receive feedback, namely through sensitivity to the outcome of their own behaviour. Within the literature,

system-detectable error has been most straightforwardly unpacked in terms of a system's sensitivity to 'action failure' (for example, see Bickhard, 1999). In this sense, error detection is second-hand: a cognitive system is not sensitive to the degree of structure-correspondence between a vehicle and the entity it represents *per se* but to the failure of action that results from that relationship. Gładziejewski frames this as follows,

> [T]he mechanism in question should be equipped with internal components whose function is to detect the fact that the action guided by the representation vehicle (through its effect on the representation consumer) fails to achieve success. (2015, p. 80)

This parallels many instances of error detection in ordinary S-representation. For example, a mountaineer might detect a lack of structural correspondence between a map and a mountain when they follow a route which the map depicts as a gentle slope only to discover a sheer cliff face. In other words, the mountaineer is sensitive to a lack of structural correspondence via their sensitivity to the failure of their own actions. This sensitivity to action failure does not itself presuppose representational content. It only presupposes that there is (i) some capacity that the system attempts to complete, and (ii) some method by which the system receives feedback on the success or failure of that capacity. System-detectable failure suggests incorrectness only when it results from a mechanism with other representation-like characteristics, such as when it results from a decouplable mechanism that guides action via structural correspondence. Once again, the mutual interplay between all four conditions for S-representation is important for justifying the representational label.

To close our initial sketch of the S-representation account, it is worth noting that there are two ways to identify a representational vehicle given the criteria for S-representation (for

a similar point see Ramsey, 2018, p. 261). The first way is to conceive of the 'whole S-representation' as the vehicle. The second way is to conceive of individual parts of the S-representation as the vehicle. In mechanistic terms, this is roughly the difference between a whole mechanism that structurally corresponds to, say, some geographical region, and the mechanism's components that individually correspond to elements of that region. This reflects a difference pertaining to ordinary S-representation too, for example, the difference between a whole cartographic map and its individual marks. A whole cartographic map may be regarded as a representation of a geographical region whilst its individual marks may be regarded as representations of features within that geographical region. This is the difference between a map standing-in for Nepal (in virtue of its global structural correspondence with Nepal), and a mark on that map standing-in for Kathmandu (in virtue of its relative position on the map).

To my mind, there is no disadvantage to viewing both the whole S-representation (i.e., the whole mechanism) and its parts as representational vehicles. However, it is important to stress that, like cartographic maps, those parts only serve as representations in so far as they bear relations to other parts within the wider S-representation. A mark on a map functions as a representation because of its relationship to other marks that together realise a structural correspondence with the represented domain (for related discussion see Ramsey, 2007, p. 78; 2018, p. 261). As such, a contour line does not function to represent because of some intrinsic feature of the contour line itself. Rather, it functions to represent because its relationship to other contour lines realises a structure that mirrors the relationship between points of elevation in the world (which in turn allows one to draw inferences about the world from the map). This holistic perspective has implications for how we understand the representational role of 'symbols' posited by frameworks like

the classical computational theory of cognition. If cognition is to be explained in terms of computations performed over symbols implemented in the brain, then as far as the S-representation account is concerned, those symbols do not represent anything in virtue of their intrinsic properties but represent (if at all) in virtue of their role as elements in a wider structure-preserving mechanism.

2.3 S-representations as mechanisms

The S-representation account and the mechanistic account of cognitive representation (see **chapter 2**) are conceptually distinct: a proponent of the S-representation account is not necessarily a proponent of the mechanistic framework and vice versa. Nonetheless, the two accounts make strong companions. This is because we can understand the vehicles which realise the functional criteria for S-representations as cognitive mechanisms. Recall that, according to the mechanistic account of representation, M is a cognitive representation iff M is a mechanism that has the causal role to stand-in for something relative to some capacity of a cognitive system. The S-representation account provides the missing specification for how this role could be realised.

The benefits of uniting the S-representation account with a mechanistic approach go two ways. On the one hand, the mechanistic account of cognitive representation is substantiated by a story of what properties a mechanism ought to possess to count as a representation; the S-representation account articulates a set of well-specified functional criteria that may be fulfilled by a class of cognitive mechanism. On the other, the S-representation account is strengthened by elucidating what sort of theoretical entity meets its criteria; a mechanistic approach ensures that S-representation ascriptions are consonant with our dominant explanatory paradigm, and in doing so, clarifies what it

means for a theory to ascribe S-representations. Hereafter, a 'representational mechanism' or 'R-mechanism' will refer to a cognitive mechanism with a causal role (relative to some cognitive capacity) that meets the four criteria for S-representation. To be a cognitive representation is to be an R-mechanism.

2.4 S-representation and classical computation

Chapter 3 argued that computation and representation are distinct functional kinds. However, this does not mean that a computing system cannot also be a representing system. The idea that the states and processes of a computing mechanism might be exploited by a cognitive system to map or model a domain, offers a clue as to how to provide representational content with a proper role within a computational approach. This is because we can start to see how the correspondence between digits and their transformation in a computing system and states of affairs relevant to behaviour could play a systematic role in explaining the capacities of a cognitive system.

Under the S-representation account, the computational processes of cognition remain 'mechanical'. In this sense, there is no difference between the causal properties of an S-representation and any old (non-S-representational) computing system. What makes the former representational is that the success or failure of the containing system depends on the degree of correspondence between the structure of a computing mechanism and some target entity, and because of this, we can reasonably characterise the functional role of that mechanism as 'standing-in for' features of that entity (more on the significance of targets and a system's success/failure in **chapter 5**). As alluded to above, parts of an R-mechanism can be identified as standing-in for parts of a domain given their role in the wider structural correspondence of that domain. When we ask *what* it is about marks on

a map that facilitate successful navigation, we must say that they accurately model features of the task-relevant environment. When we further ask *how* they model the environment, we must say that they preserve pertinent relations between features of the environment. By doing so, they stand-in for those features. An analogy to this sort of schema underlies Ramsey's (2007) defence of the role representation plays in the classical computational theory of cognition (CCTC). The CCTC commonly presents cognition as involving the construction and deployment of inner cognitive models, realised by a classical architecture. These inner models preserve the structure of some task domain. Symbols within the CCTC (over which computational operations are performed) can be understood as standing-in for parts of that domain. Ramsey writes,

> Understanding how computers work involves understanding more than the nature of their physical operations. We also want to understand what it is about those specific operations that enable the system to perform some sort of task. (2007, p. 86)

We are compelled to ask,

> "What is it about the causal/physical nature of this system that enables it to solve a particular problem?" And the CCTC answer is this: These syntactic/physical operations are successful in solving this problem because they implement a model of a problem domain, and, as such, employ elements that *stand for* various aspects of that domain. (*ibid.*, p. 86. Original emphasis.)

The CCTC posits internal models that preserve the structure of some domain (the structural-correspondence condition). This structural correspondence causally affects the system such that it determines, in part, the success or failure of a system's capacity (the action-guidance condition). These models are constructed and operated on at a temporal and spatial distance from the entity which they stand-in for (the decouplability condition), and we can easily imagine these models updating following feedback (the system-detectable error condition). In short, the CCTC plausibly posits S-representations.

Ramsey (2007) affirms the role of S-representation in the CCTC. And yet, he offers a pessimistic prognosis for S-representation.[3] This is because the dominant computational paradigms of contemporary cognitive science, in particular connectionism, eschew any substantive notion of internal structure-preserving models. In response, several authors have noted that Ramsey's (2007) analysis does not match the commitments of many connectionist approaches or contemporary cognitive neuroscience more generally. It is worth briefly considering some of the counterexamples which have been proposed.

A notable instance of this rebuke can be found in Shagrir's (2012) discussion of the contemporary 'model-based' understanding of the oculomotor system. Motor control in the oculomotor system is taken to be governed by the modelling of eye position. This is achieved because different states in the underlying 'attractor network' are responsible for encoding a unique eye position; the distance between two eye positions is systematically mirrored in the distance between states in the line attractor (which describes eye position in response to stimuli).[4] Shagrir writes, 'The state-space of the network could be seen as a *map* whose line attractor corresponds to the space of eye positions' (p. 14. Original emphasis). In other words, relations in the underlying computational process systematically mirror eye position.

[3]Ramsey's recent work expresses a more optimistic outlook. In discussing the S-representation account he writes, 'In truth, virtually every area of cognitive modelling has involved theories that appeal to representations of this sort. This includes various accounts of reasoning, knowledge representation, memory, learning, navigation, perception, language comprehension, motor control, and several other cognitive competencies' (2018, p. 263).

[4] In connectionist approaches, an attractor network is a type of recurrent network—basically, a neural network that uses feedback loops to feed information back into the system—that over time evolves toward a stable pattern (defined as a subset of possible states). Attractor networks are frequently used in contemporary computational neuroscience to model cognitive capacities.

Also responding to Ramsey (2007), Sprevak makes a related point regarding the generalisability of a model-based notion of representation. Far from being the sole remit of the CCTC, 'It is commonplace in cognitive neuroscience, connectionism, indeed all areas of cognitive science, to explain behavioural success in terms of the agent's inferences about internal models' (2011, p. 673). He offers the example of explanations that depict an agent making inferences about 'edges' in her visual field via the modelling of edges implemented by neurons in V1 (*ibid.*). Along similar lines, Grush (2008) suggests that Ramsey (2007) underplays the role of S-representation in contemporary motor-control theory. This is because of the prevalence of 'forward models'. A forward model is a model-based prediction of the agent's body position that takes the input of a motor command and outputs an estimation of body position. These outputs can then be compared to actual body position, with resulting error fed back to inform future motor commands. Grush's suggestion that forward models are potential S-representations notably anticipates Gładziejewski's (2016b) sustained defence of S-representations in the predictive processing framework (see **chapter 2** and **section 3.1** below).

I do not intend my brief sketch of these counterexamples to conclusively demonstrate the need for S-representations in our best computational explanations. Nor do I mean to take sides with regards to the effectiveness of any particular approach. I mean only to indicate that the empirical relevance of the S-representation account is not straightforwardly tied to the fate of the CCTC. We now turn to a more detailed example of S-representation in the contemporary cognitive science literature.

3.1 Cognitive S-representation in action

The S-representation account proposes a set of conditions that may be met by several otherwise distinct cognitive mechanisms. Gładziejewski describes the conditions for S-representation as a '*highly* idealized *sketch* of a *possible* mechanism' (2015, p. 69. Original emphasis). This is partly because the account offers functional criteria that may be realised by components with different structural details or with roles in different cognitive capacities. For example, psychologists working within mental model theory postulate 'mental models' (Johnson-Laird, 1983, 1998) and 'mental images' (Kosslyn, 1994) as related but distinct representations.[5] Both suggest the possibility of R-mechanisms: many purported mental models and mental images plausibly involve underlying de-couplable mechanisms that guide action by way of encoding spatial-analogical information about entities in the world. Capturing the central role of structural correspondence in mental models, Johnson-Laird says, 'the structural relations between the parts of the model are analogous to the structural relation of the world' (1998, p. 447). At the same time, mental models and mental images are importantly distinct notions within much of the literature. For one thing, they are differentiated in their anatomical properties (for example, mental model theory localizes the parietal lobe as responsible for mental models, but the occipital lobe, particularly V2, for mental images). For another, these mechanisms are involved in different cognitive processes. These examples illustrate that the S-representation account does not refer to a single kind of mechanism as individuated by scientific practice but specifies the conditions for a family of possible mechanisms whose members share the same high-level functional properties.

[5] For a recent comparison of these two frameworks that highlights some of their representational commitments, see Sima, Schultheis & Barkowsky (2013).

Our discussion so far indicates that there are two broad strategies for illustrating how the S-representation account is vindicated by the empirical literature. The first strategy involves observing the generic posits of a theoretical framework, noting that these suggest functional properties that correspond to the requirements of S-representation. The second strategy involves identifying a specific cognitive mechanism whose functional properties correspond to the requirements of S-representation. Such a mechanism may be posited by multiple frameworks. The first strategy has at least two major precedents. The first is Ramsey's (2007) discussion of the classic computational framework, already visited above. The second is Gładziejewski's (2016b) discussion of the predictive processing framework. Gładziejewski's analysis was anticipated in our sketch of the predictive processing framework in **chapter 2** where we suggested that the framework provided *prima facie* support for the relevance of 'non-natural information' in cognitive science. A complete discussion of Gładziejewski's analysis would take us too far off course. However, Gładziejewski succinctly summarises how generative models in the predictive processing framework fulfil the criteria for S-representation when he writes that,

> [C]ognitive systems navigate their actions through the use of a sort of causal–probabilistic "maps" of the world. These maps play the role of representations within the theory. Specifically, this map-like role is played by the generative models. It is generative models that, similarly to maps, constitute action-guiding, detachable, structural representations that afford representational error detection. (2016b, p. 569)

In short, the generative models of predictive processing suggest decouplable, structure-preserving mechanisms that guide action and update in lieu of error; they suggest R-mechanisms.

‘Cognitive maps’ serve as a notable example of the second strategy for illustrating how the S-representation account is vindicated in the empirical literature. Cognitive maps play a prominent role across contemporary cognitive psychology and cognitive neuroscience.[6] As cognitive maps are frequently discussed in the literature on S-representation and will serve as a useful reference point in the remainder of this chapter and the next, it is worth discussing their essential features.

3.2 Cognitive maps as S-representations

First posited by Tolman (1948), the idea of a ‘cognitive map’ (sometimes ‘mental map’ or ‘hippocampal map’) gained traction with O’Keefe & Nadel (1978) and was developed by Gallistel (1990).[7] Often studied through ethological and neurophysiological experiments on rats, cognitive maps are mechanisms, principally located within the mammalian hippocampus, that appear to play a crucial role in spatial-navigational capacities (see Bechtel, 2016, for an overview). Specifically, cognitive maps are thought to be responsible for encoding, storing and decoding information about locations and features of an organism’s spatial surroundings. In Tollman’s (1948) original experiments, rats were trained to navigate mazes and locate food. It was discovered that if a rat was trained on a circuitous route, then placed in a maze with a more direct path to the reward, the animal would take the direct route rather than the circuitous route which had been reinforced in previous trials. This was taken to indicate that the rats did not navigate solely via stimulus-response learning. In subsequent experiments, various other tools were

[6]Some do still doubt the evidence for cognitive maps. For example, see Bennett (1996); Jensen (2006).

[7] Neuroanatomy is rife with talk of sensory and motor ‘maps’ that serve as potential S-representations: for instance, ‘somatotopic maps’ and ‘retintopic maps’. It is often suggested that these serve as a kind of ‘topographic representation’, in so far as they instantiate an ordered correspondence between neural structures and a sensory surface—for example, the body to the sematosensory cortex and retina to the primary visual cortex. I leave assessing the representational credentials of these mechanisms aside, and focus on cognitive maps, as they more straightforwardly demonstrate the criteria for S-representation that I have set out.

employed to control for factors such as local environmental cues outside the maze—for instance, the 'Morris Water navigation task' in which the animal must navigate a pool of water that obscures possible auditory, visual, or olfactory cues (Morris, 1984).

Present evidence indicates that the neurological basis for these mechanisms resides in large part in the selective activation of specialised 'place cells' and 'grid cells' corresponding to specific spatial locations, which collectively function as a map of an environment (for example, see Moser, Kropff & Moser, 2008). Of significance, the firing rate and strength of connections between cells within cognitive maps are thought to correspond proportionally to the distances between features in the environment; that is, they structurally correspond. To clarify, a cognitive map does not preserve the spatial relations between elements of an environment in the spatial relations of neurons. Rather, a cognitive map preserves spatial relations within the relations between firing rates and the strength of connections between neurons. As such, a cognitive map is not a literal cartographic map. Recall that structural correspondence is a second-order relation: one set need not possess the identical spatial relations of a second set for the first set to preserve the structure of the second set. More technically, what matters is that the first set instantiates an exploitable 'geometric structure'—a set of objects satisfying those axioms, pertaining to topographical or metrical relations, that function to preserve relations between objects within another set (for discussion see, O'Brien & Opie 2004; Rescorla, 2009). Within the computational approaches that reference them, cognitive maps are expected to serve as maps in so far as the organised objects of which they are composed satisfy a set of relevant axioms (Rescorla, 2009).

Cognitive maps are candidates for R-mechanisms because they cohere with the four criteria listed above. By encoding features of the animal's environment, cognitive maps *guide action* through *structural correspondence*. The maps are also (strongly) *de-couplable* as they are thought to be involved in anticipating and planning future behaviour (Bechtel, 2016; Miłkowski, 2015b). Finally, cognitive maps are highly modifiable as the rat's environment changes, and thus involve a form of *system-detectable error* (Jeffery, 2015). The precise role of cognitive maps within the wider cognitive economy and their pervasiveness matters less for our purposes than the fact that they illustrate a plausible case of how the S-representation account might be vindicated. Note that though cognitive maps form a plausible case of an R-mechanism, proponents of the S-representation account (as an account of cognitive representation) need not carry any specific commitments about whether and where S-representations obtain (as an empirical fact).

So far, we have sketched the S-representation account and witnessed how it might be vindicated in the empirical literature. We now turn to explore two possible objections. In doing so, we will come to better understand the S-representation account, as I intend it. This discussion will also lay the groundwork for a proper examination of content through the lens of S-representation, to be taken up fully in **chapter 5**.

