Getting their Acts Together

A Coordinated Systems Approach to Extended Cognition

Submitted by **Richard Sims** to the University of Exeter as a thesis for the degree of Doctor of Philosophy in Philosophy. October 2021.

I certify that all material in this thesis which is not my own work has been identified and that any material that has previously been submitted and approved for the award of a degree by this or any other University has been acknowledged.

(signature) ____

Abstract

A cognitive system is a set of processes responsible for intelligent behaviour. This thesis is an attempt to answer the question: how can cognitive systems be demarcated; that is, what criterion can be used to decide where to draw the boundary of the system? This question is important because it is one way of couching the hypothesis of extended cognition – is it possible for cognitive systems to transcend the boundary of the brain or body of an organism? Such a criterion can be supplied by what is called in the literature a ‘mark of the cognitive’.

The main task of this thesis is to develop a general mark of the cognitive. The starting point is that a system responsible for intelligent behaviour is a coordinated coalition of processes. This account proposes a set of functional conditions for coordination. These conditions can then be used as a sufficient condition for membership of a cognitive system. In certain circumstances, they assert that a given process plays a coordination role in the system and is therefore part of the system. The controversy in the extended cognition debate surrounds positive claims of systemhood concerning ‘external’ processes so a sufficient condition will help settle some of these debates.

I argue that a Coordinated Systems Approach like this will help to move the extended cognition debate forward from its current impasse. Moreover, the application of the approach to social systems and stygmergic systems - systems where current processes are coordinated partly by the trace of previous action – promises new directions for research.

[287 words]

[Whole submission 98189 words without footnotes, references and abstract]

Table of Contents

Contents

List of Illustrations

List of Abbreviations

AA	Adams and Aizawa (2001, 2008, 2010a, 2010b; Aizawa and Adams, 2005)
AOC	Area of contention
AFCS	Automatic Flight Control System
CC	Clark and Chalmers (1998)
CIC	Content-involving cognition
CNC	Canonical neural computation (Carandini and Heeger, 2012)
COC	Coupled oscillatory containment strategy in the sheep herding experiment.
CRC	Continuous reciprocal causation
CSA	Coordinated Systems Approach (defended in this thesis)
D-cog	Distributed cognition
EXT	Component of system external to the organism
GCA	Goal contribution account of normative function
GPT	General process theory of Johanna Seibt (2018)
HEC	Hypothesis of extended cognition
HEM	Hypothesis of extended mind
HEMC	Hypothesis of embedded cognition

INT	Component of system internal to the organism
LTP	Long term potentiation involved in synaptic plasticity
MOMA	New York Museum of Modern Art
MR	Multiple realisability
MUMA	Mutual manipulability criterion for constitutive relevance
NMDA	N-methyl-D-aspartate – type of glutamate receptor mediating LTP
PEA	Principle of ecological assembly
PP	Parity principle
S&R	Search and recover strategy in the sheep herding experiment.
SAC	Starburst amacrine cell
SIF	Skilled intentionality framework (Rietveld and Kiverstein, 2014)
TSD	Task specific device (Bingham, 1988)

Acknowledgements

It takes a village to raise a PhD and I want to thank many tribes.

Firstly, the supervisor tribe: thanks for your patience, support, and careful reading of my work. Many thanks go to my lead supervisor Adam Toon who bore my manifold errors and missteps with equanimity, humour, and patience, and whose clear thinking and writing has been an inspiration. I thank my second supervisor John Dupre for introducing me to process ontology and for providing a regulative ideal towards which my writings can only aspire. The errors that remain are mine only.

Although not a supervisor, Marc Slors deserves a special mention for the many interesting and inspiring conversations about social cognition and everything else under the sun, and for helping in the organisation of the Philosophy of Coordination conference.

There is a big tribe who in conversation or more formal settings have helped shape my ideas over the last 5 years. They include (in no particular order) Giovanna Colombetti, Brian Rappert, Richard Menary, Beate Krickel, Tom Roberts, Sabina Leonelli, Joel Krueger, Denis Walsh, Kim Devereux, Achim Stephan, Sam Wilkinson, Dan Nicholson, Sabina Leonelli, Fiona Macpherson, Marcin Milkowski, Nicholas Shea, David Mitchell, Shaun Gallagher, Sarah Awad, Kate Nave, Jane Elliot, Anna Dadaian, Ian Kidd, Erman Sozudogru, Andy Pickering, Hasok Chang, Niccolo Tempini, Adrian Currie, Gemma Anderson, Sophie Gerber, Johannes Jaegher, Robert Meunier, William Bechtel, Ana-Maria Cretu, Dali Gamsakhurdashvili, Basil Wahn, Özlem Yilmaz, and Astrid Schrader.

I want to thank the Exeter tribe, especially Susan Kelly, Joel Krueger, and Kerstin Walsh for being great pastoral tutors, and my fellow PhDs for support, advice, and coffee: Javier Saurez, Thomas Bonin, Benjamin Smart, Maria Dede, Andy

Jones, Elis Jones, Jack Griffiths, Aimee Middlemiss, Flavia Fabris, Jaanika Puusalu, Elena Sherratt, Caglar Karaca, Necmiye Akcali, and Lucy Osler. Special thanks go to Chee Wong for being a warmly welcoming source of order and wisdom.

Thanks to a great tribe of correspondents for sharing their work and helping me to understand it better: Dave Chalmers, Mike Wheeler, Marco Viola, Mark Sprevak, Mazviita Chirimuuta, Wayne Christensen, Georg Theiner, Pierre Steiner, Lenny Moss, Francis Heylighen, Fred Keijzer, and Vincente Raja Galian.

On the home front I would like to thank Oskar, Ove, Patrik, Micke, and Karin in *Some Funk Junk* for putting up with the lack of practice of their keyboard player during the writing period. I shall work on my chops from now on, promise!

Special thanks go to my family tribe: David and Sylvia Sims for nurturing my interest in philosophy (and pretty much everything else) from the start, Anne Castling for taking an interest over the years, and my siblings Liz, Katy, Alex, and John for lively philosophical conversations over the course of half a century.

Special mention for my philosophical, mathematical, linguistic and musical friends and teachers: Marc, Vincent, Chris, Paul, Julian, Sue, Eric, Colin, Martin, Louis, Nina, Pam, Barbara, Dai, Nada, Janet, Sharon, Antony, James, Karsten, Maria, Gilius, Christina, Beate, Richard, Mauritz, Per, Jo, and Jaroslav for interesting conversations over the years.

I would like to thank my daughters Pashmina and Rebecca both for putting up with philosophical conversations over the dinner table and for sharing a passion for thought and puns. Finally, I want to thank my wonderful wife Caroline who has accompanied me on this journey and has generously shared her own intellectual exploration with me.

Introduction

> Our effective environment is a shifting coalition of resources and constraints, some physical, some social, some cultural, some computational (involving internal and external resources), When this shifting coalition of resources is appropriately coordinated, the tasks we set out to achieve are accomplished (Kirsh, 1999, p. 2).

This is a thesis about extended cognitive systems. A cognitive system is a set of processes arranged in such a way as to be responsible for intelligent behaviour. What makes these systems extended is that they extend beyond the boundaries of the neural architecture of the brain and even beyond the skin boundary of the organism. Supposing that extended cognition is both conceptually possible, and in fact observed in the world, is called the *hypothesis of extended cognition* (HEC). This thesis attempts to answer the question: what criteria can be used to draw a boundary around such a cognitive system and thereby provide support for a version of the HEC?

The question of HEC concerns a group of increasingly influential positions in philosophy of mind and philosophy of cognitive science known as 4E cognition: extended, embodied, enactive, and embedded cognition[1]. 4E approaches bear some sort of family resemblance but make quite distinct (even contradictory) claims and commitments in the debate[2]. While the focus of the argument will be on extended cognition, other E's will make an occasional appearance. A slightly different notion discussed later is the notion of *distributed* cognition – where a cognitive system extends not only over material things in the environment but

[1] The term 4E is traceable to a workshop at Cardiff University in July 2006 (Newen *et al.*, 2018, p. 4 fn).

[2] See, for example, Kiverstein and Clark (2009) for a good overview.

also over other individuals. Hence there will be moments in the work when I want to add a ‘D’ to the 4 E’s (see Stephan, 2018).

It has been written that, after a quarter of a century, to all intents and purposes, the extended cognition debate has reached a stalemate (see Sprevak, 2010). There is a sense that the positions in the debate have become entrenched and that the main protagonists are talking past each other. At the same time, extended cognition continues to excite research in the community both within philosophy and cognitive science and beyond, judging from the frequency of citations of the field-defining paper by Andy Clark and David Chalmers (1998 henceforth ‘CC’). The issues identified in this thesis are still live; extended cognition and 4E approaches in general are as influential as ever (see also Gallagher, 2018a). Stalemate in a live research area seems unsatisfactory, and it is a matter of importance to try to cast light on some of these questions in order to move the debate on.

That said, the problems are substantial, and it is well-travelled theoretical paths that have led to the impasse. What is needed is a new approach that learns from and distils the insights of existing accounts while avoiding their pitfalls. The aim of this thesis is to develop such an approach to answer the question: ‘what criteria can be used to demarcate cognitive systems?’ The thinking here is that demarcation criteria can then establish whether a given component belongs to the cognitive system and can therefore give support (or not) to putative cases of extended cognition.

This thesis takes up the challenge of developing a new approach with its own theoretical machinery to move the debate forward, and to suggest new lines of inquiry – this forms the substance of part II of the thesis. The success of this project depends on whether it can resolve certain sticking points in the debate that will be set out in part I. Part III applies the new machinery to these questions and comes out broadly in support of a suitably modified version of the HEC.

One of the issues, it seems to me, is that, in their drive to refinement and to say important and significant things about the nature of cognition, current approaches have been too ambitious. Indeed, in many cases they have baked rather too much in at the start, such as a commitment to a specific theory of cognition, understanding of cognitive processes, and set of paradigm cases. The result has been sophisticated theorising that leads to interesting conclusions yet remains discursively disconnected with other work that does not share the same premises. Moreover, there is a worry that these starting points are question-begging with regard to the HEC and help themselves to “proprietary cognitive criteria” (Kaplan, 2012, p. 548). In some cases, it is difficult to imagine what else can be said in the debate except ‘OK, but I do not share your starting point’. There seems little prospect of progress when the debate issues from different, and contradictory, theory-loaded starting points. For example, there is unlikely to be much common ground on the extended cognition question between (a) a position that cognition is the manipulation of neural representations and (b) that cognition is constituted by a dynamic interaction between an organism and its environment. What is interesting is if there is an understanding of cognition that makes a minimal initial commitment that would be acceptable to different players in the argument, that can make progress in the extended cognition debate.

On the other hand, cognitive agnosticism has a price. Adopting too thin a concept of cognition risks being insufficient to get a grip on the demarcation problem. In an ideal world drawing a boundary requires that there is a property that is true of processes on one side of it and not the other. I examine one such argument in chapter 2 where a boundary criterion is founded upon mechanistic notions alone without a reference to cognitive properties at all. Not taking *any* cognitive attributes into account is not likely to be a good starting point for drawing a boundary around *cognitive* systems – in fact to be able to do this at all means that a notion of cognition is smuggled in somewhere (and in my view it is).

This thesis proposes an approach that navigates a thin line between the Scylla of a heavily theory-loaded, and arguably question-begging account, and the Charybdis of not having enough grip on cognition to be useful or productive in the search for a demarcation criterion. I call it a *Coordinated Systems Approach* or CSA. It is a systems approach in the sense that it conceives a cognitive system to be a coalition of more or less autonomous processes, independent of where they are or how they are constituted, thus avoiding neurological or anthropocentric (or indeed zoocentric) bias. By taking a cognitive system to be responsible for intelligent behaviour, it commits to a specific characterisation of cognition but one that is general enough to be potentially acceptable to both sides of the debate. The new approach is, therefore, orthogonal to current positions in the debate yet learns from their insights – and I maintain that insights are to be had from all sides.