4.1 Objection 1: Isomorphism is too strong

One notable objection to characterising representation in terms of structural correspondence claims that the condition is too strong. Recall that structural correspondence is frequently cashed-out in terms of isomorphism. Strictly speaking, an isomorphism's bijective mapping demands that two systems are structurally equivalent. This strong reading of structural correspondence forbids intuitive cases of S-

representation, disallowing any divergence from perfect 1-1 correspondence (Sprevak, 2011; Shea, 2014). This is not a problem for relatively straightforward cases of S-representation with few elements in need of correspondence; for example, Galileo's classical representation of nonspatial magnitudes using geometrical figures (see Cummins, 1989, pp. 28-29). However, these strict conditions would seem to prohibit many intuitive cases of S-representation. For instance, an octagonal map could still be used to represent Nepal, despite failing to meet the conditions required to be strictly isomorphic with the geography of Nepal—think of the simple shapes used to represent territories in the board game Risk. This is an extreme example, but as Shea points out,

> [E]ven the most accurate map does not satisfy this requirement since there are always slight inaccuracies (for example due to projecting a curved world onto a flat sheet). So spatial relations on maps are simply not isomorphic in the mathematical sense to spatial relations on the ground. (2014, p. 124)

To overcome this problem, we must relax our understanding of the mapping relation implicit in the structural-correspondence criterion for S-representation. The most straightforward way to do this is to replace isomorphism with a weaker structure-preserving relation. The generic term 'homomorphism' captures this possibility (O'Brien & Opie, 2004).[8] As we saw above, not all homomorphs are isomorphs, and some allow for one-to-many mappings. Homomorphism allows for imperfect correspondence, making it possible for one object to preserve the structure of another whilst 'losing information'. 2D maps are homomorphic to 3D environments; that is, they preserve the structure of 3D environments but not perfectly so.

[8] It strikes me that fears over the appropriateness of isomorphism result from misunderstanding the way the term is often deployed by psychologists. As I understand it, psychologists have historically adopted a more liberal notion of 'isomorphism' than mathematicians, where isomorphism (in the psychologist's sense) is equivalent to homomorphism (in the mathematician's sense).

This solution allows us to define S-representation in a way that allows for 'partial structural correspondence'. Strict isomorphism sets the bar for representation too high, but we can adjust the standard for structural correspondence to a degree of reasonable attainability. Again, this feature of cognitive S-representation is paralleled in cases of ordinary S-representation. Clearly, a 2D cartographic map does not need to be strictly isomorphic to a geographical region in order to represent it. Drawing on a theme from **section 2.2** above, what matters in the case of a cartographic/cognitive map is that the cartographic/cognitive map mirrors the structure of the relevant environment such that it serves the task at hand, be it navigating a mountain range or a laboratory maze.

Chakravartty (2010) raises a similar point in relation to scientific models (for related discussion see Frigg, 2006). Chakravartty argues that strict correspondence relations such as isomorphism are idealisations used in 'epistemological theorizing' but that such idealisations are not necessary in practice:

> [B]y itself the perhaps ubiquitous failure of strictly defined mathematical similarities between representations and their targets tells us nothing about whether such similarities obtain, not strictly, but within reasonable bounds of approximation. (*ibid.*, p. 8)

In short, by relaxing the strength of correspondence that we expect to obtain between representation and represented we solve the worry that S-representation implies a condition for cognitive representation that is too demanding. Structural correspondence only matters to the degree that it serves the actions of a consuming system. Yet this reliance on a 'consuming system' invites a more serious worry.

4.2 Objection 2: Cognitive S-representations have no consumer

Representation requires a consumer for whom a vehicle represents. In cases of ordinary representation, the consumer takes the form of an interpreting agent; for example, a mountaineer in the case of a cartographic a map or a scientist in the case of a climate model. My defence of the S-representation account depends on an analogy to a type of ordinary representation. But is there really an analogous consumer in cases of cognitive S-representations? In answering, I favour a response that stresses the role of an R-mechanism for the 'whole system' whose capacity it underwrites.

There has been a tendency within the literature to treat the consumer of a cognitive representation as an identifiable sub-system that possesses a function which is dependent on a representation for success (Millikan, 1984; Papineau, 1984). Though such functions are often understood in terms of a sub-system's selected 'proper function' (see **chapter 2**), Gładziejewski (2015) provides a non-etiological spin that is well-suited to our mechanistic inclinations. For Gładziejewski, a consumer is a mechanism component with a functional role that depends on a state of affairs (external to the mechanism) that it must be 'adapted' to in order to succeed (*ibid.*, pp. 74-76). To count as a consumer, the component must be causally coupled with the representational vehicle but not causally coupled with the external state of affairs, and its functional role must depend on the representational vehicle's structural correspondence to those state of affairs (*ibid.,* p. 74).

Following Cummins (1996) and Ramsey (2007), Gładziejewski (2015) offers the following toy example to illustrate his idea of a consumer component: imagine a self-driving car that must navigate through an S-shaped track. The car can navigate through the S-shaped track in part because of a rudder attached to an S-shaped groove (an 'internal

map') inside the car. The rudder moves along the groove which in turn steers the wheel that moves the front wheel. In this example, Gładziejewski understands the steering wheel to be the consumer of the representation. The steering wheel must be 'adapted' to the S-shaped track to play its role, but the steering wheel is not causally-coupled with the track. Instead, the steering-wheel depends on a correspondence between the representation (the groove) and the track: 'The steering wheel makes use of the representational vehicle – the groove – to succeed at performing its function' (*ibid.*, p. 75).[9] It is important for Gładziejewski that the groove 'indirectly guides the car (the system as a whole) by directly "guiding" the steering wheel (internal consumer)' (*ibid.*, p. 75). Note that selected functions are not required to make sense of the steering wheel as a consumer in this example: what matters is that the car's steering through the track (the capacity being explained) causally depends, in part, on the representation structurally corresponding to the track such that the steering wheel turns appropriately (though see footnote 9 below).

Following Gładziejewski's definition, I think it is reasonable to characterise the steering wheel as a consumer component. However, I am uncertain whether such a consumer component is necessary or of central importance to the S-representation account. Recall that cognitive maps earn their S-representational credentials in virtue of their functional role relative to causing a cognitive capacity—say, successful spatial-navigation in a rat. Given this, it is unclear that some separate consuming subsystem is either identifiable or required for cognitive S-representations to count as being 'consumed'. What matters is

[9] Though our description of the car involves structural correspondence, action guidance and (weak) decouplability, it does not involve 'system-detectable error'. Strictly speaking, then, the system does not possess an S-representation under the four criteria listed above. It is easy to supplement the example with the system-detectable error criteria (as Gładziejewski later does; 2015, pp. 80-81). However, note that it is because examples like this might intuitively appear to involve representation even in the absence of system-detectable error that one may want to drop the system-detectable error condition, as suggested in **section 2.2** above.

the role played by an R-mechanism in causing the rat to navigate. In this way, the whole system—the rat in this case—can be understood as the consumer, for it is the rat's capacity that depends on the mechanism playing a representational-role. So long as an R-mechanism plays a causal role in a capacity of that system, *qua* S-representation, then we have sufficient grounds to think that a representation is being consumed.

In drawing an analogy to ordinary maps and models, it is important to stress that conscious interpretation is not required for an R-mechanism to be consumed. Nor do we need to posit any inner interpreting 'homunculi' of increasingly simpler sub-systems, as some have imagined (Dennett, 1982).[10] This is because, *sans* conscious interpreter or inner homunculus, 'representation' continues to accurately describe the role of an R-mechanism within a cognitive system. To illustrate, imagine we were to replace a human map reader of the sort associated with ordinary S-representations with a robot dubbed 'NavBot'. In addition to some basic motor and navigational competencies (such as steering around environmental obstacles), NavBot is notable for the fact that it can scan ordinary cartographic maps that correspond to its environment and move itself to certain 'goal locations' depending on the properties of the map. Specifically, if NavBot detects an 'X' it will move to a corresponding location in the world. NavBot can do the following:

> **(i)** Scan the map and detect an 'X' mark.
>
> **(ii)** Estimate a location in the world (L_1) that corresponds to X's coordinates on the map.

[10] The type of strategy I offer here resembles a version of what Ramsey labels the 'mindless strategy' (2007, p. 193). Ramsey coins this term when discussing ways to avoid a regress of 'little inner minds' in accounting for the use of cognitive representation (*ibid*., p. 190).

(iii) Determine coordinates on the map that correspond to NavBot's own location in the world (L_2).

(iv) Plan a navigational route from L_2 to L_1.

(v) Move to L_1.

NavBot's ability to reach L_1 depends on a structural correspondence between X on the map and NavBot's environment. When asked *what* role the map plays (in causing NavBot to reach L_1), we can reasonably say it stands-in for features of NavBot's environment (such as the spatial distance between L_1 and L_2). When asked *how* it plays this role, we must say that the map preserves the structure of the environment through the relations of its own elements; for example, the proportional spatial distance between L_1 and L_2 in the environment corresponds to the proportional distance between NavBot's coordinates and X on the map. This role of the map for NavBot parallels the role of a map for an ordinary agent; NavBot consumes a cartographic map in a manner analogous to an ordinary agent. NavBot's map thus serves as a representation. Furthermore, there is no change in the role of the map-as-representation for NavBot if we (a) move the map inside NavBot's head and (b) change the map from an ordinary cartographic map to a digital map of the sort commonly used in 'robotic mapping' (for an introduction to robotic mapping see Thrun, 2002). The explanatory utility of characterising NavBot as a representation consumer remains unshaken by internalising and digitalising the map.

If the present analysis is correct, then the explanatory role played by the robot's consumption of the map (*qua* representation) is not undermined by the fact that NavBot is not consciously using the map as a representation of its environment. Nor do we need to invoke an internal homunculus that uses the representation. What matters is the role

the map serves in causing NavBot's behaviour and determining the success or failure of its actions. Returning to a theme established in **chapter 1**, if we assume that NavBot must be a conscious interpreter for its map to count as a representation then we eliminate the possibility of subpersonal cognitive representation playing an explanatory role from the outset. Such demanding standards are not warranted. What should be gained from ascribing representation to cognitive systems is some appreciable explanatory purchase—some contribution to our grip on the target phenomena. R-mechanisms achieve this by capturing a distinct functional role that so closely resembles a class of ordinary representation. Ramsey similarly rejects the need for conscious interpretation to justify the legitimacy of cognitive S-representation (as genuine representation) when he writes,

> Is S-representation comparable to full-blown conscious thoughts? No, it is a technical notion of representation based on our commonsense understanding of things like maps, invoked by a theory to explain cognition in a certain way. (2007, p. 89)

Of course, we can grant that the personal level marks a unique domain of representation; we could even insist that ultimately, from some privileged metaphysical standpoint, only those representations involving conscious agents are *real* representations. This does not preclude there being an epistemic role for representation in scientific explanations pitched at the subpersonal level (see **chapter 1**). In the present case, talk of representation consumption usefully demarcates a significant type of functional contribution that a mechanism makes toward the capacity of a cognitive system when that capacity causally depends on a mechanism that meets the requirement for S-representation.

The above example alone is unlikely to sway those antagonistic towards the possibility of naturalising content at the subpersonal level. Unlike genuine representation, the

thought goes, NavBot's mechanisms do not cause NavBot's capacities *in virtue of* their representational content (Hutto & Myin, 2013, 2017). Thus, NavBot does not really consume a representation. The next chapter will be fully devoted to showing how R-mechanisms possess content by any reasonable standard. First, it will prove useful to visit a proposal from Ramsey (2016): that the 'functional role dimension' and the 'content grounding dimension' are conceptually distinct dimensions of cognitive representation that a complete account must address.

5.1 The functional role vs. content grounding distinction

Chapter 2 presented Ramsey's (2007) JDC as an effective method for adjudicating legitimate ascriptions of cognitive representation. Ramsey (2016) subsequently clarifies what he takes the role of the JDC to be—a sort of 'meta job description' (the distinction is touched upon in Ramsey, 2007 and again in 2018). This provides Ramsey with an opportunity to draw a distinction between the 'functional role dimension' and the 'content grounding dimension' of representation, and in doing so, illustrate the limitations of an account like S-representation. Similar distinctions between a representation's 'functional role' and a representation's 'content grounding' is sometimes implicit in the literature.[11] However, Ramsey's (2016) paper provides the most explicit and thorough exposition. For Ramsey, both the functional role and content grounding dimensions must be addressed to provide a complete account of representation. I agree with Ramsey that both dimensions must be addressed. However, in what follows, I will argue that the S-representation account has the resources to address both, at least when situated within a mechanistic

[11] In a recent example, Thomson and Piccinini define a representation as something that has 'semantic content' *and* has 'an appropriate functional role', where that functional role is to 'as serve as a "stand in" for X so as to guide behavior with respect to X' (2018, p. 193). Taken at face value, this gives the impression that serving as a stand-in and having content come apart.

framework. The 'mechanistic account of content' that I will come to defend suggests that the 'representational target' and 'representational content' of a token R-mechanism fall out of facts pertaining to the context in which that R-mechanism serves as an S-representation (see **chapter 5**). This, it seems to me, is sufficient for addressing the content grounding dimension. As a consequence, we avoid needing to supplement the S-representation account with, say, a traditional causal-historical theory of content, such as teleosemantics.[12] The following discussion will build the foundations for this view.

According to Ramsey's analysis, the 'functional role dimension' concerns what it is for something to function as a representation, whilst the 'content grounding dimension' concerns the conditions for determining a given representation's content; what makes a representation about *x* and not *y*. Ramsey writes of the functional role dimension:

> Here we are talking about those conditions and features of a state or structure that give rise to its having a representational function (in the teleological sense). Many have argued that representation is a functional kind and I believe this assessment is correct. (2016, p. 4)

Ramsey writes of the content grounding dimension:

> Here, rather than explaining a certain type of role, we are interested in a certain type of relation; namely, the content relation that exists between a representation and its intentional object. Our question is, what conditions or properties or relations make a representation about what it is about? (*ibid.*, p. 4)

[12]This attitude towards Ramsey's distinction is paralleled in the next chapter. There I differentiate between two 'problems of content'. The 'hard problem of content' and the 'content determination problem' bear a correspondence to the functional role dimension and content grounding dimension respectively. Though I think these two problems are conceptually independent and both must be addressed, I also think that a solution to the content determination problem falls out of our best response to the hard problem of content.

To help illustrate the intended distinction, observe that one could know that something functions as a representation (for example, a map), without knowing what it represents (for example, what it is a map of). At the same time, one could know something is a representation, without knowing what functional properties make it a representation. Ramsey writes,

> [O]ne can learn that something is a representational device of some sort, and that it has a specific content by virtue of certain causal links, but remain uncertain about how it is actually used by the cognitive system it serves. (*ibid.,* p. 6)

Ramsey concludes that a complete account must answer two questions:

> 1) what makes it the case that a physical state or structure (such as a neurological state or structure) functions as a representational state (and not something else)? and, 2) what makes it the case that something functioning as a representational state has the content it does (and not some other content)? (*ibid*., p. 5)

All this leads Ramsey to suggest that a 'map/model notion' (*ibid.,* p. 11) addresses the former but requires a supplementary theory of content grounding to address the latter. I take it that the map/model notion encompasses the S-representation account.

Ramsey's distinction between the functional role and content grounding dimension of representation is useful for at least two reasons. Firstly, it reinforces the need to specify the conditions under which a cognitive entity functions as a representation, as the S-representation account does. I agree with Ramsey that 'in various accounts of mental representation, the functional role of representing is more or less blurred together with the issue of content, or, alternatively, ignored altogether' (*ibid.,* p. 5). Secondly, the distinction highlights that a complete account of cognitive representation must address

how to think about the content of a token representation. It is right that merely specifying generic functional criteria, as the S-representation account does, will not tell us how a token representation has its content determined. Nonetheless, there are two related considerations that should heighten our caution when affirming the distinction between the functional role and content grounding dimensions.

The first consideration to keep in mind is that, as presented, Ramsey's distinction potentially obscures another useful distinction: between 'representational content' and 'representational target'. Ramsey sometimes appears to treat these synonymously:

> [W]e have seen that while maps and models may provide a good sense of how something functions as a representation, there is nevertheless a problem of content indeterminacy. How do we specify the target of the broader map/model structure?' (2016, p. 9)

The value of the target/content distinction will depend on one's method for addressing the content grounding dimension. However, it is crucial to the mechanistic account of content that I will present in **chapter 5**. Very roughly, a representation's content concerns how it presents the world to its consumer, whereas a representation's target concerns how the world is (think of the difference between the way a map depicts the topology of the Himalayas and the actual topology of the Himalayas). Successful representation occurs when there is sufficient alignment between content and target (for related discussion see Cummins, 1996).

Importantly, terms sometimes used to describe content, such as 'intentional object' (Ramsey, 2016, p. 4), are somewhat ambiguous between target and content as just defined. Furthermore, some of the entities that Ramsey (2016) identifies as a

representation's content might sometimes be better described as a representation's target according to the mechanistic account of content. Also drawing on the example of cognitive maps, Ramsey writes,

> On the one hand, neurons in the hippocampus are supposed to comprise a map of the environment. That suggests the map/model notion of representation is in play [...] On the other hand, neurons are described as representing places in the environment because their activation levels co-vary with proximity to these locations. That suggests the causal/informational notion of representation is at work; that neurons are representations because their activity is strongly correlated with environmental cues. So which is it? (*ibid.*, p. 11)

Ramsey contends that both are at work: the cells function as elements in a map, but their contents are determined by the way their activation nomically depends on proximity to locations in the organism's environment. Of course, the selective activation of neurons in response to stimuli reveals something vital about how those neurons serve the system. I agree with Ramsey that reliable activation in proximity to places in the environment is evidence that the neurons represent those places. However, in allowing this, we must be careful to disambiguate target and content. According to the mechanistic account of content, whilst the map targets the organism's environment, the map's content may sometimes fail to refer to the actual environment, instead referring to a possible environment that would need to be actual for the mechanism to generate behavioural success. This distinction between content and target might strike some as definitional nit-picking, but it plays a significant part in allowing for misrepresentation (see **chapter 5**).