Not being too specific about cognition, and basing the argument on relatively uncontroversial premises, risks that it results in a framework that does not say very insightful or refined things about cognitive systems. But I am hoping to capture the key features: the first things one might notice about any system that promises to be cognitive, and that the method works across a wide range of systems and time and space scales. It is more likely to be general features that will identify a wide class of systems, than refined ones.

The term ‘system’ is used primarily in an epistemological sense in this thesis albeit with some metaphysical implications. A system is a collection of connected processes that provides the explanatory resources for understanding a set of cognitive phenomena. The idea is that to understand intelligent behaviour in the world we take the system as being what is required for the explanation. To explain a phenomenon, one should be able to draw upon the resources provided by the system and it should not be necessary to refer to significant items outside it. Such a system is explanatorily self-sufficient. For example, if we want to understand the different behaviours of a bacterium in two situations in which a sugar gradient is present or absent, we might take the boundary of the organism itself as a first approximation to delimiting the system of interest. Later in the investigation we might discover that it is a subsystem of the organism that allows us to explain the requisite suite of behaviours.

Later in the thesis we shall tackle the objection that taking a system to be ‘merely’ an explanatory resource amounts to an instrumentalism regarding systems. My response is inspired by David Chalmers’ assertion about common sense psychology that “the classification of [mental states] can depend on our explanatory purposes” (2011, p. xii); the same mind can be extended or not depending on the explanatory context. Indeed, Andy Clark himself suggests that one could flip back and forth between such views like a Neckar cube (2011a, Chapter 9). I attempt to make this idea more precise by framing it in terms of the set of tasks towards which the behaviour of the system is directed. What counts as the system depends on which set of tasks we have in view. This is not to say that there are no systems, or system boundaries in the world, but rather, to paraphrase John Dupre[3], that there are an awful lot of them. Explanatory interests pick out the relevant boundary relative to the appropriate set of tasks. Once the task set is fixed by the investigation then the system responsible for their performance – the production of goal-directedness – is fixed by the requirement of that it possess a certain kind of functional coherence; it is somewhat more than a propitious collection of elements that happens to explain a phenomenon.

In this respect there are strong links with the cybernetics and complex system theory literature. But unlike many writings in these traditions, the view taken here of systems as coalitions of processes suggests the master metaphor of ‘flow’ – a river or ants’ nest, rather than a machine with fixed components. Indeed, I

[3] Private communication.

contend that it is implicit reliance on a machine metaphor that gets some of the existing approaches into trouble.

Some might regard the introduction of goal-directedness as rather too bold and already breaking the promise regarding running the argument from non-controversial premises. Science, we are told, despises teleology, and seeks to explain the world without it. This may be true of physics and chemistry; and only if we ignore thermodynamics and those systems that organise themselves to avoid its effects[4]. But it is exactly these systems, such as biological organisms studied by the special sciences, that are likely, in one way or another, to be cognitive. To these people, I say, look around you at the myriad cognitive systems and note what they have in common – they are all exhibiting behaviour, that is, they are producing goal-directed action. They are not acting according to chance, neither are they acting like rocks and rivers just going with the thermodynamic flow. Behaviour really is distinguishable from chance, and from inert entropic processes, because it possesses a special modal profile. As the philosopher of biology Denis Walsh points out, purposive events are robust across range of different initial conditions and mechanisms, while chance events are not (2015a, p. 193). We can place the mouse on different points on a slope and it will go for the cheese independent of where it starts out. We can place a ball on the slope, and its path depends critically on the starting point. As in cybernetics, and complex system theory that is its inspiration, goal-directedness is a key idea in this thesis. The task then is to find a mode of description of systems that captures the basic character of their goal directedness, but one that is sufficiently powerful to answer the central questions in the extended cognition debate. It is another balancing act. The CSA is an attempt to do just this.

[4] See the classic work on the role of thermodynamics in modern science by Ilya Prigogine and Isabelle Stengers (1984).

Since this thesis aims to establish the HEC, its implications are intimately bound up with those of extended cognition more generally and it is answerable to the same questions. For example, why does it matter whether the HEC is true? Of course, there are the immediate implications in philosophy of mind and cognitive science. If the HEC is true, then pursuing empirical work in psychology solely in a white-room environment may not be the best course of action. The use of machine learning and computational models in cognitive or computational neuroscience may need to extend to systems including bodily, environmental, social, and cultural resources.

But the debate also touches areas of philosophy outside philosophy of mind and cognitive science. Whether one takes an extended mind perspective or not may well determine what one regards as salient in terms of norms in ethical or moral judgments. For example, if part of the environment counts as part of the cognitive apparatus of the individual, interference with that part may well carry the same ethical or moral significance as interfering with a person’s neural structures. The question of whether to move a person with Alzheimer’s from her home is surely bound up with the question of whether that home should be regarded as constituting part of her cognitive system required for successful day-to-day functioning. It does matter whether one takes an extended perspective or not (see Clark, 2017).

Regarding the specific contribution of the CSA, there may be interesting consequences across many different fields. In pedagogy, for example, there are questions about the interactive role of objects, architectural space and learners in a classroom (see for example Barab and Plucker, 2002). Since the CSA does not make a principled distinction between systems involving a single agent or multiple agents there are circumstances where a classroom and its occupants may be regarded as a single cognitive system. Implications follow for classroom organisation, and indeed for the architecture of schools. There are clearly parallels here with a host of other social, financial, and legal institutions.

In 4E research, the CSA casts light on problems that are perhaps intractable using the explanatory tools of mechanisms. Notable amongst these are questions concerning loosely bound cases of social cognition in, for example, institutions like the legal system or everyday situations like supermarket shopping. This work ties in with that of Marc Slors on symbiotic cognition (2019, forthcoming a) and suggests new questions and avenues for research.

Finally, an important avenue for more research is the realisation that many of the interesting systems discussed in this thesis fall under the general heading of stygmergic systems – systems whose future actions are coordinated by the results or traces of previous performances. These not only include systems found in nature such as eusocial insects and flock behaviour, but also swarm robotics, engineering applications and wider systems like markets in economics. The application of CSA to these areas is touched upon in part III of this thesis but the area is ripe for further research, not least because so many systems in the world are stygmergic.

In this connection, it is precisely because the CSA furnishes such a general criterion for cognition untethered to a human, or zoocentric paradigm, that it can be used to investigate the cognitive capacities of plants – an exciting and relatively new field spanning plant biology and cognitive science. Such an investigation is beyond the scope of this work but, as I state in the conclusion, is a fertile area for further work.

The thesis is divided into three main parts. Part I is a review of the main arguments in the literature for the HEC and an examination of the sticking points in these arguments. Part II consists of the building of the theoretical machinery the CSA that is the main product of the thesis. Part III sees the new machinery being put to practical use in terms of the main examples and arguments in the debate.

While there are precursors to many of the arguments in this thesis in the literature, the entirety of the construction – the way the puzzle pieces are put together – is, to the best of my knowledge, a new contribution to the debate and my hope is that it does indeed break the impasse.

Ric Sims, September 2021, Gävle, Sweden.

Chapter 1

Extended Cognition

> The truth is, the Science of Nature has been already too long made only a work of the Brain and the Fancy: It is now high time that it should return to the plainness and soundness of Observations on material and obvious things. (Hooke, 1665, p. 13)

1.1 Preliminaries

In this chapter I shall focus the discussion on the hypothesis of extended cognition (HEC). This amounts to the claim, roughly speaking, that cognition may be constituted by processes that run over environmental artefacts and structures as well as neural ones. Built into the extended viewpoint, at least implicitly, is the idea that this is not just a conceptual possibility, but that there are actual cases of extended cognition in the world. Establishing the credentials of the HEC therefore consists in both doing conceptual work in showing that there is no inherent contradiction in positing an extended cognitive process, but also analysing real or imagined examples and showing that, on occasion, they are found in the world. Some authors go further and claim that cognition is ordinarily or typically extended, and that the extended view properly captures the character of cognitive processes and gives a new and productive theoretical background to existing data in cognitive science and psychology (Theiner, 2011, p. 25). While the findings of this thesis may lend some weight to the radical view, I shall take as a starting point the more modest proposal that it is possible that cognitive processes are extended and that, in some cases, they are.

The main aim of the chapter is to present the relevant aspects of the debate as I see it currently, and as far as it is relevant to the research question. Since the aim of this research is to provide a framework in which to tackle some of the problems of existing accounts, I offer apologies for focussing on these problems and giving less space to aspects of existing theory that work well (such as elements of so-called second- or third-wave theories). These will, in some cases, serve as a jumping off point for the theoretical developments in the rest of the thesis and I shall come back to them in part III of the thesis.

I propose to tackle the task of this chapter in four stages. The first is to understand the substance of the main claims for extended cognition and the arguments for them in what is regarded by many as the canonical text CC (Clark and Chalmers, 1998). The second stage discusses objections that are levelled, not at the specific arguments of the paper itself, but perhaps more at the *kinds* of arguments that CC use. These are methodological arguments about how we should go about analysing cognitive systems. These objections are potentially more problematic for the defender of HEC because they threaten the very methodological foundations of the theory. Having done this preliminary work, it is useful to group the questions underlying the discussion into *Areas of Contention* (AoC) that indicate the key areas of disagreement. How one responds to these AoC defines, roughly speaking, one's 4E position. The AoC can be used to orient the debate and indicate the likely obstacles to a successful resolution of the demarcation problem.

Finally, at the end of the chapter I shall give the barest outline of the track of the argument through the whole thesis. At this point it is hoped that the reader will have a general idea of the nature of the problems to be overcome and some rough picture of the strategy to be followed.

Before I discuss the layout of the chapter in more detail, a brief word about the term *cognition* is in order. One of the key areas of contention is the question of what theory of cognition should fix the reference of cognitive terms in the argument. As mentioned in the introduction, this is potentially a question-begging matter, because a background theory of cognition can skew the debate one way or another. If one side insists that cognition takes place on a neural substrate, for example, then extended cognition arguments do not get off the ground. Therefore, a definition of cognition is one of the variables in this work. Nonetheless as a rough and ready working definition to get the investigation started, I shall take cognition to be a characteristic of a system able to produce intelligent, that is, goal-directed behaviour. This is not so distant from the definition given in standard cognitive science textbooks. Here, for example is a list of attributes of a cognitive system taken from David Vernon's textbook (2014, p. 8). A cognitive system…

1. is autonomous.
2. acts to pursue goals.
3. adapts to changing circumstances.
4. learns from experience.
5. anticipates outcomes.

This focus of this thesis is on attribute (2) that the system acts to pursue goals and can change its actions in response to circumstances. In Chapter 4 this will be understood to include (3) adapting to changing circumstances. In acting, the system possesses a repertoire of behavioural responses that can, in the interesting cases, be extended and made to operate more efficiently through *learning* (4) broadly construed. I do not directly address autonomy (1) or anticipating outcomes (5). Moreover, I make no assumptions that the existence of phenomenal consciousness or qualia are essential conditions on a state or process being cognitive.