The second consideration I wish to raise with regards to the functional role vs. content grounding distinction concerns what theory of content determination it primes us for. One might think that if the functional role of representation comes apart from its content, then an account like S-representation will need to be supplemented with a traditional causal-

historical theory of content, such as teleosemantics or a Dretskean causal-informational theory (Ramsey, 2016, 2018, mentions both).[13] However, such theories are not the only options on the table, and if we are to distinguish between function role and content grounding then we must ward ourselves against thinking that the distinction entails the necessity of a causal-historical theory, especially if we wish to ensure that S-representation is consonant with mechanistic explanation (see **chapter 2** and **chapter 5** for related problems with causal-historical theories of content). To help understand this, it is worth considering Ramsey's dalliance with a more controversial thesis.

I take Ramsey's (2016) chief concern to be establishing that representational function and representational content are conceptually independent. However, in response to Neander's (2009) dismissal of the idea that something could count as a representation without representing anything, Ramsey briefly raises the idea that it might be possible (in principle) for function and content to be physically independent—in particular, for a creature to possess states that function as representations without possessing any representational content (2016, pp. 6-7). Regardless of how serious Ramsey considers this possibility, it is worth examining because it helps to highlight diverging attitudes toward content determination.

To illustrate his point, Ramsey appeals to Davidson's (1987) Swampman thought experiment. In brief, Davidson imagines a freak lightning strike hitting a swamp comprised of a primordial-soup of elemental ingredients. The energy from the blast

[13] I do not mean to suggest that Ramsey is committed to addressing the content dimension with a causal-historical theory of content. Indeed, elsewhere he mentions alternative theories of content that appeal to the way a representation is used, similar to the account I defend in **chapter 5** (Ramsey, 2007; see also Ramsey, 2016, p. 9, footnote 6). I merely intend to highlight that, as presented, the functional role/content grounding dimension distinction invites the possibility of combining the S-representation account with causal-historical theories, and that we ought to be wary of such a move.

happens to break and bond the alchemical broth in such a way that it spawns forth an uncanny creature. This being is physically identical and indistinguishable from an ordinary person. Because this 'Swampman' is physically identical to an ordinary person, its internal states are physically identical to an ordinary person's. From this starting point, we are supposed to consider the internal vehicles of Swampman as indistinguishable from an ordinary person's. Ramsey points out that 'many have claimed to hold the intuition that such a being would have internal states that are functionally similar to normal representations, but that would nonetheless lack content' (2016, pp. 6-7). This supports the possibility of a mechanism that meets the functional criteria for S-representation but lacks representational content.

Before passing judgement on whether Swampman's states have content, consider that the case is softened if we allow that Swampman has internal states that are merely 'functionally similar' and not identical to an ordinary person's (Ramsey, 2016, p. 6). Swampman does not have any etiological functions; thus, if Swampman's internal states have functions at all, presumably they have functions in the causal-role sense frequently invoked in mechanistic explanations (see **chapter 2**).[14] And yet, Swampman's states have identical causal roles to an ordinary person's states. Thus, Swampman's states are functionally identical to an ordinary person's states. In response, one might insist that Swampman's states do not possess content and that content makes a causal difference; but that would not support the proposition that the conditions for functioning as a representation could be met without meeting the conditions for representational content. For the thought experiment to hold water we must consider the following possibility: can

[14] A Dretskean causal-historical theory of content somewhat complicates matters. According to Dretske's learning based approach, Swampman would not begin with representational content but could gain representational content once he 'gets going' and acquires a learning history. For a similar point, see Shea (2018, pp. 22-23).

Swampman possess states that function identically to that of an ordinary person's without possessing representational content?

Chapter 2 suggested that a successful account of cognitive representation ought to provide representational content with causal relevance; after all, some eliminativists claim correctness conditions at the subpersonal level contribute nothing to our explanation of a system's behaviour to support their cull of subpersonal representation (for example, see Hutto & Myin, 2013, 2018). A plausible way to do this is to show that content is relevant to the causal role of a cognitive mechanism. To concede that Swampman's cognitive mechanisms function as representations but lack content (and are therefore not full-blown representations), is to concede that content is not causally relevant to our explanation of Swampman's behaviour. In turn, this undermines the idea that content is of causal relevance in explanations of cognition. And yet, content does appear to be of causal relevance in our explanations of Swampman's behaviour. This is because the degree of correspondence between Swampman's 'R-mechanism' and some state-of-affairs causally determines the success/failure of Swampman's behaviour (just as it does for a non-Swamp person), and this is key to the causal relevance of content.

Consider another case closer to home. Imagine a cognitive neuroscientist is studying the brain of a rat in order to understand its capacity to navigate a maze and locate a reward. Let's name our subject 'Original Rat'. During her observations, our scientist concludes that Original Rat's hippocampus contains internal cognitive maps that possess the functional properties required to be an R-mechanism. These cognitive maps structurally correspond to the animal's environment and allow it to successfully navigate its environment. Suddenly, a freak lightning strike results in a peculiar explosion in the

experimental physics lab next door causing the generation of a physically identical being. Let's designate this unworldly doppelgänger 'Swamp Rat'. Let's further imagine that Original Rat is destroyed during the event, only to be instantaneously replaced by Swamp Rat in precisely the same location. Distracted by the incident, the experimenter is none the wiser and assumes the creature in front of her to be Original Rat.

Swamp Rat has identical behaviours to Original Rat—both 'narrowly', in terms of its bodily properties, and 'broadly', in terms of its interaction with the environment. Swamp Rat has a physically indistinguishable anatomy and physiology to Original Rat, including an indistinguishable brain with the same cellular organisation and activation patterns, and is located in the very same environment. As such, Swamp Rat has identical navigational capacities to Original Rat. The scientist continues to identify Swamp Rat's 'R-mechanisms' as representations of the creature's environment and, as far as the scientist is concerned, the relative correspondence between these mechanisms and the creature's environment continues to play the same explanatory role. The intuition I wish to pump is that content is as relevant to explaining Swamp Rat's capacities as it is to explaining Original Rat's capacities. This is because there is an identical causal role being played by a mechanism that meets the requirements for R-mechanism in determining the success or failure of a system's capacity. If correct, this indicates that a causal-historical theory of content is inappropriate and unnecessary when it comes to R-mechanisms.[15] We will return to Swamp Rat in **chapter 5**.

[15] Swamp Rat's origins are as preposterous as Swampman's. Real cognitive mechanisms result from mundane biological and learning histories. My aim is not to deny the importance of such histories in creating mechanisms, only to put pressure on the idea that content determination is best understood in terms of etiological factors instead of what a mechanism presently contributes to a system's capacities.

In summary, the functional role vs. content grounding distinction is useful for highlighting what a complete account of representation must address. Nevertheless, we must be careful to appreciate the potential for a further distinction between a representation's content and a representation's target, as some theories distinguish the two. Furthermore, we need not assume that the S-representation account requires a traditional causal-historical theory of content, as such theories seem problematic for a mechanistic approach.

5.2 Representation tokens and contextual functions

We are almost ready to pull our threads together and propose a complete account of content for R-mechanisms. To close our discussion of the functional role vs. content grounding distinction, and further lay the foundations for the account of content that I will defend, I wish to draw attention to a contrast between properties of a 'representation type' on the one hand and properties of a 'representation token' on the other. Representation types are a class of mechanism. The S-representation account can be characterised as articulating functional criteria for a representation type. Representation tokens are individual members of a representation type. Representation tokens realise the generic properties of their representation type within a wider mechanistic context. Recognising this wider mechanistic context is important because whilst the S-representation account cannot itself tell us what a token R-mechanism represents—because it only specifies generic functional criteria—it may be that the token semantic properties of a given R-mechanism become transparent when we observe the broader circumstances under which it serves as an S-representation. In this way, the content dimension is addressed by observing the exploitation of a given S-representation in context. This further supports the notion that the S-representation account does not require a supplementary causal-historical theory of content. The remainder of this section

will unpack the importance of context further, before pursuing my theory of content more fully in the following chapter.

Token R-mechanisms exist within a wider mechanistic context in two senses. Firstly, a token R-mechanism is likely to be embedded within a higher-level mechanism; for example, a cognitive map forms part of a greater spatial-navigation mechanism. Secondly, a given R-mechanism serves as an S-representation relative to a cognitive capacity that is embedded within a task environment; for example, a rat's capacity to locate food might take place within a laboratory maze. Thus, token R-mechanisms are posited within the context of a higher-level mechanism and a task environment. Let's examine these contextual factors in turn.

There is a precedent in the literature for emphasising the contextual nature of many mechanisms. **Chapter 2** introduced the idea of 'contextual function'. A contextual function is part of a contextual explanation; one that considers the broader environment of a mechanism. This is necessary when situating a mechanism within a higher-level mechanism (Craver, 2001; Craver 2013). The underlying idea is that when a mechanism plays a causal role in a higher-level mechanism, function attribution must reference entities external to the mechanism itself. As Craver summarises,

> Contextual explanations are characteristically outward looking and upward looking. They are outward looking because they refer to components outside of the item to be explained and they are upward looking because they contextualize that item within the behaviors of a higher-level mechanism. (2013, p. 153)

For example, to describe the heart as functioning to pump blood is to describe the heart contextually. This is because describing the heart as functioning to pump blood locates

its pumping activity within a higher-level mechanism—a wider set of spatiotemporal components comprising the cardiovascular system. Following Craver once more, 'A description of the heart's mechanistic role function is contextual to the extent that it makes explicit reference to objects other than the heart itself and its parts.' (*ibid.*, p. 152). R-mechanisms are also likely to be located within higher levels that underlie a capacity. Take rat navigation. A cognitive map will be connected to perceptual processing and motor mechanisms that allow a rat to sense its surroundings and physically move. Together, these mechanisms form components within a higher-level spatial-navigation mechanism. In this way, R-mechanisms play their S-representational role within the context of a higher-level mechanism.

The contextual nature of cognitive mechanisms does not stop at the bounds of the higher-level mechanisms that comprise the cognitive system. A mechanism's environment provides background conditions and influences on the mechanism that are frequently important for understanding the mechanism's part in an *explanandum*. This is especially true of many biological and cognitive mechanisms because the capacities of cognitive systems typically concern the system's interaction with its environment. As Bechtel writes, 'In biological systems, even the behavior of the parts themselves is often affected by the organization and environment in which they function and learning about such behavior requires studying the part in such a context' (2009, p. 560). Cognitive science seeks eventually to explain the capacities of whole cognitive systems. Such capacities concern a system's engagement with its environment, implying some relation between the parts of the system and features of its task environment. Indeed, cognitive capacities as diverse as route planning, facial recognition, mind reading, object categorisation, self-relative position tracking and so on, imply entities external to the system itself but located

within the environment in which the capacity takes place. In this way, R-mechanisms play their S-representational role within the context of a task environment.

The context of a token R-mechanism (the higher-level mechanism and task environment it is situated within) cannot be discerned merely by examining the generic properties of S-representation. However, in an important sense, the higher-level mechanism and task environment that contextualise a token R-mechanism are not independent of its S-representational role. Rather, such context arises from the individual conditions under which a mechanism serves as an S-representation. The same can be said of many mechanisms. The fact that the heart pumps blood is determined by the conditions under which it plays its pumping role, that is, within the higher-level cardiovascular system—something we could not determine only by knowing that a mechanism functions as a pump—and the way a receptor neuron responds to edges in the environment is determined the conditions under which it plays its receptor role—something we could not determine only by knowing that a mechanism functions as a receptor. Distinguishing between the generic properties of a representation type (i.e., S-representation) and how a token mechanism realises the functional criteria for that representation type (e.g., a token cognitive map) helps to frame Ramsey's (2016) observation that one can know how something functions as a representation (for example, as an S-representation) without knowing what it represents. The S-representation account is an account of the generic functional properties belonging to a type of mechanism.

The contextual properties of a token R-mechanism inform the particular character of its functional role, adding detail beyond the generic properties shared by all R-mechanisms. In closing, I suggest that this wider mechanistic context helps us to understand the

individual semantic properties of a token representation. Essentially, what a given R-mechanism represents—strictly speaking, what I label its 'target' (see **section 5.1** above; **chapter 5**)—can be understood by consideration of its causal role in context. For example, where a cognitive map contributes to a rat's capacity to navigate, the function of the mechanism is to stand-in for its present environment. This is because it is the rat's present environment, and not anything else, that the token R-mechanism must structurally correspond to for the capacity that causally-depends on that mechanism to succeed. As we shall see, the content of a token R-mechanism can be seen to fall out of its ability to succeed or fail to match its target. To this extent, the S-representation account supplies the resources to address the content grounding dimension.

6. Conclusion

The preceding two chapters undermined many of the traditional reasons for assuming that representation plays a role in explanations of cognition. I argued that some notions of cognitive representation did not identify theoretical entities with the function to represent and that computational explanations of cognition were not necessarily representational explanations of cognition. This chapter began the constructive part of this thesis. By drawing an analogy to a type of ordinary representation that includes cartographic maps and scientific models, the S-representation account provides a set of functional criteria under which a cognitive entity counts as standing-in for something else on behalf of a cognitive system. Moreover, the S-representation account accords with a more general mechanistic account of cognitive representation. Synthesising these two accounts lays the foundation for a well-defined, empirically-driven notion of representation.

The next chapter continues the constructive project by returning to a theme established in **chapter 1** and touched on repeatedly throughout the thesis: the role of representational content at the subpersonal level. By drawing on the functional properties of S-representation within a mechanistic framework, I will show how the representationalist can provide a naturalistically respectable account of content at the subpersonal level.

Chapter 5

The Two Problems of Content[1]

1. Introduction

The S-representation account promises to provide cognitive representation with an explanatory role that accords with the increasingly popular mechanistic understanding of explanation in cognitive science. Hard-line anti-representationalists are unlikely to be swayed, remaining suspicious of so-called representational mechanisms (R-mechanisms) and their ability to bear genuine representational content. In this final chapter, I build on the view that began to emerge in **chapter 4** and tackle the issue of representational content head on. Though I will concur with some existing trends in the literature, I will elaborate and adjust where necessary, and show in detail how we can provide a naturalistically respectable understanding of content that resists the eliminativist's strongest objections.

There are two overlapping but distinct 'problems of content' implicit within the literature, a distinction that subsequent discussion should remain sensitive to. The first problem concerns the justification for attributing representational content at the subpersonal level (i.e., what makes a subpersonal mechanism possess content in the first place?). I label this the 'hard problem of content' (HPC), following Hutto & Myin (2013, 2017). The second problem concerns the conditions that determine a token representation's particular content (i.e., what makes a representation about *x* and not *y*?). I call this the 'content

[1] Portions of this chapter appear in Lee (forthcoming b).

determination problem'. This chapter argues that the notion of an R-mechanism provides the resources to remedy both problems.[2]

I will argue that the HPC dissolves when we acknowledge the role correctness conditions play in accurately describing the functional character of an R-mechanism, given the requirement of structural correspondence between an R-mechanism and some state of affairs for the realisation of a cognitive capacity. This mirrors the causal relevance of correctness conditions for ordinary S-representation. Building on this response to the HPC, I further argue that the content of a token R-mechanism refers to the state of affairs that would need to actual for it to realise a cognitive capacity and generate behavioural success. The degree of overlap between an R-mechanism's content and an R-mechanism's 'target' underwrites its semantic evaluability, and engenders the possibility of misrepresentation. I label this view the 'mechanistic account of content' (MAC). As I frame it, the MAC is not an independent theory of content determination *per se*. Rather, the MAC grounds content determination in the features of an R-mechanism already implicit within the S-representation account. It does this by framing content determination in terms of how a token R-mechanism plays its S-representational role within the context of realising a cognitive capacity embedded within a task environment. These responses to the two problems of content are mutually reinforcing. They both rely on (and support the adequacy of) R-mechanisms and combined yield a thorough account of subpersonal content.

[2] Ramsey's (2016) 'functional role' and 'content grounding' distinction can be recast in light of these two problems (see **chapter 4**). The HPC concerns the functional role dimension and the content determination problem concerns the content grounding dimension. Mirroring Ramsey's analysis, I think the traditional debate has concentrated on the latter at the expense of the former. This is unfortunate because our response to the HPC can inform our response to the content determination problem, as we shall soon see.

The chapter proceeds as follows. **Section 2** returns to examine the concept of 'representational content' in greater detail, outlines the HPC and demonstrates how the S-representation account surmounts this first problem of content. **Section 3** examines the content determination problem. I will briefly discuss the possibility of pursuing a 'hybrid view' (first touched upon in **chapter 4**), that combines the S-representation account with a traditional causal-historical theory of content such as teleosemantics. I close by defending the MAC as an alternative to traditional causal-historical theories.