Now to the layout of this chapter. The canonical exposition of the extended mind hypothesis CC is discussed in section 1. The responses it provoked, the most well-known of which are due to Fred Adams and Ken Aizawa (henceforth collectively AA), are outlined in section 2. AA's objections to the HEC are serious, especially the worry about so-called *cognitive bloat* – the extending of the system unreasonably far out into the world. However, the strongest arguments against

HEC come from views of cognition that share the view that environmental resources may fulfil an important role in cognition but nonetheless hold that such resources remain external to the cognitive system. These methodological objections are discussed in the second half of the chapter. In this vein, section 3 is devoted to the argument of Robert Rupert that any explanation of a cognitive phenomenon via an extended system can be translated into one in an embedded (non-extended) context without loss of explanatory power. Equally worrying from an extended mind perspective is Mark Sprevak's argument that the method of *inference to the best explanation* (IBE)*,* used standardly in the natural sciences, both supports the extended mind argument and its main rival Rupert's theory of embedded cognition in equal measure prompting some commentators to pronounce the debate a stalemate. This is discussed in section 4.

Section 5 pulls the discussion together and proposes that it can be summarized by five independent areas of contention and section 6 sketches the bare outline of the strategy taken in this thesis to answer questions underlying the areas of contention.

1.2 The hypothesis of extended cognition

The idea of extended cognition has precedents in the literature. Nonetheless, it is fair to say that Andy Clark and David Chalmers' *Analysis* paper CC remains the jumping-off point for many of the arguments in the field[5]. Indeed, 'linking to

[5] Merlin Donald's Origins of the Modern Mind (1991), Edwin Hutchins' Cognition in the Wild (1995a), and the work of David Kirsh (1995, 1999; Kirsh and Maglio, 1994) are sometimes regarded as close precursors, but CC is widely regarded as starting the debate in earnest (cited 2232 times in the web of science database as of Aug 2021), and has spawned a series of sympathetic refinements and critical commentaries that are field-defining. More distant precursors are Vygotsky and his notion of external scaffolding and proximal development (1978), Heidegger (1962), and Dewey (1958, 1998) and their notions of the world-involving nature of cognition.

Clark and Chalmers' is often a methodological first step for establishing a new or differently-nuanced line of enquiry in 4E cognition (see, for example, Colombetti and Roberts, 2015; Gallagher and Crisafi, 2009; Gertler, 2007; Miyazono, 2017; Rowlands, 2009; Wheeler, 2011). On the other side of the coin, it is also the starting point for arguments that reject 4E claims (see, for example, Adams, F. and Aizawa, 2001, 2008, 2010a; Rupert, 2009a, 2010a, 2010b, 2011, 2012, 2016; Walter, S., 2010; Weiskopf, 2008, 2010). This thesis follows the precedent by examining the main examples in the paper and the arguments that they illustrate.

CC aim to answer the question "where does the mind stop and the rest of the world begin?" (1998, p. 7). The first half of their paper focuses on two main points: the possibility that cognitive processes may take place in the world outside the skin of the agent and how such processes might be identified.

Clark and Chalmers express this idea in the following manner:

> [Under certain conditions] the human organism is linked with an external entity in a two-way interaction, creating a *coupled system* that can be seen as a cognitive system in its own right. All the components in the system play an active causal role, and they jointly govern behaviour in the same sort of way that cognition usually does. If we remove the external component the system's behavioural competence will drop, just as it would if we removed part of its brain. Our thesis is that this sort of coupled process counts equally well as a cognitive process, whether or not it is wholly in the head. (1998, pp. 8–9 emphasis original).

I read this as amounting to the conjunction of four claims:

System claim: In the right conditions, an organism causally coupled to an external entity constitutes a cognitive system.

Functionalism claim: The causal linkage is active in such a system, and taken together, the roles played by these causal linkages govern behaviour in a functionally similar manner to what we normally consider to be cognition.

Manipulability claim: A test for the external component being part of the system is that its removal will result in lower behavioural competence.

Location claim: The coupled system counts as a cognitive system whether or not it is wholly in the head.

I shall take the conjunction of these four claims as constituting the canonical Hypothesis of Extended Cognition (HEC). It is this hypothesis that I shall investigate in this thesis and a version of which I shall end up defending. At the end of chapter 6 I shall return to these claims and show how the CSA supports a modified version of them.

It is fair to say that at first HEC was not well received. The original CC paper was rejected many times over a period of years before it was finally accepted. And we can see why. HEC is profoundly radical by the lights of traditional cognitive science and philosophy of mind.

Broadly speaking, much philosophy of mind and cognitive science, at the end of the millennium and to some extent today, is internalist. The mind is taken to be something that depends in a metaphysically essential way, or *supervenes*, on the brain - a physical organ inside the head of the organism. If cognition is essentially mental, and the mental is internal then cognition is essentially internal. As Mark Rowlands puts it, traditional Cartesian internalist cognition is the view that the machinery for cognition is located within the subject, and that its possession does not logically depend on any feature external to the subject (2003, pp. 12–18). This violates all four of the HEC's claims as stated in CC; the HEC is simply incompatible with Cartesian internalism. Rowlands suggests that Cartesian internalism underwrites the dominant role of representations in cognitive science through two claims:

(1) Mental representations are structures instantiated in the brains of cognising animals - structures that make claims on the world.

(2) Cognitive processes consist in the application of transformational rules to mental representations. (Rowlands, 2010, p. 30).

Since mental representations are internal and cognitive processes consist in the transformation of mental representations then cognitive processes are physically located inside the organism (and most likely in the brain).

Positions such as this challenge the HEC. Friends of the HEC must provide arguments showing that (1) and (2) do not adequately capture the nature of cognition. To see how they do this let us look at the initial moves made by CC in their article where they appeal to an intuition pump in the form of the game of Tetris.

1.2.1 Tetris and the Parity Principle

Imagine a person sitting in front of a computer screen playing Tetris. This is a computer game from the 1980's in which the player tries to fit shapes constructed from four grid squares called 'zoids' into slots in a wall of squared patterns. The aim of the game is to try to fit the zoid as efficiently as possible and not leave any holes. This game was brought to the attention of philosophers of cognitive science by David Kirsh and Paul Maglio in their influential 1994 paper. They observe that competent players of the game surprisingly engage in extra rotations of the pieces prior to fitting them into the wall, even though this manipulation cost valuable time (1994). Kirsh and Maglio hypothesise that being able to rotate the shapes is an advantage in playing Tetris. CC build on this observation and invite us to consider three cases: in the first the shapes can be rotated in the subject's imagination, in the second the shapes can be rotated on the screen by pressing a key on the keyboard, and in the third the subject's brain is fitted with a plug-in hardware module that performs the rotation operation. CC ask the question how much cognition is involved in each case. They claim that, to all intents and purposes, the three cases are similar. Whether the rotation is performed in the head, in the hardware module, or on the screen, the contribution to cognition is the same. The intention is to leverage our intuition that, in the case of mental rotation, the cognitive system resides in the brain of the subject, while in the case where it is performed on the screen, it extends to encompass the computer and the screen. The situation involving the plug-in module is an intermediate case. CC borrow the term *epistemic action* from Kirsh and Maglio to describe the rotation operation performed by the player – an action that is not strictly required for the game but that seems to play a cognitive role. CC consider the action of rotation to be part of the cognitive processing of the player. They suggest that such epistemic actions are commonplace and give similar examples in the literature: the use of pencil and paper in a long multiplication task (Rumelhart and McClelland, 1986), the rearranging of Scrabble tiles to aid word recall (Kirsh, 1995) or use of a nautical slide rule in navigation (Hutchins, 1995a). Kirsh and Maglio take the player to be offloading a cognitive task on the environment, but CC take the more radical view that the player recruits the environment into the system. This difference in terminology marks a crucial theoretical distinction underlying the discussion in this chapter.

What permits the inference to the claim that epistemic actions are part of the cognitive system rather than just an external tool used by the system? This is licensed in CC by an appeal in the 'functionalism claim' of HEC to a principle of functional equivalence. CC call it the *parity principle*:

> Epistemic action, we suggest, demands the spread of *epistemic* credit. If, as we confront some task, a part of the world functions as a process which, *were it done in the head*, we would have no hesitation in recognising as part of the cognitive process, then that part of the world *is* (so we claim) part of the cognitive process (1998, p. 8 emphasis original).

Through the Parity Principle, CC establish a standard for individuating parts of cognitive systems - by their overall function rather than by the details their specific implementation. A process is cognitive if it performs the right sort of function. And how do we know what the right sort of function is? We compare it with a process that, were it in the head, we would decide that it was a non-controversial case of cognition. For example, if a process performed the cognitive action of multiplying two numbers together and we would regard an in-head performance of this task as cognitive, then the process is cognitive even if it involves external elements such as pencil and paper. Applying it to the Tetris examples, the rotation on the screen in the second case as cognitive since we would plausibly regard the functionally equivalent mental operation as being cognitive. The same applies to the silicon plugin in the third case. The point is that it shouldn't matter where such processes take place.

However innocent it may seem, the Parity Principle embodies a lot of theoretical choices, and in many ways is key to the argument in CC. Its implications are monumental – as CC put it later – we should see the rearranging of scrabble tiles on the tray not as an action but as a thought (1998, p. 10). Because of its centrality, the theoretical justification of the Parity Principle constitutes grounds for potential objections to the HEC. For example, its use demonstrates a wholesale commitment to a type of extended functionalism (see Wheeler, 2010); describing cognitive processes at the level of their functional roles is the right way to go rather than, say, describing the details of their mechanistic implementation. Granted, functionalism was the dominant theoretical framework for cognitive science in the latter part of the 20th century and perhaps still is, but not all players in the debate necessarily accept the jump to the extended version. I shall say more about the parity principle in the next section.

1.2.2 Otto and Inga

The second half of CC takes a slightly different tack than the first. Instead of addressing the question of what aspects of the environment belong to a cognitive system they ask whether structures in the environment could be vehicles for mental states (see Hurley, 2010, p. 102). A *vehicle* is a syntactic structure that carries representational *content*. So written language is a vehicle, the word ‘dog’ is a physical state – an inscription that because of its physical properties stands for something – in this case a four-legged animal. This is why the position taken in CC is often referred to as *vehicle* externalism – it is the vehicle that is external to the organism, rather than the content externalism of Putnam and Burge (Burge, 1986; Putnam, 1973). CC concoct the famous Otto-Inga thought experiment to make plausible their intuition that, indeed, such environmental vehicles might partially constitute mental states.

The case of Otto and Inga has had a lot of coverage. I shall describe it in detail because it will be useful to refer to this example in the chapters to come. This is a scenario in which two people, by hypothesis, both harbour the desire to go to an exhibition at the Museum of Modern Art in New York. The first person called Inga (‘in’-ga – she represents the internalist view) holds a dispositional belief about the location of the gallery in her biological memory. The second person, Otto (‘out’-o – he represents the externalist view), has mild Alzheimer’s and refers to a notebook that he has constantly to hand to direct him to the gallery. CC claim that the two situations are functionally equivalent with respect to belief-desire psychology. The notebook plays the same functional role in the case of Otto as biological memory plays in the case of Inga in guiding the action of going to the museum. They appeal to the Parity Principle to assert that if the functional role of the notebook is equivalent to that of Inga’s biological memory in that it is the vehicle for a representation constituting an intentional state (believing say) then it is cognitive.

Clearly, Otto walked to 53rd Street because he wanted to go to the museum, and he believed the museum was on 53rd Street. And just as Inga had her belief even before she consulted her memory, it was a dispositional belief, it seems reasonable to say that Otto believed the museum was on 53rd Street even before consulting his notebook. In relevant respects the cases are analogous: the notebook plays for Otto the same role that memory plays for Inga. “The information in the notebook functions just like the information constituting an ordinary non-occurrent belief; it just happens that this information lies beyond the skin.” (1998, p. 13).

For CC, then, the notebook and Inga’s biological memory are on a par. But since Inga’s memory contains, amongst other things, her dispositional beliefs, then Otto’s notebook constitutes some of his dispositional beliefs too. Inga accesses her memory to access her belief, while Otto accesses his notebook. That would mean that the vehicle for his beliefs was physically external to him. As Clark likes to put it in conversation and interviews: “we should not prejudice our judgments of when something is cognitive by its location” (2017).