2.1 What (exactly) is representational content?

Previous chapters suggested that representational content must be of causal relevance to a consumer of any genuine representation. **Chapter 4** then claimed that the S-representation account delivers adequate criteria for a cognitive mechanism to play a genuinely representational role. Those sceptical of cognitive representation will naturally demand to know precisely how so-called R-mechanisms play a functional role such that their carrying content is causally relevant to their supposed consumer. At this stage, it is worth pausing to recall and refine our understanding of representational content.

We have seen many times already that content relates to the 'semantic properties' associated with representation. As a useful starting point, Brown writes, 'A state with content is a state that *represents* some part or aspect of the world; its content is the way it represents the world as being' (2014. Online, section 1. Original emphasis). Most importantly, a representation's content refers to some state of affairs ('the way it represents the world') such that the representation is 'semantically evaluable'. A representation is semantically evaluable in so far as it bears correctness conditions, i.e.,

in so far as there are states of affairs under which the representation counts as correct or incorrect.

Rescorla reflects this orthodoxy, summarising that a mental state has content if it is 'associated with *veridicality conditions*: conditions for veridical representation of reality' (2016, p. 17. Original emphasis). What all diverging theories of content share, according to Rescorla, is the assumption that 'many important mental states are evaluable as veridical or non-veridical' (*ibid*., p. 17). Though I take Rescorla's definition to be essentially correct, we should be careful not to assume that representational content requires a mental state to be associated with a literally true state of affairs. After all, many ordinary representations correctly represent counter-factual histories, imaginary worlds or possible futures—though we need not consider these 'veridical representations of reality'. For example, I might write an accurate 'what if' history considering world events if JFK had not been assassinated, draw a faithful map of Middle-Earth, or construct a precise model of future climate change. Following earlier chapters, I will settle for the term 'correctness conditions'. This term allows that successful representation is not necessarily strictly veridical. Correctness conditions also permit that a representation's semantic evaluation (its 'correctness value') is not necessarily binary (in other words, a representation can be more or less correct), and is general enough to encompass truth, accuracy or other preferences for defining semantic measures of success. In summary, something possesses representational content just in case it is associated with correctness conditions, that is, conditions for the correct representation of some state of affairs.

Recall that the need to account for representational content at the subpersonal level is tied to the need to address the possibility of misrepresentation (see **chapter 2**). If something

possesses correctness conditions, then it is possible for it to be incorrect. As with ordinary representation, incorrectness is supposed to be causally relevant to cognitive representation. For example, a mountaineer's map is capable of misrepresenting features of the Himalayas such that this misrepresentation causally affects the outcome of the mountaineer's actions. The challenge is to show how misrepresentation could be causally relevant in the case of a subpersonal cognitive representation in a manner analogous to ordinary representation. Shea puts the problem in the following way:

> Whether a representation is correct or incorrect depends on factors outside the organism, which seem to make no difference to how the representation is processed within the organism (e.g. to how activity of some neurons causes activity of others). Yet its truth or falsity, correctness or incorrectness, is supposed to make a crucial explanatory difference. (2018, p. 10)

In summary, any successful account of content for R-mechanisms must address the conditions under which an R-mechanism counts as correct or incorrect in a way that makes its correctness or incorrectness causally relevant to a cognitive system. If we cannot do this, then we have not accounted for content at the subpersonal level.

It will prove instructive to anticipate some of the conclusions about content in R-mechanisms drawn below by returning to what Menary labels the 'Peircean principle' (2007, p. 95), touched upon in **chapter 1**. Peirce's original analysis of representation identified three necessary components: a representational vehicle (or 'sign'), the represented object, and an interpreter or consumer that exploits some salient property of the vehicle (for related discussion, see von Eckardt, 1993; Ramsey, 2007). Under the present account, the vehicle is an R-mechanism. The consumer is the wider containing cognitive system whose behavioural success causally depends on the R-mechanism. Drawing on the previous chapter, the position I outline below frames the represented

object in terms of an R-mechanism's 'representational target'. A representational target is that which an R-mechanism must structurally correspond to for it to fulfil its function and realise the capacity of the consuming system (*ceteris paribus*). Given this mapping of terms within the R-mechanism story to the Peircean triadic analysis of representation, how should we understand content determination in R-mechanisms? The answer is that the content of a token R-mechanism refers to the state of affairs that would need to obtain for the R-mechanism, such that it is, to realise the capacity and generate behavioural success. Misrepresentation is made possible by the disconnect between content and target. Before properly attending to this understanding of content determination, however, let's return to the idea that there can be no correctness conditions at the subpersonal level.

2.2 The hard problem of content

In **chapter 1** we saw that Hutto & Myin (2013, 2017) argue that there can be no subpersonal cognitive representation because representation requires content, and there is no such thing as content at the subpersonal level. I put pressure on an a priori interpretation of this argument that assumes the notion of representation does not apply at the subpersonal level as a matter of respecting strict category boundaries. However, the onus remains on an account of subpersonal representation to demonstrate how content is afforded a legitimate role within a theory of cognition. With the S-representation account now on the table, let's re-examine Hutto & Myin's argument.

Hutto & Myin propose that any form of representationalism aimed at the level of 'basic cognition' cannot explain how a so-called representation could 'be about' something in a 'non-spooky' way (2013, p. 66). To account for content, the thought goes, we must be able to provide a naturalistically respectable explanation of how a subpersonal cognitive

entity could possess correctness conditions. Hutto & Myin's starting assumption has been shared throughout this thesis: that an effective account of representation cannot take representational content as a given. They write,

> Naturalistic theories with explanatory ambitions cannot simply help themselves to the notion of information-as-content, since that would be to presuppose rather than explain the existence of semantic or contentful properties. (2013, p. 67)

If we accept that representation is a functional kind, then justified attributions of representational content are predicated on content playing a part in the functional role (*qua* representation) of a subpersonal cognitive mechanism. The question is whether it is possible to provide an account which offers content that part. Hutto & Myin's answer is a resolute no. My answer is a resounding yes.

2.3 What *kind* of content?

Recall that Hutto & Myin's argument rests on the observation that many so-called representations are best understood in purely 'informational' terms, and that information and representational content are conceptually distinct (2013, p. 67). Hutto & Myin hold that the mental states and processes of basic cognition, which representationalists treat as contentful, are better understood as information bearing. This is because their essential explanatory role is cashed out in terms of covariance (not 'standing-in for' in the sense of bearing correctness conditions). This is 'information-as-covariance' and is to be understood in the sense that *x* carries information about *y,* just in case *x* lawfully or reliably covaries with *y*. In other words, information-as-covariance is 'natural information'.

Hutto & Myin hold that covariance is insufficient to ground representational content (2013, p. 67). Let's examine this in a little more detail. Hutto & Myin believe that there is no property of covariation that, by itself, can secure correctness conditions. In their terms, something doesn't 'say' or 'mean' anything 'just in virtue of instantiating covariance relations' (*ibid*., p. 67). Entities that merely covary with some feature fail to capture the unique explanatory potential that paradigmatic representations play; to covary with some state-of-affairs, by itself anyway, is not to stand-in for that state-of-affairs. It is from this starting point that Hutto & Myin derive their 'Covariance Doesn't Constitute Content Principle'. They summarize their view by writing,

> [I]f we stick to the notion of information-as-covariance there are no grounds for thinking that the world, standing apart from agentive systems, contains anything that could be called informational *content*. If covariation is assumed to be the only worldly source of informational content, then, in light of the Covariance Doesn't Constitute Content Principle, we have, as yet, no explanation for the natural occurrence of informational content in the world. (2013, p. 71. Original emphasis.)

In keeping with earlier arguments, I agree with Hutto & Myin's principle: covariation is indeed insufficient for grounding representational content in any significant, theoretically interesting sense. And yet, this does not show that there is no alternative notion of information at the subpersonal level from which to source representational content.

The 'covariance doesn't constitute content principle' can be understood as an unsurprising consequence of the well-established distinction between two notions of information. **Chapter 2** suggested that we ought to (a) distinguish between two notions of information in cognitive science, natural and non-natural, that map onto the Gricean notions of natural and non-natural meaning, and (b) consider that representational content (as an informational notion) is equivalent to non-natural information. The content sceptic

should be satisfied with the distinction between the two kinds of information as it helps articulate their worry: natural information plays an explanatory role in cognitive science, but there is no justification for positing non-natural information. Hutto & Myin hold that genuine non-natural information can only occur within and between encultured agents, for it is only within this intersubjective (personal level) arena that the norms required for genuine content arise. The S-representation account challenges this view.

2.4 The hard problem and propositional attitudes

To better appreciate the potency of the S-representation account as a response to the HPC, it is worth briefly exploring an unnecessary assumption about the type of representation that cognitive science is committed to. Though Hutto & Myin intend the HPC to threaten all forms of representation at the level of 'basic cognition', at times they talk as if the ascription of 'propositional content', by which they mean semantically evaluable content, is tied-up with ascribing propositional attitudes. At one stage, they affirm their agreement with eliminative materialism that we ought to avoid projecting the properties and structure of everyday language onto our explanations of basic cognition (2013, p. 13). They write, 'Like classic eliminativism, REC denies that basic mentality and cognition should be modelled in terms of propositional attitudes' (*ibid*. p. 13). They go on to say,

> [R]EC does not claim that propositional attitude explanations are never appropriate. REC holds that some organisms have more than one way of getting around cognitively, and that some organisms—language users, at least—are capable of genuinely contentful, representational modes of thinking and reasoning. (*ibid.*, pp. 13-14)

Read at face value, this description implies REC's opposition to a form of representationalism committed to propositional attitudes, which is to say, agents bearing belief/desire style relations to inner syntactic states. However, such an assumption fails

to capture the full sweep of representation types posited in scientific practice. In response, the anti-representationalist could insist that only propositional attitude representations serve as a foundation for the legitimate ascription of content. Therefore, in the absence of such representations, there is no justification for ascribing content.[3] However, I see no reason to assume this. After all, there are plentiful examples of ordinary iconic, abstract and other content-bearing representations (for example, maps, photographs, portraits etc.) on the basis of which one might attempt to model cognitive representation.

The takeaway message is that the fate of representation in cognitive science does not hinge on the fate of propositional attitudes. This is vital because one may reject the relevance of propositional attitudes for 'basic cognition' whilst holding onto the value of cognitive representation in an alternative guise. One may even embrace core aspects of Hutto & Myin's philosophy—for example, conceding that propositional attitudes are a phylogenetically and ontogenetically advanced development heralded by the advent of sociocultural practices (Hutto, 2008; Hutto & Myin, 2013; 2017), whilst also maintaining that subpersonal representation is to be found in another form. In summary, the legitimate ascription of content at the subpersonal level does not hinge on the legitimate ascription of propositional attitudes.

[3] We can thus imagine a form of anti-represenationalism which combines traditional eliminative materialism with a conceptual analysis that ties all content ascription to propositional attitude ascriptions.

2.5 The hard problem and S-representation

Hutto & Myin are correct to question versions of representationalism that help themselves to semantic properties rather than explaining their presence and relevance within a theory of cognition. Nevertheless, I previously suggested that we leave open the possibility that some account of cognitive representation could, in principle, provide adequate and 'non-spooky' justification for the attribution of semantic properties at the subpersonal level (see **chapter 1**). We are now in a better position to appreciate that the notion of R-mechanism shows how this is possible. In short, R-mechanisms systematically determine the outcome of a capacity or task (of the sort studied by cognitive science) based on the degree of structural correspondence between a mechanism and some state of affairs (what I call the 'representational target'). This, I contend, satisfies any reasonable demands on what it takes for a subpersonal entity to count as bearing correctness conditions.

Our starting premise is that there is no good a priori reason to suppose that non-natural information must necessarily derive from personal level conventions (see **chapter 1**). In Piccinini & Scarantino's words, 'What matters for something to bear nonnatural information is that, somehow, it stands for something else relative to a signal recipient' (2011, p. 24). In my words, what matters for something to bear representational content is that, somehow, it stands-in for something else relative to a consuming system (in a way that affects its behaviour in some significant respect). By meeting the functional properties for S-representation, R-mechanisms do just that.

To appreciate how R-mechanisms achieve this, it is crucial to recall that where an R-mechanism plays a causal role in a cognitive capacity, the occurrence of that cognitive capacity causally depends on the degree of structural correspondence between the

mechanism and some target state of affairs (see **chapter 4** and **section 3** below). Because of this, it is natural to describe an R-mechanism as 'standing-in for' the state of affairs that it is required to correspond to, and in turn, to describe it as possessing correctness conditions that are met if the required correspondence occurs. For example, imagine that a rat's capacity to navigate from a starting location to a reward location within a maze involves an R-mechanism in the form of a cognitive map. In order for the rat to navigate to the reward location, for its capacity to occur, the R-mechanism must structurally correspond to the maze (to a sufficient degree). Furthermore, we can imagine that if the rat fails to locate the reward due to insufficient structural correspondence, then its cognitive map will update. This will increase the probability that the required structural correspondence now obtains, and the rat will subsequently succeed in navigating (more on system-detectable error below). Thus, the structural correspondence between cognitive map and maze is causally relevant to the rat's behaviour. Given this, the cognitive map serves a causal role akin to an ordinary map, standing-in for the maze on behalf of the rat. In turn, the map possesses correctness conditions that correspond to the conditions under which it succeeds or fails to stand-in for its target: the map counts as correct when its structure sufficiently matches the maze's structure, and incorrect when it does not. In summary, once we acknowledge the relationship between an R-mechanism, a target and behavioural success, attributing correctness conditions to an R-mechanism becomes informative—tracking those conditions under which it succeeds or fails to play its representation-like role.

It is worth emphasising that this emergence of correctness conditions from the interplay between vehicle, target and behavioural success is analogous to the way ordinary S-representations suggest correctness conditions. An ordinary map's structure must

correspond to, say, the topology of the Himalayas for a mountaineer to successfully conduct a hike. In turn, the map is said to possess correctness conditions that correspond to the states of affairs under which it succeeds or fails to stand-in for its target: the map counts as correct when it sufficiently mirrors the topology of the Himalaya's, and incorrect when it does not. In this way, content is justifiably ascribed to R-mechanisms because their functional role is analogous to the very functional role of ordinary S-representations that supplies their correctness conditions.

Recall that the justifiable attribution of correctness conditions is conceptually tethered to accounting for misrepresentation. Where a system's success in realising some capacity causally depends on an R-mechanism, misrepresentation occurs because of a mismatch between the structure of the mechanism and the structure of the state of affairs it must correspond to for the capacity to succeed. For example, the relative success of a rat's capacity to navigate its environment depends on the extent to which its cognitive map corresponds to features of its environment. We can view misrepresentation as a form of malfunction, albeit one that does not depend on selection or learning history: an R-mechanism malfunctions when it fails to cause a cognitive capacity in the here-and-now. The way a cognitive capacity's failure is explained by a mechanism's lack of correspondence to some state of affairs is analogous to the way many everyday failures are explained by an ordinary S-representation's lack of correspondence to some state of affairs. For instance, a mountaineer might fail to trek the Himalayas because portions of her map fail to sufficiently correspond to the structure of the relevant geographical features of the Himalayas. The causal relevance of misrepresentation in R-mechanisms thus parallels the causal relevance of misrepresentation in ordinary S-representations.

A crucial implication of the present analysis is that correctness conditions appropriately capture how a system containing an R-mechanism *works* (for related discussion, see Bartels, 2006). Returning to the quote from Shea (2018) above, it remains true that 'factors outside the organism' make no difference to how the representation is processed within the organism *per se,* but a representation's correctness or incorrectness does play a crucial explanatory role in so far as it captures the efficacy of a mechanism in relation to a cognitive system embedded within a task environment. In the absence of invoking correctness conditions, we lose an informative perspective on the role an R-mechanism plays in causing a capacity to succeed or fail (or something in between). Gładziejewski & Miłkowski make a similar point, arguing that 'similarity' is causally relevant to an S-representation's functional contribution towards behaviour. It is worth quoting the authors in full when they write,

> [A]n S-representation cannot do its job (i.e., enable success) without being structurally similar to the target. Here, the pattern of relations between components of the S-representation plays a crucial role. For example, a map—be it artifactual map or neurally-realized cognitive map—needs to stand in a structural resemblance relation to the terrain if it is to perform its S-representational job; and any figure placed within a map can act as an S-representational surrogate only insofar as it stands in certain relations to other figures or lines on the map. (2017, p. 348)

In short, the degree of correspondence or similarity between an R-mechanism and some target state of affairs is not incidental but critical for determining the outcome of an embedded system's behaviour. Once more, the requirement of structural correspondence ensures correctness conditions are causally relevant at the subpersonal level.

Despite any apparent temptation to think that R-mechanisms parallel the criteria for correctness conditions in ordinary S-representations, Hutto & Myin are explicit in their

rejection of (so-called) cognitive S-representations achieving genuine representation-hood. This is because attributing representational content adds nothing to the explanatory work achieved by attributing structural correspondence. Hutto & Myin write,

> [F]unctional isomorphisms are all that need to be exploited for the purposes of mapping and navigating. Yet it is not at all obvious why the exploitation of such correspondences need entail the existence of representational contents. (2017, p. 159)

Tonneau similarly states that,

> Whenever behavior is explained by appealing to an internal isomorph of the environment, what does the explanatory job is the notion of isomorphism and not that of representation. (2012, p. 342)

I agree that a mechanism which structurally corresponds to 'another state or process does not, by itself, explain why such stand-ins should be thought to instantiate or bear representational content' (Hutto & Myin, p. 2017, p. 158). And yet, to describe the S-representation account in this way—that is, merely in terms of structural correspondence—is to undersell the account. The S-representation account captures the possibility of a *decouplable* mechanism *guiding action* in a way that systematically determines the degree of a system's success in achieving some capacity through 'functional isomorphisms' (*structural correspondence*), reinforced by the role of *system-detectable error*. When we consider all four conditions in play, dismissing cognitive S-representations as mere internal isomorphs of the environment becomes harder to sustain. Once more, I suggest that the conditions for cognitive S-representation mirror those that enable ordinary S-representations to bear content. Cognitive S-representations are no more or less mere internal isomorphs of the environment than cartographic maps are mere external isomorphs of the environment.