1.2.3 Hypothesis of the extended mind (HEM)

CC are motivated to take the HEC further to form a hypothesis about mind itself that I shall call the *hypothesis of the extended mind*: If an external vehicle plays the same functional role as an internal intentional state *could* do (such as a belief or a desire) then the external vehicle can be considered part of the subject’s mental machinery or literally part of her mind.

The HEM seems to me to be a bigger commitment than that of the HEC of the previous section since it comes with more baggage. Mind has connotations of conscious thought and agency for a start. There is an inevitable comparison with the mind we are all most familiar with – our own. When Douglas Adams chose to call the supercomputer in his *Hitchhikers Guide to the Galaxy* ‘Deep Thought’ it was surely with human thought in view rather than the cognitive system of an amoeba (1996). Minds are where human-like thinking goes on – they are more impressive than mere cognitive systems. Then there is the fact that the HEM is couched in terms of a belief-desire psychology. Otto desires to go to MOMA and believes (via the notebook) that 53rd street is where he should go. The implication of the vehicle/contents distinction in this case is that these folk-psychological notions correspond to mental representations. In Inga's case they are representations whose vehicle is constituted by biological structures whereas for Otto they are constituted by the notebook.

The view of mind we have from this example is that it is something that has mental states and engages in mental processes which are something like ordered transitions between mental states. It is a complex physical entity that is composed of parts and that its states are entirely constituted by the arrangement of its parts. The HEC contribution is that some of these parts might be external to the brain or the body of the organism meaning that the same can be said of some mental states. Moreover, and this is the important bit, mind is located wherever the mental states are, and for HEM, these states, wherever they are, are mental states of a person[6].

HEM makes additional claims to those of the HEC made in the first part of the chapter. For example, it takes for granted that belief-desire psychology is the right way to fix the reference of cognitive concepts. I shall remain neutral on this question for the time being since my strategy is to try to construct a theory from a minimal set of theoretical commitments. But I note that logical structure of the

[6] Talking about mind inevitably invokes connotations of conscious phenomenal experience. I have not said anything here about consciousness and in this thesis, I intend to bracket it, not because it is not important, but because I think I can address the main questions without it.

argument in CC seems to allow for the separation of the HEC and the more theoretically loaded HEM. For example, Robert Wilson (2004) takes the position that HEC is true but rejects HEM. Of course, HEM logically entails HEC, so a rejection of HEC is also a rejection of HEM. This is the route taken by most critics of HEM. They reject HEM *and* HEC as a whole package. The next section examines these arguments.

1.3 Parity and its discontents

Since the parity principle (PP) does the heavy lifting in CC's argument, many of the objections to it implicate the PP at some point. I shall deal with issues that directly concern the PP in this section and then look at issues that concern general systemhood in the next. There are five main clusters of objections to the HEC based on the Parity Principle:

1. cognitive bloat
2. mark-of-the-cognitive
3. Otto two-step and the ultimate control argument.
4. Grain parameter problems
5. Skewed benchmarks

I shall deal with each in turn. Before I do this, I want to say a bit more about the parity principle on which they are based.

What sort of job does the PP do and how literally should we interpret it? Mark Sprevak refers to the Parity Principle as an "egalitarian equal-treatment principle" (2009, p. 505). It is a kind of common-sense guide to the sorts of things we might take as cognitive rather than being a strict formal characterization. Appropriating Rawls (1971) Clark describes it as a "veil of metabolic ignorance" (2005, p. 2 fn, 2011a, p. 77). For Rawls, the 'veil of ignorance' in his *Theory of Justice* is a way of removing our prejudices from the table in order that the reasons that we give in the argument can be a basis for justice or fairness, without knowledge of our social location. For Clark, the adjudication of cognitiveness should take place without knowledge of metabolic location. The PP is “best seen as a heuristic (a rough-and-ready tool) for identifying some plausible cases of cognitive extension” (Clark, 2011a, p. 48)[7].

There are some subtleties here. As far as I can tell no-one in the literature has remarked that as it is stands in CC, the principle is a sufficient, but not a necessary condition for a process being cognitive. Later on, we shall see that some authors use it to show that certain processes are *not* cognitive, that is, as a necessary condition.

What are we to make of the phrase “If, as we confront some task…” (1998, p. 8)? It is tempting to interpreting the PP as saying something about cognition in relation to the performance of tasks with the implicit normativity that this implies. CC do not develop this idea, but it motivates the task-based arguments of this thesis and the basic starting point that cognition is goal directed. Similarly the phrase “(…) a part of the world functions as a process (…)” (1998, p. 8), suggests that CC, at least subliminally, regard cognition as constituted by processes (rather than, say, machinery). These are hints that I would like to follow up in later chapters (especially chapters 3 and 5). Pointing out that part of the world plays a role in the system through a process is important because it heads off the misconception that systems are static entities that are somehow ‘intrinsically’ cognitive rather than cognition being a dynamic property of processes.

[7] Georg Theiner suggests that in addition to the above epistemic role it plays an additional metaphysical role in which it individuates cognitive states by comparison with known internal cognitive states (2011).

Crucially, the principle appeals to intuitions about what kinds of process we might regard as cognitive if we found them inside the head and whether we would hesitate before pronouncing them cognitive. There are subtleties here that have led to some confusion. Mike Wheeler shows that some writers on the parity principle misinterpret it to mean that the external component must have the same functional profile as an *existing* internal component (2019, p. 84). An example is AA: “(…) [CC] contend that the active causal processes that extend into the environment *are just like the ones found in intracranial cognition*” (Adams, F. and Aizawa, 2001, p. 56 quoted in Menary 2007 p. 56 emphasis added). Other commentators such as Robert Rupert (2009a, p. 32) have taken CC to task on this appeal to intuition and accused them of making cognition into a response-dependent concept; that what makes something cognitive is my reaction to the claim rather than something inherent in the system. We shall return to these objections later in the chapter.