Of course, we must concede that it is possible to describe the interaction between a mechanism that meets the criteria of S-representation and a wider containing system without appealing to correctness conditions. However, this itself is an uninteresting observation. After all, it is possible to describe just about anything without appealing to correctness conditions including ordinary S-representations. For example, in principle, we could describe the interaction between a map and a mountaineer in terms of atomic particles and physical forces (for a similar point, see Ramsey, 2007, pp. 33-34). If the issue is whether representational content *can* be exorcised from explanations of cognition, then the anti-representationalist wins by fiat. The more interesting question is whether ascribing correctness conditions contributes anything of explanatory value from a certain level of description (and I think Hutto & Myin agree: see, for example, 2017, chapter 1). In the case of S-representation, this contribution consists in relating the systematic relevance of structural correspondence between a class of action-guiding mechanism and some state of affairs to the success or failure of a cognitive system's capacities.

Ultimately, what matters is that the category 'R-mechanism' individuates a class of potential mechanisms whose functional contribution to the cognitive economy closely resembles the way a class of ordinary representation contributes to the behaviour of agents in everyday life. In fact, even if we were to capitulate to the content sceptic's conviction that R-mechanisms do not count as possessing genuine representational content—because, despite everything that has been said, we dismiss the relationship between R-mechanisms, targets, and cognitive capacities as sufficiently like the relationship between ordinary representations, targets, and an agent's actions—we would still be compelled to mark out mechanisms that meet the functional criteria for S-

representation as an interesting class precisely because their functional role so closely resembles a type of ordinary representation. And yet, it is precisely this close resemblance to a type of ordinary representation that motivates the classification of these mechanisms *as representations*, at least for the explanatory purposes of appropriately describing the distinctive functional character of our theoretical posits—what more could one want?

In closing, I want to emphasise that the system-detectable error criterion reinforces the part that strength of structural correspondence plays in grounding correctness conditions. If a theory of a system's behaviour posits an S-representation with a feedback component, whereby the system adjusts its behaviour based on a mismatch between the structure of an R-mechanism and some state of affairs relevant to a task, then such a mismatch provides further justification for thinking that error, therefore correctness conditions, therefore content, contributes to our understanding of how the mechanism works. Thus, versions of the S-representation account that feature a system-detectable error component present an even stronger case for representational content than those that do not.

At this stage, the content sceptic may concede that the S-representation account provides content with causal relevance in the abstract. Nonetheless, they will insist that we have failed to show how a given representation has its particular content determined. In other words, they will ask how we are supposed to think of a token R-mechanism's particular content—what makes a token R-mechanism about *x* and not *y*? Until we address this, the justifiable attribution of content at the sub-personal level will be threatened.

3.1 Causal-historical theories of content determination

Perhaps the most obvious strategy that proponents of the S-representation account might deploy to address content determination is to opt for a 'hybrid view', to borrow from Ramsey (2016), that combines the S-representation account with a traditional causal-historical theory of content, such as teleosemantics (for example, Millikan, 1989b, 1990). We do not have the space to discuss every possible iteration of such a hybrid view. Instead, I will settle for noting a general worry with this strategy. This worry was anticipated in the previous chapter in the discussion of Ramsey's (2016) distinction between the 'functional role' and 'content grounding' dimensions of representation and concerns the inappropriateness of etiology for fixing content.

The underlying worry is that causal-historical theories place undue attention on the proximal conditions, evolutionary history or other etiological factors that are in some sense responsible for the existence or tokening of a representation. **Chapter 2** argued against purely selectionist accounts of function, suggesting their emphasis on a system's history is of limited value to mechanistic explanation. Consideration of a mechanism's etiology is interesting in its own right. It is also of heuristic value when considering what possible causal roles a mechanism might play. However, mechanistic explanation is principally concerned with how a phenomenon is causally constituted or produced by the organisation of and interaction between spatiotemporal parts and their operations. If cognitive representation is to be understood as a class of mechanism, then we should doubt the adequacy of etiological approaches. Consideration of 'swamp cases' in **chapter 4** reinforced this suspicion, suggesting that the causal relevance of content is separable from a system's history because R-mechanisms are explanatory, *qua* representation, even for swamp creatures. In summary, if representational explanations primarily figure in

mechanistic explanations of system capacities, then we can reasonably question whether the (evolutionary or ontogenetic) history of the mechanism is of central importance to content determination.

In contrast to traditional causal-historical theories, I suggest that we shift our focus away from a representation's origins and towards how a representation figures in the constitution or production of a consuming system's capacities—the very thing that justifies attributing correctness conditions at the subpersonal level in the first place (see **section 2** above). Rather than wedding content determination to a set of causal antecedents, we should locate content within those very structural relations of a representational mechanism that are available for exploitation by a cognitive system. This chimes with a criticism of causal-historical views courtesy of Anderson & Rosenberg, who write,

> [L]et us say, by way of situating our own account of representational content, that we find the various causal approaches too *input focussed,* meaning they give too much importance to the ways in which the environment affects the organism to endow its states with representational meaning (2008, p. 56. Original emphasis.)

The authors add that none of the traditional causal-historical theories, 'give sufficient weight to the full range of what a subject *does* with its representations' (*ibid*. Original emphasis). It is similar intuitions that drive the account of content I defend below. The conditions that cause a representation to exist or token bear a non-necessary relationship to what a representation does for a consuming system. In turn, what a representation does for a consuming system should be central to fixing its content.

As a final note on the hybrid view, it is worth considering 'content pluralism' as a potential complementary position. Traditional debate frames the content determining relation of a token representation as determinate and exclusionary. Hence, theories of content determination are usually taken to be in competition (for example, teleosemantics vs. causal-informational theory). However, according to content pluralism, multiple content-bestowing relations co-exist. To my knowledge, content pluralism has been underexplored within the context of both the S-representation account and representational mechanisms more generally. Future research ought to examine the plausibility of content pluralism.

There are two possible versions of content pluralism. The first version says that different types of representation have their contents determined by different relations (Shea, 2013, presents a view close to this). For instance, one type of representation may have its content determined by selection history whilst another has its content determined by learning history. The second version says that the same token representation may simultaneously bear multiple content determining relations, and thus bear different contents relative to different relations (I briefly explore a related position in relation to the semantic view of computation in Lee, forthcoming a). For instance, the very same representation may have one content in virtue of its selection history and another in virtue of its learning history. Notably, if either version of content pluralism is correct, and my worries about traditional causal-historical theories are misguided, then perhaps the mechanistic account of content determination defended below can co-exist alongside causal-historical accounts, as one part of a larger pluralism about content determination.

3.2 The mechanistic account of content determination

According to the 'mechanistic account of content' (MAC), an R-mechanism's content refers to the state of affairs that the R-mechanism does structurally correspond to and would need to be actual for the capacity in question to be realised and behavioural success to occur. Equally important is an R-mechanism's 'target'. An R-mechanism's target is the actual state of affairs that the R-mechanism must structurally correspond to for the capacity in question to be realised and behavioural success to occur. The degree to which an R-mechanism counts as correct is determined by the degree to which its content and target overlap (see **section 2.5** above).[4] The remainder of this section will elaborate the MAC, beginning with the dual significance of targets and contents.

For the MAC, representational contents require representational targets. The target of a token R-mechanism is fixed by a brute fact about the capacity it serves and the task environment which embeds that capacity. For example, the target of a cognitive map that is responsible for allowing a rat to navigate from a start location to a reward location within a maze is, roughly, the maze. This is because it is the maze that the cognitive map must correspond to for the capacity it serves to be realised. In other words, the maze is the state of affairs that the representation must emulate for the rat to succeed at the task at hand. (see **chapter 4**). As Bechtel again helps to illustrate, for cognitive science,

> The focus is not on the material changes within the mechanism, but rather on identifying more abstractly those functional parts and operations that are organized such that the mechanism can interact appropriately in its environment […] That is, cognitive scientists identify mental operations and

[4] Again, correctness need not be all-or-nothing. There may be degrees of correctness measured by the extent to which a representation's relative structural correspondence to its target causes partial or total behavioural success—for example, when a rat partially navigates towards a reward hidden in a maze before losing its way.

> consider how they contribute to the individual's functioning in its environment. (2008, p. 23)

Of course, a cognitive map may or may not correspond to its actual environment. The MAC states that the content of an R-mechanism refers to the state of affairs that it does correspond to and would need to be actual for the R-mechanism to realise the capacity in question, that is, the state of affairs that would need to be the target for it to generate behavioural success (for related discussion, see Bickhard, 1999; Gładziejewski, 2015). This state of affairs is determined by the structure of the R-mechanism's action-guiding parts and the capacity it serves. To illustrate, imagine that a rat is attempting to locate food within its environment. As it happens, the rat is located within an S-shaped maze. Unfortunately, the rat possesses a cognitive map whose structure corresponds to a T-shaped maze (causing the rat to plan and move as if it was located in a T-shaped maze). In this scenario, the content of the cognitive map refers to a T-shaped maze, that is, the state of affairs that would need to be actual for the R-mechanism to generate behavioural success. Notice that whilst the R-mechanism will structurally correspond to very many things (see **chapter 4**), it is only the T-shaped maze that is relevant to the capacity which the R-mechanism plays a part in. Put otherwise, it is only the T-shaped maze that reflects a relevant counter-factual (more on this below). Understanding both representational contents and representational targets in terms of the correspondence required for the

realisation of a cognitive capacity provides the MAC with a mechanistic footing whilst capturing the importance of a system's wider embeddedness.[5]

The correctness or incorrectness of a token R-mechanism results from the degree of overlap between its content and its target. This is consistent with the intuition that S-representations, whether cognitive or ordinary, succeed when there is an appropriate alignment between 'the way the world is presented' to the consumer, and 'the way the world is'. It is important to stress that whilst an R-mechanism's content is partly determined by its structure once it enters a representational relationship, structure alone carries no semantic significance. It is only in relation to a representational target that a representational content bears any explanatory significance. In this way, representational targets and representational contents are co-defining.

The MAC presents an alternative to many causal-historical theories by relating content determination to the causal relevance of a mechanism for the success or failure of a cognitive capacity in the here-and-now. Correct representation is tethered to a subpersonal entity fulfilling its functional role (as, say, teleosemantics would suggest), but such fulfilment is cashed out in terms of a mechanism's ability to realise some capacity of a containing cognitive system. At the same time, the MAC circumvents the worry that S-representation trivialises content (Sprevak, 2011; Morgan, 2014). The trivialisation

[5] In Lee (forthcoming b) I make a distinction between the 'content' and 'contextual content' of an R-mechanism. The content of an R-mechanism refers to the formal structure of its action-guiding parts. The contextual content of an R-mechanism is equivalent to what I call 'content' in the present chapter. Characterising content in this way is arguably useful because it helps to underscore what is semantically similar about two structurally identical mechanisms playing a role in two different tasks. However, I have since questioned the necessity of this distinction and worry that my previous formulation of content invites accusations of a vehicle/content conflation. The key, I think, is to retain the idea that structure is a determinant of content—hence, the 'semantic continuinty' between two structurally identical mechanisms—without treating the content of a mechanism as referring to formal structure itself. In any case, I here take what I previously called 'contextual content' to be the explanatorily primary notion of content when it comes to R-mechanisms.

concern is built on the assumption that if structural correspondence obtains between an S-representation and some state of affairs, then the content of that S-representation must refer to that state of affairs. The MAC says that this characterisation is misguided. Once again, the content of an S-representation does not refer to anything that it shares structure with. The content of an S-representation refers only to that which it shares structure with and is relevant to the behavioural success of a cognitive system (for related discussion and an alternative appraisal, see Cummins, 1996).

With a preliminary sketch of the MAC complete, it will prove helpful to further illustrate the account using an analogy to ordinary S-representation. Imagine that you visit an IKEA store to buy a new office desk. At the beginning of your journey through the cavernous trove of furniture and household appliances, you pick up a store map. You use this map to guide your way to the appropriate department and locate your purchase. The fact that the structure embodied in the map corresponds to the store explains your ability to navigate. Imagine now that the night before your visit a disgruntled manager (with a penchant for performative irony) has taken out their frustration over recent corporate restructuring by restructuring the layout of the IKEA store itself. Now when you use your map to guide your way toward the office desks you arrive at the standing lamps instead. This time the lack of correspondence between the map and the store explains your failure to navigate. According to the MAC, the content of the map remains the same across both scenarios—because the conditions under which the map would cause you to locate the office desks are identical—whilst the target differs—because the actual location of the office desks diverges. Put otherwise, the correctness conditions of the map remain the same across the two scenarios, whilst the correctness values of the map differ.

Now imagine that you visit an IKEA store in a foreign country and that you bring your old map with you. Despite myriad superficial differences between the foreign store and the original store, it transpires that you are equally capable of navigating the foreign store using your old map. In fact, you can successfully navigate any IKEA store of similar size (in the absence of sabotaging managers). This is because each of these IKEA stores instantiates the same structure to a sufficient approximation. Given the example, one might be tempted to say that the content of the map must refer to any generic IKEA store of a certain size. Indeed, it is no accident that these maps allow for the successful representation of any generic IKEA store given the coordination behind the construction of IKEA stores and the maps used to navigate them. Notice, however, that one could, in principle, successfully navigate *any* environment that instantiates a similar structure but is *not* an IKEA store—say a virtual simulation of an IKEA store, or a hardware store with the same layout (more on this shortly). Keeping successful action front and centre, the MAC suggests that we think of content as referring to whatever state of affairs reflects the possibility of behavioural success, depending on the task in question.

Let's turn now to another example of cognitive representation to help further illustrate the MAC. Imagine that we are explaining a cognitive capacity in terms of a state estimate of the sort posited by Bayesian sensorimotor psychology. Assume that we are interested in explaining the capacity of a system to estimate distances between its hand and objects in the environment, in this instance, a punnet of blueberries that the system is trying to reach. Let's grant, for illustrative purposes, that part of the underlying mechanism is an R-mechanism. Call the actual distance between the hand and the blueberries D1. The *target* of the representation is D1. This is because D1 is the actual condition that the state needs to estimate for the capacity to succeed (understood in terms of what the state must

structurally correspond to). Suppose that given its structure, the R-mechanism causes the system to misestimate the distance between hand and blueberries (the R-mechanism fails to structurally correspond to D1). Instead, the R-mechanism structurally corresponds to a shorter distance in the agent's environment (conditions that instantiate the vehicle's structure), causing the system to reach this distance. Call this point D2. The MAC says that the *content* of the representation refers to D2. This captures the fact that if D2 were the target—if the R-mechanism was required to structurally correspond to D2 because the blueberries were at D2—then the behaviour would succeed. The R-mechanism will structurally correspond to many things, but only D2 reflects a relevant counter-factual.

The action-centeredness of the MAC equips us with a framework for better understanding the appeal of ascribing content to Swamp Rat (see **chapter 4**). Recall that Swamp Rat possesses a mechanism that is physically identical to an R-mechanism in Ordinary Rat (i.e., a cognitive map). I claimed that Swamp Rat's mechanisms play the same explanatory role (*qua* representation) for Swamp Rat as they do for Ordinary Rat. From that premise, I tentatively concluded that Swamp Rat's mechanisms possess content. Failing to allow this would endanger the causal relevance of content to the production or constitution of cognitive capacities of the sort studied by cognitive science. From the perspective of the MAC, the mechanisms in Swamp Rat and Ordinary Rat have identical semantic properties in accordance with the causal role of their underlying mechanisms. Take some capacity that Swamp Rat and Ordinary Rat both display, like navigating from a starting location in a laboratory maze to a location elsewhere in the laboratory maze containing a reward. Assume that the relative success of this behaviour depends on the structural correspondence between an R-mechanism and, roughly, the layout of the laboratory maze. Further assume that both creatures, being physically indistinguishable

and located within identical environments, succeed and fail to navigate under equivalent circumstances. According to the MAC, the target and content of the mechanisms in both creatures are the same. After all, the creatures display equivalent capacities within the same task environment, and their mechanisms are structurally identical. Though Ordinary Rat inherited its mechanisms through the normal processes of evolution and learning, and Swamp Rat inherited its mechanisms through the unusual process of a cosmic miracle, the indistinguishable behaviour of both creatures is explained by mechanisms playing the same representational role, with the same representational content.

By dispensing with traditional causal-historical theories, one may wonder just how widely the MAC construes the scope of 'correct representation'. The short answer is that correct representation begins and ends with successful action. A correct representation is rarely the result of blind luck. And yet, blind luck, should it ever occur, is just as good as mundane biological history when it comes to realising a cognitive capacity. Our analogy to ordinary S-representation helps to demonstrate this. Of course, it is no accident that a map of IKEA correctly represents multiple targets. Multiple IKEA stores all bear the same layout because those stores were produced according to a shared plan—that is, there are additional mechanisms that reliably coordinate the production of representation and multiple targets ensuring that the structure of the former systematically matches the structure of the latter. And yet, an IKEA map could feasibly facilitate the navigation of a shopper in an independent hardware store that just so happened, by freak coincidence, to share the same layout as an IKEA store. To the extent that the structural correspondence between map and layout causally explains how one successfully navigates the hardware store, the map can be said to correctly represent the hardware store.

As an aside, note that the MAC can still make sense of why one might say that the map represents IKEA when asked—even following successful navigation of the hardware store. This is because when asked what something represents in ordinary life, one is often being asked about its originally intended or typical target, given the reliable (but non-necessary) connection that holds between the target that a map was designed to stand-in for or usually stands-in for, the set of targets the map will successfully stand-in for, and the target a given user is intending the map to stand-in for.

The above IKEA map example highlights that the MAC permits 'lucky correctness' (in principle). This is a major difference between it and many traditional causal-historical theories. Of course, freakish coincidences are rare. We should expect additional mechanisms and biological histories to explain how a cognitive representation comes to correctly stand-in for, say, one's local environment or the distance between one's hand and an object. Correct cognitive representations do not usually pop into existence following freak lightning strikes (let's assume) but are reliably produced and coordinated by other mechanisms and are shaped by an organism's evolutionary and learning history. However, these important facts do not themselves determine the criteria for evaluating correctness and misrepresentation. As repeatedly suggested already, the history of a mechanism does not account for the role a mechanism plays in realising a capacity in the here-and-now. Hence R-mechanisms, like other mechanisms, may feature in successful explanations even when one remains ignorant about their history. In summary, for the MAC, though correct representation is likely the reliable product of additional mechanisms and a system's history, the criterion for evaluating correctness is to be identified with the extent to which an S-representation allows a consumer to complete a

capacity or task that causally depends on the S-representation's action-guiding structure. In a slogan, 'correct representation = successful action'.

In closing this introduction to the MAC, it is worth acknowledging the debt that the account owes to Cummins (1989, 1996). A complete exegesis comparing the MAC to Cummins' positive account of representation is beyond the scope of this thesis. Nonetheless, it is worth noting Cummins' (1996) suspicion of S-representation deriving from his belief that it is beholden to a 'use-based theory' of content. For Cummins, all such theories are flawed because, by identifying the content of representation with how representation is used (as he thinks they do), use-based theories conflate content and target. S-representations, therefore, cannot create the mismatch between content and target required to make sense of representational error. Cummins argues that what is needed is a theory that ties content to intrinsic features of the representation. Error is made possible because of the potential mismatch between these intrinsic features and their 'application'. In other words, error occurs when the 'target of tokening' fails to satisfy the representation's content (*ibid*., p. 6). This reasoning reflects the same ethos evident in the MAC. It is true that the MAC may be thought of as a use-based theory in so far as it claims that a mechanism derives its semantic properties from the way it is exploited by a consuming system, and it is only that structure which is relevant to action that contributes towards content determination. However, following Cummins' insights, the MAC also allows for error by drawing a distinction between the content of an R-mechanism and its target. In this way, the MAC avoids the pitfall that Cummins thinks befalls all use-based theories (for related discussion see Ramsey, 2007, pp. 104-107).

3.3 Clarifications and criticisms of the MAC

Reviewing some possible objections to the MAC will help to further clarify and strengthen the account. Perhaps the most obvious complaint is that the MAC has things backwards: surely it is a prior theory of content that explains where and why a mechanism counts as correct. But the MAC appears to portray the opposite picture: the MAC uses a prior notion of correspondence—the structural correspondence between a mechanism and some state of affairs required to realise a capacity and generate behavioural success—to explain how we should think about content. However, the very insistence of things needing to be in this order—for a theory of content, independent of the functional role played by a mechanism, to specify where correctness begins and ends—is precisely the narrative that the MAC resists. This owes in part to the observation of the explanatory role played by those theoretical posits which inspire the S-representation account, for example, cognitive maps. Prior to our theorising about content, scientific explanations of cognitive capacities that draw on 'iconic representations', like cognitive maps, already imply the importance of correspondence between a mechanism and some target given the way their structure determines the containing system's behaviour (Bechtel, 2008). The MAC says that we ought to think of subpersonal content in terms of the correctness conditions already implicit in these explanations of cognition.

One might suspect that the problem of content determination originated in a tension inherent in the traditional lingua-form representations associated with the language of thought (see **chapter 2**). It was often taken that the semantics of (apparently representational) cognitive states was not relevant to their causal powers within a cognitive system. This belief is reflected in Fodor's (1980) 'methodological solipsism', a position which holds that cognitive processes and states are to be construed with reference

only to formal operations performed over internal vehicles which are themselves individuated by their functional relationship with one another (see Wilson, 2004, for more recent discussion). At the same time, these vehicles were often assumed to possess representational content given their role as the reduction base for propositional attitudes. Again, however, the very identification of representations like cognitive maps (which inspire the S-representation account), involves an implicit attribution of content. To borrow from Bechtel, 'neuroscientists have tended to make the relation to the content central to the identification of vehicles and so have not faced the challenge of reconnecting the content to the vehicle.' (2008, p. 161). He goes on to say,

> The connection between content and vehicle is fundamental to how representations are characterized in neuroscience; consequently, neuroscience representations are more clearly grounded in the causal nexus relating organisms to their environments than are those advanced in cognitive science. (*ibid.*, p. 186)

Though Bechtel does not defend any detailed account of content in particular, his observation captures the same spirit that inspires the MAC.[6] Take another example. In Bayesian psychology, 'state estimates' are first identified as a theoretical posit because the framework deems it necessary to posit probabilistic estimations of environmental conditions. These conditions reflect the features of a token system's environment that the structure of the underlying mechanism needs to match for the capacity to succeed. There exists an implicit notion of target and content in the positing of a state estimate; a notion

[6] I adopt a more liberal understanding of cognitive science than Bechtel does, one which encompasses both neuroscientific and non-neuroscientific theories of cognition that posit similar representations (such as the classical computational theory of cognition). Bechtel (2008) also subscribes to a far more permissive notion of representation than the one defended in this thesis.

of what needs estimating and what is estimated—where success or failure is explained with recourse to the degree of correspondence between the two.[7]

The present discussion does draw attention to a more grievous doubt that arises from situating content determination within functional mechanisms: the threat of indeterminacy. Historically, representational content has been assumed to be fixed and determinate (for discussion, see Egan, 2014). This is often taken to mean that a representation's content is objective and does not vary with scientific practice (i.e., does not change depending on divergent explanatory goals). Such objectivity seems key to ensuring cognitive representation's naturalistic credentials. However, general anxieties about the determinacy of mechanism properties threaten to infect our conceptualisation of content in R-mechanisms.

There are two issues that one may have concerning the determinacy of content in R-mechanisms. Firstly, a component may plausibly play the role of representation in several mechanisms, meaning that the component's target and content will vary with the wider mechanism it serves. To my mind this is the least serious of the two issues, as though it implies that semantic properties are less fixed than traditional naturalised accounts of content assume, these nevertheless result from objective properties of the mechanism's wide functional relations. This is consistent with the kind of interrelatedness and reuse of neural structures that have become a fixture of contemporary cognitive science (for example, see Anderson, 2010). This was discussed in **chapter 2**. Secondly, whether a

[7]Once again, the need for a target/content distinction is strengthened by the possibility of system-detectable error. In those mechanisms capable of system-detectable error, as suggested in many Bayesian approaches, the system updates its future estimations to better match a set of target conditions. The capacity for updating an estimate implies the notion of a state bearing a content in need of updating (to better serve some system goal).

vehicle has content at all depends on whether there is a phenomenon that is explained by the capacity of an underlying mechanism with a functional role satisfying the criteria for S-representation. However, according to some, the *explanandum* capacity of a mechanistic explanation is fixed by the perspective of an observer (Craver, 2013). Therefore, the function of a mechanism depends on an *explanandum* fixed by the perspective of an observer. For example, the function of the heart is to pump blood relative to the circulatory system but to make *thump-thump* noises relative to diagnosing cardiovascular disease. Mechanisms do not have any essential, definitive functions, but only causal roles relative to the phenomena of interest to agents. Thus, R-mechanisms only possess their representational function (and thus content) given the explanatory inclinations of observers. If correct, the worry goes, R-mechanisms and the MAC threaten the objectivity of representational content.

I will note two related points in response to the more serious second worry. The first point is that the MAC avoids content becoming radically indeterminate, even if the more serious worry is correct. This is because once an *explanandum* is fixed (for example, navigation), and the resulting causal mechanisms identified (for example, a cognitive map), the function to represent becomes objective, and the specification of the target and the content-determining structure becomes fixed irrespective of any agent's judgement. This is true for any mechanism function. For example, relative to its role in blood circulation, the function of the heart to pump blood is fixed. It is for this reason that I think the term 'perspectivalism' (adopted by Craver, 2013) overemphasises the arbitrariness of functions under the causal-role account.

The second point is that worries about indeterminacy arise from broader issues concerning the mechanistic framework. To what extent mechanisms and their functions are indeterminate depends on this broader (and live) debate. Therefore, the proponent of the MAC may accept that content fixation is indeterminate, but only insofar as functional mechanisms are indeterminate more generally. Nonetheless, some mechanists have sought to modify the mechanistic framework to secure a more objective foundation for mechanism functions. For instance, Piccinini (2015, chapter 6) and Maley & Piccinini (2017) argue for a contemporary naturalistic account that seeks to provide objective criteria for 'teleological functions' in mechanisms. For Maley & Piccinini, a mechanism's function is the 'stable contribution by a trait (or component, activity, property) of organisms belonging to a biological population to an objective goal of those organisms' (p. 244). Let's examine this account further.

Maley & Piccinini reject etiological accounts of function for reasons similar to those offered in earlier chapters. These include concerns about the opaqueness of a system's selection history and the irrelevance of selection history to the present causal powers of a trait (2017, pp. 238-239). However, they also reject the adequacy of the standard causal-role account because they think it threatens to overproliferate functions and leads to a counterintuitive perspectivalism (*ibid.*, p. 240). By contrast, Maley & Piccinini seek an objective and 'ontologically serious' foundation for 'teleological functions' in mechanisms (*ibid.*, pp. 236-237). They maintain that, on their account, 'functions are an aspect of what a system *is*, rather than an aspect of what we may or may not say about that system' (*ibid.*, p. 237. Original emphasis). I will call the kind of functions Maley & Piccinini discuss 'objective goal functions'.

Maley & Piccinini's account contains many subtleties, but the fundamental idea is that a mechanism's function must contribute to a system's goals. Science seeks to explain the causal contribution of a living system's parts to those goals, ascribing functions accordingly.[8] Goals are understood as states that the system is organised to bring about and a mechanism's function is its contribution towards that goal. More exactly, systems have 'objective goals' understood in terms of survival and inclusive fitness.[9] These are states 'toward which the energy expenditure, via mechanisms, must work in order for organisms to exist' (2017, p. 243). A mechanism's function is a stable contribution to survival or inclusive fitness. For example, the objective goal function of the heart is to pump blood (as opposed to making *thump-thump* noises) because pumping blood is the causal contribution of the heart to the survival of the containing organism. Another way to think about this is that living systems behave in such a way as to contribute to their survival and inclusive fitness, and a mechanism's function is its role in bringing about that behaviour (*ibid.*, p. 247). For instance, when a rat seeks out food within its environment it contributes to its own survival; the function of a cognitive map is to represent its environment because that is its role in bringing about that survival-contributing behaviour. Also note that Swamp Rat possesses objective goal functions, in so far as it is organised (just like Ordinary Rat) to survive and reproduce. As such, Swamp Rat's cognitive maps also have the objective goal function to represent. In summary, the objective goal account offers an alternative to etiological approaches to function and appears to accord with much function ascription within scientific practice (*ibid.*, p. 253).

[8] The subtleties of Maley & Piccinini's arguments allow, for example, the extension of their account to include artefacts and human aims (so-called 'subjective goals') that may be orthogonal to strictly biological ends.

[9] 'Inclusive fitness' refers to the theory that a gene may contribute towards its own selective success through its contribution to the reproduction of other organisms who possess copies of that gene (Hamilton, 1964). This provides a means for the evolution of altruistic behaviour.

Maley & Piccinini are careful to distinguish their position from standard causal-role accounts as well as etiological accounts (2017, pp. 240-241). However, the core of Maley & Piccinini's theory is not in conflict with the causal-role account. Instead, I suggest, it supplements it. After all, objective goal functions are still a kind of causal role (Krickel offers a characterisation of objective goal functions that emphasises this fact; 2018, p. 45). Maley & Piccinini's account does not undermine the importance of causal roles for grounding function so much as limit the sorts of causal roles that count as objective goal functions. From this starting point, we can imagine a modified version of Maley & Piccinini's account that accepts the base causal-role account—a mechanism's functional role is its causal role relative to the production or constitution of some *explanandum*—whilst individuating an interesting subset of causal roles, namely, those which contribute towards an organism's 'objective goals'. It is just these objective goals, let's grant, that are of interest to biology and cognitive science. Under this interpretation, hearts really do have the function to make *thump-thump* noises relative to diagnosing heart disease (a conclusion Maley & Piccinini resist; 2017, p. 240). However, unlike pumping blood, making *thump-thump* noises does not count as an objective goal function—the sort of function that biologists are interested in.[10] The function of the heart to pump blood remains 'objective', in so far as pumping blood is objectively its causal role relative to the organism's objective goal. This interpretation of Maley & Piccinini's account conforms with the pluralist approach to accounts of function outlined in **chapter 2**.[11]

[10] Hearts do have other 'objective goal functions' besides pumping blood, such as thermoregulation. The objective goal account does not prohibit mechanisms from possessing multiple objective goal functions. For discussion, see Piccinini (2015, p. 103).

[11] One major advantage of this interpretation over the standard objective goal account of functions presented by Maley & Piccinini is that it still permits us to continue to talk about the functions of maladaptive mechanisms (see **chapter 2**).

Maley & Piccinini's objective goal account offers a promising alternative to etiological approaches whilst constraining the sort of causal roles that count as functions relative to the interests of biology and cognitive science. I do not intend this brief sketch to conclusively demonstrate the determinacy of content given a mechanistic approach to cognitive representation. Rather, I intend only to gesture towards the possibility of further cementing the objectivity of content by modifying the base causal role account of mechanism functions. Ultimately, these issues go beyond the scope of R-mechanisms and the MAC. The account of representation I defend takes for granted the *explananda* of cognitive science (i.e., paradigmatic cognitive capacities) which, once fixed, serve to ground the criteria for R-mechanisms and their content, leaving it to accounts like those defended by Maley & Piccinini to explain the significance of such *explananda*.

In closing our discussion of the MAC, it is worth returning to worries about representation raised by certain proponents of 4E cognition. Recall from **chapter 2** that the action-oriented representation (AO-representation) account questions those concepts of cognitive representation that suggest 'objective' and 'action-neutral' contents, instead emphasising the importance of ecological constraints on any plausible form of representation, and the possibility of action-relevant 'imperative' contents (for example, see Clark, 1997; Mandik, 2005). Such contents are, in some sense, related to the real world needs and behaviours of a resource-bound system. Is the MAC friendly to an action-oriented approach to representation?

Future research will be devoted to further exploring the relationship between R-mechanisms, the MAC, and 4E approaches to cognition in general. For now, I will settle for noting a *prima facie* sense in which the account of representation on offer is friendly

to AO-representation. I already stressed above that the MAC takes the exploitation of cognitive representation for action seriously in so far as it adopts the success/failure of cognitive capacities as its focal point for characterising content determination. According to the MAC, contents require targets, and targets are specified by the needs of the acting system consuming the representation—needs that may be highly species- and individual-relative. The content of an R-mechanism refers only to that which is relevant to the realisation of behavioural success. In turn, whilst intrinsic structure helps determine content, it only the structure of action-relevant parts of the mechanism that is semantically relevent, that is, those parts that affect the system's processing and motor outputs. In this way, the MAC places action front and centre.

It is also worth simply noting that that the R-mechanism account is consonant with more general lessons from 4E cognition concerning the non-representational nature of at least some cognition, and the likely interplay between both representational and bodily/environmental resources for those capacities that do involve representation. Finally, it strikes me that the action-oriented flavour of the MAC is enhanced when combined with an account of function like Maley & Piccinini's. If the objective goal account is correct, then an R-mechanism's function to represent is firmly rooted in the survival and inclusive fitness of an ecologically embedded system. An R-mechanism and its content are thus tethered to the needs of the active system they play a role within. To summarise, R-mechanisms and the MAC are well placed to deliver an account that not only articulates a clear set of conditions under which ascriptions of subpersonal cognitive representation make a robust explanatory contribution but do so in a way that accords with the considerations of 4E cognition.

4. Conclusion

This chapter distinguished between two closely related problems about representational content. The ‘hard problem of content’ concerns the justification for attributing representational content at the subpersonal level. The ‘problem of content determination’ concerns how a token representation acquires its particular content. The hard problem becomes easier when we consider that correctness conditions emerge from the need for structural correspondence in cases where a cognitive capacity depends on an R-mechanism. In turn, the content determination problem is assuaged by observing that token R-mechanisms have certain target states of affairs that they must structurally correspond to in order to count as correct—as determined by the capacity they serve and the wider task environment. Under the ‘mechanistic account of content’, a representation’s content refers to the state of affairs that must obtain for it to realise a cognitive capacity and generate behavioural success. The relative overlap between target and content underwrites an R-mechanism’s degree of correctness or incorrectness and thus allows for the possibility of misrepresentation. Together, these responses to the two problems of content firmly plant the semantic properties of cognitive representation within causal-mechanistic explanation.

A principal virtue of the S-representation account is that traditional problems associated with naturalising content begin to dissipate. The justifiable ascription of semantic properties at the subpersonal level and the content determination of a token representation are grounded in ordinary facts concerning the causal role played by a class of mechanisms in routine cognitive capacities. Content no longer looks spooky, and so, the compulsion to wholly exorcise representation from cognitive science is dispelled.

Thesis Conclusion

This thesis examined the role of representation in explanations of cognition. Such a longstanding issue may sometimes seem like a perennial philosophical dispute, with so many competing intuitions that it's unclear whether a solution is even possible. And yet, this thesis indicates that a solution is finally within our grasp. The mechanistic approach that I have advocated takes the best lessons from both sides of the debate to hone our understanding of cognitive representation. Cognitive representation can play a role in explanations of cognition: a cognitive representation is a type of cognitive mechanism—one that meets the criteria for S-representation. In turn, a representational explanation is a type of mechanistic explanation.

Let's summarise our findings. I have argued that we should hesitate before assuming the notion of cognitive representation is a category error, remaining open to the possibility that 'representation' informatively describes cognitive activity at the subpersonal level. As we saw, exploring whether ascriptions of representation at the subpersonal level contribute to a scientific theory is not the same thing as settling the final ontological status of cognitive representation. Whether and how representation plays a part in our best scientific theory of cognition is of primary importance to cognitive science, not whether subpersonal cognitive representations are ultimately real in a strong, metaphysical sense.

If my analysis is correct, then several traditional ways of thinking about subpersonal cognitive representation fail to secure its explanatory significance. Receptors do not possess a functional role that is distinctly representation-like. Neither do action-oriented considerations, by themselves, show that subpersonal entities serve as representations.

Adopting the intentional stance also fails to support representationalism, remaining neutral on the sorts of internal mechanistic processes that cognitive science investigates. Finally, a computational approach to cognition does not automatically imply a representational approach to cognition. Physical computation does not presuppose representation and there's nothing about the structure of computational explanations in cognitive science that necessitates representation.

Despite these negative results, there remains at least one notion in philosophy and cognitive science that does supply representation with a substantive explanatory role. According to the S-representation account, a cognitive representation is a type of internal map or model-like entity that guides the actions of a cognitive system by mirroring the structure of the world. This provides a clear set of empirically plausible functional criteria that is consonant with a broader mechanistic framework of explanation in cognitive science. The resulting picture of a 'representational mechanism'—a cognitive mechanism that meets the functional criteria for S-representation—provides the grounds for a robust and distinctive form of representational explanation. In this way, if a theory of cognition posits representational mechanisms, then it is a representational theory in a well-defined and substantive sense.

The notion of a representational mechanism dissipates many traditional worries associated with naturalising cognitive representation. In particular, it accounts for representation's paradigmatic semantic properties at the subpersonal level. The functional criteria met by a representational mechanism allows for the attribution of semantic properties in a naturalistically respectable fashion. At the same time, these functional criteria supply an intuitive way to think about content determination. The mechanistic

account of content unpacks the particular semantic properties of a token representational mechanism in terms of what state of affairs needs to be actual for the mechanism to realise a capacity and generate behavioural success.

Questions inevitably remain. Moving forward, I intend to strengthen the mechanistic approach to representation by developing some suggestions touched upon throughout this thesis. Three main research areas stand out. The first research area concerns whether the burgeoning conceptual engineering literature can help assuage worries over ascribing representation at the subpersonal level. Building on the ethos of this thesis, I think it will prove fruitful to move further towards a normative approach to psychological concepts that examines how our understanding of representation can serve (and be revised by) scientific needs. The second research area concerns the compatibility of the account outlined in this thesis with 4E approaches which I have only been able to touch upon briefly. Such approaches traditionally vary in their acceptance of cognitive representation. The mechanistic approach promises to help taxonomize and adjudicate the representational commitments of different 4E approaches. On a related note, the final research area concerns ecologically-focussed theories of mechanism function. I will further explore how the objective goal account can complement the mechanistic approach to cognitive representation whilst sustaining a broadly pluralist attitude toward theories of function. I also intend to investigate connections between the objective goal account and similar notions of function within the cybernetics tradition. Pursuing these research areas will test the boundaries of the mechanistic approach to cognitive representation, and in doing so, promises to advance our understanding of explanation in cognitive science.

Bibliography

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences*, 33(4), 245-266.

Anderson, M. L., & Rosenberg, G. (2008). Content and action: The guidance theory of representation. *Journal of Mind and Behavior*, 29(1-2), 55-86.

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states?. *Psychological Review*, 116(4), 953.

Barrett, D. (2014). Functional analysis and mechanistic explanation. *Synthese*, 191(12), 2695-2714.

Bartels, A. (2006). Defending the structural concept of representation. *Theoria. Revista de Teoría, Historia y Fundamentos de la Ciencia*, 21(1), 7-19.

Barve, A., & Wagner, A. (2013). A latent capacity for evolutionary innovation through exaptation in metabolic systems. *Nature*, 500(7461), 203-206.

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

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

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

Bechtel, W. (2016). Investigating neural representations: The tale of place cells. *Synthese*, 193(5), 1287-1321.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences,* 36(2), 421-41.

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

Bennett, A. T. (1996). Do animals have cognitive maps?. *Journal of Experimental Biology*, 199(1), 219-224.

Bennett, M., & Hacker, P. (2007). Selections from philosophical foundations of neuroscience. In M. Bennett, D. Dennett, P. Hacker, & J. Searle. (Eds.), *Neuroscience and philosophy: Brain, mind and languag*e. New York: Columbia University Press.

Bermúdez, J. L. (2009). Mindreading in the animal kingdom? In R. Lurz (Ed.), *The Philosophy of Animal Minds* (pp. 145–164). New York: Cambridge University Press.

Bermúdez, J. L. (2014). *Cognitive science: An introduction to the science of the mind* (2nd ed.). Cambridge: Cambridge University Press.

Bickhard, M. H. (1999). Interaction and representation. *Theory & Psychology*, 9(4), 435-458.

Block, N. (1987). Advertisement for a semantics for psychology. *Midwest Studies in Philosophy*, 10(1), 615-678.

Block, N. (1990). The computer model of the mind. In D. Osherson, & E. Smith (Eds.), *An invitation to cognitive science, vol. 3: Thinking*. Cambridge, MA: MIT Press.

Boden, M. A. (2006). *Mind as machine: A history of cognitive science*. Oxford: Oxford University Press.

Borg, E. (2018). On deflationary accounts of human action understanding. *Review of Philosophy and Psychology*, 9(3), 503-522.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, 47(1), 139-159.

Brown, C. (2016). Narrow mental content. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2016 ed.). Retrieved from https://plato.stanford.edu/archives/sum2016/entries/content-narrow/

Brun, G. (2016). Explication as a method of conceptual re-engineering. *Erkenntnis*, 81(6), 1211-1241.

Carnap, R. (1936). Testability and meaning. *Philosophy of Science*, 3(4), 419–71.

Chakravartty, A. (2010). Informational versus functional theories of scientific representation. *Synthese*, 172(2), 197-213.

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

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

Chiel, H. J., & Beer, R. D. (1997). The brain has a body: Adaptive behavior emerges from interactions of nervous system, body and environment. *Trends in Neurosciences*, 20(12), 553-557.

Churchland, P. S. (1981). Eliminative materialism and the propositional attitudes. *The Journal of Philosophy*, 78(2), 67-90.

Churchland, P. S., & Sejnowski, T. J. (1990). Neural representation and neural computation. *Philosophical Perspectives*, 4, 343-382.

Clark, A. (1997). *Being there: Putting brain, body, and world together again*. Cambridge, MA: MIT Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3), 181-204.

Clark, A. (2014). *Mindware: An introduction to the philosophy of cognitive science* (2nd ed.). Oxford: Oxford University Press.

Clark, A. (2016). *Surfing uncertainty*. Oxford: Oxford University Press.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7-19.

Clark, A., & Toribio, J. (1994). Doing without representing?. *Synthese*, 101(3), 401-431.

Coelho Mollo, D. (2018). Functional individuation, mechanistic implementation: The proper way of seeing the mechanistic view of concrete computation. *Synthese*, 195(8), 3477-3497.

Craik, K. J .W. (1943). *The nature of explanation*. Cambridge, UK: Cambridge University Press.

Crane, T. (1990). The language of thought: No syntax without semantics. *Mind and Language*, 5(3), 187-213.

Crane, T. (2016). *The mechanical mind: A philosophical introduction to minds, machines and mental representation* (3rd ed.). New York: Routledge.

Craver, C. (2001). Role functions, mechanisms and hierarchy. *Philosophy of Science* 68(1), 31-55.

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

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

Craver, C. (2013). Functions and mechanisms: A perspectivalist view. In P. Huneman (Ed.), *Functions: Selection and mechanisms* (pp. 133-158). Dordrecht: Springer.

Craver, C., & Darden, L. (2013). *In search of mechanisms: Discoveries across the life sciences*. Chicago: University of Chicago Press.

Cummins, R. (1975). Functional analysis. *The Journal of Philosophy*, 72(20), 741-65.

Cummins, R. (1983). The nature of psychological explanation. Cambridge, MA: MIT Press.

Cummins, R. (1989). *Meaning and mental representation*. Cambridge, MA: MIT Press.

Cummins, R. (1991). The role of representation in connectionist explanations of cognitive capacities. In W. M Ramsey, S. P. Stich, & D. E Rumelhart (Eds.), *Philosophy and connectionist theory* (pp. 91-114). Hillsdale, NJ: Lawerence Erlbaum.

Cummins, R. (1996). *Representations, targets, and attitudes*. Cambridge, MA: MIT Press.

Davidson, D. (1987). Knowing one's own mind. *Proceedings of the American Philosophical Association,* 60(3), 441–58.

Dehaene, S., & Cohen, L. (2007). Cultural recycling of cortical maps. *Neuron*, 56(2), 384-398.

Dehaene, S. (2009). *Reading in the brain: The new science of how we read*. New York: Penguin.

Dennett, D. (1969). *Content and consciousness*. London: Routledge & Kegan Paul.

Dennett, D. (1982). Styles of mental representation. *Proceedings of the Aristotelian Society*, 83, 213-226.

Dennett, D. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Dennett, D. (1991). Real patterns. *The Journal of Philosophy*, 88(1), 27-51.

Dennett, D. (2018). Reflections on Tadeusz Zawidzki. In B. Huebner (Ed.), *The philosophy of Daniel Dennett* (pp. 57-61). Oxford University Press: New York.

Dewhurst, J. (2018). Individuation without representation. *The British Journal for the Philosophy of Science*, 69(1), 103-116.

Dietrich, E., & Markman, A. B. (2003). Discrete thoughts: Why cognition must use discrete representations. *Mind & Language*, 18(1), 95-119.

Downey, A. (2018). Predictive processing and the representation wars: A victory for the eliminativist (via fictionalism). *Synthese*, 195(12), 5115-5139.

Drayson, Z. (2012). The uses and abuses of the personal/subpersonal distinction. *Philosophical Perspectives*, 26(1), 1-18.

Drayson, Z. (2014). The personal/subpersonal distinction. *Philosophy Compass*, 9(5), 338-346.

Dretske, F. (1981). *Knowledge and the flow of information*. Cambridge, MA: MIT Press.

Dretske, F. (1988). *Explaining behavior: Reasons in a world of causes*. Cambridge, MA: MIT Press.

Dupré, J. (1995). *The disorder of things: Metaphysical foundations of the disunity of science*. Cambridge, MA: Harvard University Press.

Dupré, J. (2002). *Humans and other animals*. Oxford: Oxford University Press.

Dupuy, J. P. (2009). *The mechanization of the mind: On the origins of cognitive science*. Cambridge, MA: MIT Press.

Eberle, R., Kaplan, D., & Montague, R. (1961). Hempel and Oppenheim on explanation. *Philosophy of Science*, 28(4), 418-428.

Egan, F. (1995). Computation and content. *The Philosophical Review*, 104(2), 181-203.

Egan, F. (2010). Computational models: A modest role for content. *Studies in History and Philosophy of Science* (special issue on Computation and Cognitive Science). 41, 253-259.

Egan, F. (2014). How to think about mental content. *Philosophical Studies*, 170(1), 115-135.

Figdor, C. (2014). Verbs and minds. In M. Sprevak, & J. Kallestrup (Eds.), *New waves in philosophy of mind* (pp. 38-53). London: Palgrave-Macmillan.

Figdor, C. (2017). On the proper domain of psychological predicates. *Synthese*, 194(11), 4289-4310.

Figdor, C. (2018). *Pieces of mind: The proper domain of psychological predicates*. Oxford: Oxford University Press.

Floridi, L. (2005). Is semantic information meaningful data? *Philosophy and Phenomenological Research*, 70(2), 351–370.

Floridi, L. (2010). *Information: A very short introduction*. Oxford: Oxford University Press.

Fodor, J. A. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.

Fodor, J. A. (1980). Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and Brain Science*, 3(1), 63–73.

Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. Cambridge, MA: MIT Press.

Fodor, J. A., & Pylyshyn, Z. W. (1981). How direct is visual perception? Some reflections on Gibson's "ecological approach". *Cognition,* 9(2), 139-196.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2), 3-71.

Forge, J. (1980). The structure of physical explanation. *Philosophy of Science*, 47(2), 203-226.

Franceschini, N., Pichon, J. M., & Blanes, C. (1991). Real time visuomotor control: From flies to robots. *Fifth International Conference on Advanced Robotics 'Robots in Unstructured Environments'*, 2, 931-935.

Franklin, S. (1995). *Artificial minds*. Cambridge, MA: MIT Press.

Fresco, N. (2014). *Physical computation and cognitive science*. Heidelberg: Springer.

Frigg, R. (2006). Scientific representation and the semantic view of theories. *Theoria. Revista de Teoría, Historia y Fundamentos de la Ciencia*, 21(1), 49-65.

Friston, K. (2009). The free-energy principle: A rough guide to the brain?. *Trends in Cognitive Sciences*, 13(7), 293-301.

Gallistel, C. R. (1990). *The organization of learning*. Cambridge, MA: MIT Press.

Garson, J. (2013). The functional sense of mechanism. *Philosophy of Science*, 80(3), 317-333.

Garson, J. (2015). *The biological mind: A philosophical introduction*. London: Routledge.

Garson, J. (2016). *A critical overview of biological functions*. Dordrecht: Springer.

Garson, J. (2017). A generalized selected effects theory of function. *Philosophy of Science*, 84(3), 523-543.

Garson, J. (2018). How to be a function pluralist. *British Journal for the Philosophy of Science*. 69(4), 1101-1122.

Garzón, F. C. (2008). Towards a general theory of antirepresentationalism. *The British Journal for the Philosophy of Science*, 59(3), 259-292.

Gergely, G. (2002). The development of understanding self and agency. In U. Goswami (Ed.), *Blackwell Handbook of Childhood Cognitive Development* (pp. 26-46). Oxford: Blackwell.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naive theory of rational action. *Trends in Cognitive Sciences*, 7(7), 287-292.

Gładziejewski, P. (2015). Explaining cognitive phenomena with internal representations: A mechanistic perspective. *Studies in Logic, Grammar and Rhetoric*, 40(53), 63-90.

Gładziejewski, P. (2016a). Action guidance is not enough, representations need correspondence too: A plea for a two-factor theory of representation. *New Ideas in Psychology*, 40, 13-25.

Gładziejewski, P. (2016b). Predictive coding and representationalism. *Synthese*, 193(2), 559-582.

Gładziejewski, P., & Miłkowski, M. (2017). Structural representations: Causally relevant and different from detectors. *Biology & Philosophy*, 32(3), 337–355.

Glennan, S. (2002). Rethinking mechanistic explanation. *Philosophy of Science*, 69(3), 342-353.

Godfrey-Smith, P. (1992). Indication and adaptation. *Synthese*, 92(2), 283–312.

Godfrey-Smith. P. (1993). Functions: Consensus without unity. *Pacific Philosophical Quarterly*, 74(3), 196-208.

Godfrey-Smith, P. (2009). Triviality arguments against functionalism. *Philosophical Studies*, 145(2), 273–295.

Goodman, N. (1968). *Languages of art: An approach to a theory of symbols*. London: Oxford University Press.

Gould, S. J. (1997). Evolution: The pleasures of pluralism, *New York Review of Books*. Retrieved from https://www.nybooks.com/articles/1997/06/26/evolution-the-pleasures-of-pluralism

Gould, S. J., & Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society. London, Series B*, 205(1161), 581-598.

Grice, H. P. (1957). Meaning. *The Philosophical Review*, 66(3), 377-388.

Grush, R. (2008). Representation reconsidered by William M. Ramsey. *Philosophical Reviews.* Retrieved from https://ndpr.nd.edu/news/23327-representation-reconsidered

Hamilton, W. D. (1964). The genetical evolution of social behaviour. II. *Journal of Theoretical Biology*, 7(1), 17-52.

Hardcastle, V. (1999). Understanding functions: A pragmatic approach. In V. Hardcastle (Ed.), *Where biology meets psychology: Philosophical essays* (pp. 27-43). Cambridge, MA: MIT Press.

Hardcastle, V. (2002). On the normativity of functions. In A. Ariew (Ed.), *Functions* (pp. 144-156). Oxford: Oxford University Press.

Haslanger, S. (2012). *Resisting reality: Social construction and social critique*. Oxford: Oxford University Press.

Haugeland, J. (1991). Representational genera. In W. M Ramsey, S. P Stich, & D. E. Rumelhart (Eds.), *Philosophy and connectionist theory* (pp. 61-89). Hillsdale, NJ: Lawrence Erlbaum.

Haugeland, J. (1998). *Having thought: Essays in the metaphysics of mind.* Cambridge, MA: Harvard University Press.

Hempel, C. (1942). The function of general laws in history. *Journal of Philosophy*, 39, 35–48.

Hempel, C., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, 15, 135–175.

Hohwy, J. (2013). *The predictive mind*. Oxford: Oxford University Press.

Hornsby, J. (2000). Personal and sub-personal; A defence of Dennett's early distinction. *Philosophical Explorations*, 3(1), 6-24.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology,* 160(1), 106–154.

Huebner, B. (2018). Introduction to the philosophy of Daniel Dennett. In B. Huebner (Ed.), *The Philosophy of Daniel Dennett* (pp. xi-xxxiv). Oxford: Oxford University Press.

Hurley, S. L. (2001). Perception and action: Alternative views. *Synthese,* 129(1), 3-40.

Hurley, S. L. (2002). *Consciousness in action*. London: Harvard University Press.

Hutto, D. (2008). *Folk psychological narratives: The sociocultural basis of understanding reasons.* Cambridge, MA: MIT Press.

Hutto, D., & Myin, E. (2013). *Radicalizing enactivism: Basic minds without content*. Cambridge, MA: MIT Press.

Hutto, D., & Myin, E. (2014). Neural representations not needed: No more please. *Phenomenology and the Cognitive Sciences*, 13(2), 241–256.

Hutto, D., & Myin, E. (2017). *Evolving enactivism: Basic minds meet content*. Cambridge, MA: MIT Press.

Jeffery, K. J. (2015). Spatial cognition: Entorhinal cortex and the hippocampal place-cell map. *Current Biology*, 25(24), 1181-1183.

Jensen, R. (2006). Behaviorism, latent learning, and cognitive maps: Needed revisions in introductory psychology textbooks. *The Behavior Analyst,* 29(2), 187-209.

Johnson, K. (2004). On the systematicity of language and thought. *The Journal of Philosophy*, 101(3), 111-139.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press.

Johnson-Laird, P. N. (1998). Imagery, visualization, and thinking. In J. Hochberg (Ed.), *Perception and cognition at the century's end* (pp. 441-467). San Diego, CA: Academic Press.

Kaiser, M. I. (2017). The components and boundaries of mechanisms. In S. Glennan, & P. Illari (Eds.), *The Routledge handbook of mechanisms and mechanical philosophy* (pp. 116-130). New York: Routledge.

Kaplan, D. M., & Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of Science*, 78(4), 601-627.

Kitcher, P. (1984). Species. *Philosophy of Science,* 51(2), 308-333.

Kosslyn, S. M. (1994). *Image and brain: The resolution of the imagery debate*. MIT Press, Cambridge, MA.

Krickel, B. (2018). *The mechanical world: The metaphysical commitments of the new mechanistic approach*. Cham, Switzerland: Springer.

Kukla, R. (2018). Embodied stances: Realism without literalism. In B. Huebner (Ed)., *The philosophy of Daniel Dennett* (pp. 3-31). Oxford: Oxford University Press.

Ladyman, J. (2009). What does it mean to say that a physical system implements a computation? *Theoretical Computer Science,* 410(4-5), 376-383.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to western thought*. New York: Basic Books.

Lee, J. (2017). Mental representation and two kinds of eliminativism. *Philosophical Psychology*, 31(1), 1-24.

Lee, J. (forthcoming a). Mechanisms, wide functions and contents: Towards a computational pluralism. To appear in *The British Journal for the Philosophy of Science.*

Lee, J. (forthcoming b). Structural representation and the two problems of content. To appear in *Mind and Language.*

Lettvin, J. Y., Maturana, H. R., McCulloch, W. S., & Pitts, W. H. (1959). What the frog's eye tells the frog's brain. *Proceedings of the Institute of Radio Engineers*. 47, 1940-1951.

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

Malafouris, L. (2013). *How things shape the mind*. Cambridge, MA: MIT Press.

Maley, C. (2018). Towards analog computation. *Minds and Machines,* 28(1), 77-91.

Maley, C., & Piccinini, G. (2017). A unified mechanistic account of teleological functions for psychology and neuroscience. In D. Kaplan (Ed.), *Explanation and integration in mind and brain science* (pp. 236-256). Oxford: Oxford University Press.

Mandik, P. (2005). Action-oriented representation. In A. Brook, & K. Akins (Eds.), *Cognition and the brain: The philosophy and neuroscience movement* (pp. 284-305). Cambridge: Cambridge University Press.

McBeath, M. K., Shaffer, D. M., & Kaiser, M. K. (1995). How baseball outfielders determine where to run to catch fly balls. *Science*, 268(5210), 569–573.

McDowell, J. (1994). The content of perceptual experience. *The Philosophical Quarterly*, 44(175), 190-205.

McLaughlin, P. (2000). *What functions explain: Functional explanation and self-reproducing systems*. Cambridge: Cambridge University Press.

Menary, R. (2007). *Cognitive integration: Mind and cognition unbounded*. Basingstoke: Palgrave MacMillan.

Miłkowski, M. (2013). *Explaining the computational mind*. Cambridge, MA: MIT Press.

Miłkowski, M. (2015a). The hard problem of content: Solved (long ago). *Studies in Logic, Grammar and Rhetoric*, 41(1), 73-88.

Miłkowski, M. (2015b). Satisfaction conditions in anticipatory mechanisms. *Biology & philosophy*, 30(5), 709-728.

Miłkowski, M. (2017). The false dichotomy between causal realization and semantic computation. *Hybris*, 38, 1–21.

Miłkowski, M., Clowes, R., Rucińska, Z., Przegalińska, A., Zawidzki., T, Krueger., J., Gies., A., McGann, M., Afeltowicz., Ł., Wachowski, W., Stjernberg, F., Loughlin, V., & Hohol, M. (2018). From wide cognition to mechanisms: A silent revolution. *Frontiers in Psychology*. 9. Online. doi: 10.3389/fpsyg.2018.02393

Millikan, R. G. (1984). *Language, thought, and other biological categories: New foundations for realism*. Cambridge, MA: MIT Press.

Millikan, R. G . (1989a). In defense of proper functions. *Philosophy of Science*, 56(2), 288–302.

Millikan, R. G. (1989b). Biosemantics. *The Journal of Philosophy*, 86(6), 281-297.

Millikan, R. G. (1990). Compare and contrast Dretske, Fodor, and Millikan on teleosemantics. *Philosophical Topics*, 18(2), 151-161.

Millikan, R. G. (1993). *White queen psychology and other essays for Alice*. Cambridge, MA: MIT Press.

Millikan, R. G. (1995). Pushmi-pullyu representations. *Philosophical Perspectives*, 9, 185-200.

Millikan, R. G. (2017). *Beyond concepts: Unicepts, language, and natural information*. Oxford: Oxford University Press.

Mole, C., & Zhao, J. (2016). Vision and abstraction: An empirical refutation of Nico Orlandi's non-cognitivism. *Philosophical Psychology*, 29(3), 365-373.

Morgan, A. (2014). Representations gone mental. *Synthese*, 191(2), 213-244.

Morris, R. (1984). Developments of a water-maze procedure for studying spatial learning in the rat. *Journal of Neuroscience Methods*, 11(1), 47–60.

Moser, E. I., Kropff, E., & Moser, M. B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annual Review of Neuroscience*, 31, 69-99.

Myin, E., & Hutto, D. (2015). REC: Just radical enough. *Studies in Logic, Grammar and Rhetoric*, 41(1), 61-71.

Neander, K. (1991). The teleological notion of 'function'. *Australasian Journal of Philosophy*, 69(4), 454-468.

Neander, K. (2009). Teleological theories of mental content. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy* (Winter 2009 ed.). Retrieved from http://plato.stanford.edu/archives/win2009/entries/content-teleological/

O'Brien, G., & Opie, J. (2004). Notes toward a structuralist theory of mental representation. In H. Clapin, P. J. Staines, & P. Slezak (Eds.), *Representation in mind: New approaches to mental representation. Perspectives on cognitive science* (pp. 1-20). Amsterdam, Netherlands: Elsevier.

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford: Clarendon Press.

Orlandi, N. (2014). *The innocent eye: Why vision is not a cognitive process*. Oxford: Oxford University Press.

Papineau, D. (1984). Representation and explanation. *Philosophy of Science*, 51(4), 550–72.

Peirce, C. (1998). *The essential Peirce. Volume 2.* Peirce Edition Project (Eds.), Bloomington, IN: Indiana University Press.

Piccinini, G. (2007). Computational modelling vs. Computational explanation: is everything a Turing Machine, and does it matter to the philosophy of mind? *Australasian Journal of Philosophy*, 85(1), 93-115.

Piccinini, G. (2008). Computation without representation. *Philosophical Studies*, 137(2), 205-241.

Piccinini, G. (2015). *Physical computation: A mechanistic account*. Oxford: Oxford University Press.

Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, 183(3), 283–311.

Piccinini, G., & Scarantino, A. (2010). Computation vs. information processing: Why their difference matters to cognitive science. *Studies in History and Philosophy of Science Part A*, 41(3), 237-246.

Piccinini, G., & Scarantino, A. (2011). Information processing, computation, and cognition. *Journal of Biological Physics*, 37(1), 1-38.

Popper, K. (1959). *The logic of scientific discovery*. London: Hutchinson.

Pöyhönen, S. (2014). Intentional concepts in cognitive neuroscience. *Philosophical Explorations*, 17(1), 93-109.

Putnam, H. (1992). Philosophy and our mental life. In B. Beakley, & P. Ludlow (Eds.), *The philosophy of mind: Classical problems/contemporary issues* (pp. 91-98). Cambridge, MA: MIT Press.

Quine, W. V. (1948). On what there is. *The Review of Metaphysics*, 2(1), 21-38.

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

Ramsey, W. M. (2016). Untangling two questions about cognitive representation. *New Ideas in Psychology*, 40(A), 3-12.

Ramsey, W. M. (2018). Maps, models and computational simulations in the mind. In M. Sprevak, & M. Colombo (Eds.), *The Routledge handbook of the computational mind* (pp. 259-271). New York: Routledge.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, 20(4), 873-922.

Rescorla, M. (2009). Cognitive maps and the language of thought. *The British Journal for the Philosophy of Science*, 60(2), 377-407.

Rescorla, M. (2012). How to integrate representation into computational modeling, and why we should. *Journal of Cognitive Science*, 13(1), 1-37.

Rescorla, M. (2014). The causal relevance of content to computation. *Philosophy and Phenomenological Research*, 88(1), 173-208.

Rescorla, M. (2015). Bayesian perceptual psychology. In M. Matthen (Ed.), *The Oxford Handbook of the philosophy of perception* (pp. 694-716). New York: Oxford University Press.

Rescorla, M. (2016). Bayesian sensorimotor psychology. *Mind & Language*, 31(1), 3-36.

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

Sawyer, S. (2018). The importance of concepts. *Proceedings of the Aristotelian Society*. 118(2), 127-147.

Saxberg, B. V. H. (1987a). Projected free fall trajectories. I. Theory and simulation. *Biological Cybernetics*, 56(2-3), 159–175.

Saxberg, B. V. H. (1987b). Projected free fall trajectories. II. Human experiments. *Biological Cybernetics*, 56(2-3), 177–184.

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

Scarantino, A. (2015). Information as a probabilistic difference maker. *Australasian Journal of Philosophy*, 93(3), 419-443.

Schulz, A. W. (2018). *Efficient cognition: The evolution of representational decision making*. Cambridge, MA: MIT Press.

Schwartz, P. H. (2004). An alternative to conceptual analysis in the function debate. *The Monist*, 87(1), 136-153.

Schweizer, P. (2017). Cognitive computation sans representation. In T. Powers (Ed.), *Philosophy and computing: Essays in epistemology, philosophy of mind, logic, and ethics* (pp. 65-84). Cham, Switzerland: Springer.

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(3), 393–416.

Shagrir, O. (2012). Structural representation and the brain. *British Journal for the Philosophy of Science*. 63(3), 519-545.

Shagrir, O. (forthcoming). In defense of the semantic view of computation. To appear in *Synthese*.

Shea, N. (2013). Naturalising representational content. *Philosophy Compass*, 8(5), 496-509.

Shea, N. (2014). Exploitable isomorphism and structural representation. *Proceedings of the Aristotelian Society,* 112(2 pt. 2), 123-144.

Shea, N. (2018). *Representation in cognitive science*. Oxford: Oxford University Press.

Shores, C. C. (1985). Propositional attitudes and the language of thought. (Unpublished doctoral dissertation). University of Washington, Washington, Seattle.

Sima, J. F., Schultheis, H., & Barkowsky, T. (2013). Differences between spatial and visual mental representations. *Frontiers in Psychology*, 4(240). Online. doi: 10.3389/fpsyg.2013.00240

Sodian, B. (2011). Theory of mind in infancy. *Child Development Perspectives*, 5(1), 39-43.

Sprevak, M. (2010). Computation, individuation, and the received view on representation. *Studies in History and Philosophy of Science*, 41(3), 260–270.

Sprevak, M. (2011). Representation reconsidered by William M. Ramsey. *British Journal for the Philosophy of Science*, 62(3), 669-675.

Sprevak, M. (2013). Fictionalism about neural representations. *The Monist*, 96(4), 539-560.

Steiner, P. (2010). The bounds of representation: A non-representationalist use of the resources of the model of extended cognition. *Pragmatics & Cognition*, 18(2), 235-272.

Stich, S. (1983). *From folk psychology to cognitive science*, Cambridge, MA: MIT Press.

Strasser, A. (2010). A functional view toward cognitive representations. In D. Ifenthaler, P. Pirnay-Dummer, & N. M Seel (Eds.), *Computer-based diagnostics and systematic analysis of knowledge* (pp. 15-25). New York: Springer.

Swoyer, C. (1991). Structural representation and surrogative reasoning. *Synthese*, 87(3), 449-508.

Thompson, E. (2010). *Mind in life: Biology, phenomenology, and the sciences of mind*. Cambridge, MA: Harvard University Press.

Thomson, E., & Piccinini, G. (2018). Neural representations observed. *Minds and Machines*, 28(1), 191-235.

Thrun, S. (2002). Robotic mapping: A survey. In G. Lakemeyer, & B. Nebel (Eds.), *Exploring artificial intelligence in the new millennium* (pp. 1-35). San Francisco: Morgan Kaufman.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, 55(4), 189-208.

Tonneau, F. (2012). Metaphor and truth: A review of Representation Reconsidered by W. M Ramsey. *Behavior and Philosophy*, 39/40, 331-343.

Turing, A. M. (1936/2004). On computable numbers, with an application to the *entscheidungsproblem*. In B. J, Copeland (Ed.), *The essential Turing* (pp. 58–90). Oxford University Press: Oxford.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 49, 433-460.

van Gelder, T. (1995). What might cognition be, if not computation?. *The Journal of Philosophy*, 92(7), 345-381.

Varela, F. J., Thompson, E., & Rosch, E. (2016). *The embodied mind: Cognitive science and human experience* (Rev. ed.). Cambridge, MA: MIT Press.

Villalobos, M., & Dewhurst, J. (2017). Why post-cognitivism does not (necessarily) entail anti-computationalism. *Adaptive Behavior*. Online. doi: 10.1177/1059712317710496

Villalobos, M., & Dewhurst, J. (2018). Enactive autonomy in computational systems. *Synthese*, 195(5), 1891-1908.

von Eckardt, B. (1993). *What is cognitive science?* Cambridge, MA: MIT Press.

Ward, D., Silverman, D., & Villalobos, M. (2017). Introduction: The varieties of enactivism. *Topoi*, 36(3), 365– 375.

Wexler, B. E. (2006). *Brain and culture: Neurobiology, ideology, and social change*. Cambridge, MA: MIT Press.

Wheeler, M. (2005). *Reconstructing the cognitive world: The next step*. Cambridge, MA: MIT Press.

Williams, C. (1966). *Adaptation and natural selection*. Princeton, NJ: Princeton University Press.

Wilson, R. A. (2004). *Boundaries of the mind: The individual in the fragile sciences: Cognition*. Cambridge: Cambridge University Press.

Wilson, A., & Golonka, S. (2013). Embodied cognition is not what you think it is. *Frontiers in Psychology,* 4. Online. doi: 10.3389/fpsyg.2013.00058

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

Zawidzki, T. W. (2011). How to interpret infant socio-cognitive competence. *Review of Philosophy and Psychology*, 2(3), 483-497.

Zednik, C. (2011). The nature of dynamical explanation. *Philosophy of Science*, 78(2), 238-263.

Zhao, J., & Cakal, S., & Yu, R. (submitted). Statistical regularities merge object representation. Available on request from http://zhaolab.psych.ubc.ca/publications.html