1.3.1 Cognitive bloat

Dan Weiskopf (2010) notes that establishing the HEC is a Goldilocks problem. Set the bar for cognitive systemhood too high, and the thesis ceases to be attractive since only the usual internalist suspects qualify. Set the bar too low, and systems seep out gratuitously into the world and the HEC fails to identify the appropriate scientific kinds. The term *cognitive bloat* refers to the second possibility - the worry that the parity argument lets too much of the environment into the system for the idea of system to do useful work. AA (2001, p. 57) call this the threat of *pancognitivism* where everything becomes part of a cognitive system. If Otto’s notebook can be considered part of Otto’s cognitive system, then what about his phone or his copy of Encyclopedia Britannica? If Otto regularly accesses the internet on his phone and looks things up using the Google search engine, does that mean that Otto’s cognitive system encompasses the whole internet?

There is a sense in which this objection captures an intuition that cognitive systems are localised and contain implicitly trusted resources. If I am out in a café, then I am not in the same location as my copy of the Encyclopedia Britannica. Moreover, I cannot really count the contents of some dubious website among my beliefs if I do not trust what is written there. CC anticipate this objection and try to neutralise it by introducing further 'glue and trust' conditions that express these intuitions:

1. The notebook is a constant in Otto's life – in cases where the information in the notebook would be relevant, he will rarely take action without consulting it.
2. The information in the notebook is directly available without difficulty.
3. Upon retrieving information from the notebook, he automatically endorses it.
4. The information in the notebook has been consciously endorsed at some point in the past, and indeed is there as a consequence of this endorsement (1998, p. 17).

These conditions parallel the situation of traditional internal cognitive capacities. Inga's ability to navigate to MOMA is due to her memory about the location of the museum and is more or less reliably and directly available to her. The contents of her memory are generally automatically endorsed. These conditions remove some of the more fanciful scenarios of supposed extended mind. Maybe Otto does not carry Encyclopaedia Britannica around with him or does not endorse all the content of websites on the internet[8].

[8] There are other issues with endorsement, for example, Rupert's observation that past conscious endorsement of a notebook entry introduces a dependence on history that does not fit well with Clark and Chalmers' avowed insistence that the external features of a cognitive system are active, playing a role in the here and now, and have a direct effect on the organism and its behaviour

Cognitive bloat is such a central concern here that a mere section of discussion would not do it justice. There is a sense in which the whole thesis is an attempt at a response. It is a topic to which I shall regularly return and will test the success of the CSA project. For now, I shall point out that some commentators are satisfied that the 'glue and trust' conditions prevent the extent of the system getting out of hand, for example Shaun Gallagher (2018a). My immediate concern, and potentially more serious for the friends of HEC, and with implications for the bloat question, is the contention that the parity principle appeals to our intuitions about what kinds of processes are cognitive. It seems to imply that we already have in place a way of grounding our intuitions regarding this question. I shall look at this issue in the next subsection.

1.3.2 Mark of the cognitive

The example of Otto and Inga is intended to support HEM and relies on establishing that the Otto plus notebook system is on a par with Inga's biological capacities when it comes to the task of navigating to MOMA. HEM can be undermined by casting doubt on the Parity Principle which establishes the functional equivalence. One way of doing this is by providing an independent criterion for a state or process to be cognitive - a so-called *mark of the cognitive*. If it can be shown that there is a case where the external process does not satisfy the mark of the cognitive criterion while the internal process does then the Parity Principle is faulty and the functional parity argument for the HEC falls.

This is the line taken by AA (Adams, F., 2018; Adams, F. and Aizawa, 2001, 2010a, 2010b; Aizawa and Adams, 2005). While they offer no complete theory

(Clark and Chalmers, 1998, p. 9; Rupert, 2004, pp. 402–405). It is interesting to note that in later work Clark drops this last condition (2010a, p. 46).

of the mark of the cognitive themselves, they point to two necessary conditions that need to be satisfied by such a criterion (2001, p. 48).

(1) Cognitive states must involve intrinsic, non-derived content

(2) Cognitive processes possess a fine-grained causal structure that is similar to that of actual human cognitive processes (2001, p. 51).

The common thrust of these conditions is that cognitive processes and states are individuated by their causal structure and not by some functional equivalence to other cognitive processes. AA appeal to a distinctiveness argument here. They argue that cognition is not common in the world. Where it occurs, it must supervene on a distinctive lower-level process. They argue that "(r)oughly speaking, lower-level processes should be as distinctive as the higher-level processes they realise" (2008, p. 68)[9]. Cognition is an unusual phenomenon in the world, therefore the causal processes that bring it about must be too. That is why the causal details of this lower-level process matter[10]. AA engage with functionalism at this point. They do not reject the possibility that there are cognition-supporting substrates that involve non-neural elements - it is just that, in fact, there aren't any (2008, p. 69). This is not so much a direct repudiation of the Parity Principle as the claim that the HEC is contingently unlikely.

Let us deal with these conditions in turn. AA's first condition asserts that, roughly speaking, the content of a mental state, what it is about, cannot be derived from

[9] This distinctiveness argument is vulnerable to arguments from a sufficiently strong emergentism where higher-level processes emerge from the right kind of organisation of lower-level processes. It is exactly this argument that I develop in part II.

[10] Wheeler offers a counterexample to the distinctiveness argument when he points out that evolution sometimes produces the same function in radically different biological structures - what is called 'evolutionary convergence'. The implication here is that biological functions are generally multiply realisable, and AA would have to give evidence to show that cognition is the radical exception (2010, p. 250).

another intentional state, such as a public symbol system or a set of social conventions or norms that require interpretation. AA write: "(…) it is not by anyone's convention that a state in a human brain is part of a person's thought that the cat is on the mat" (2001, p. 48). In positing this condition AA (2001, p. 51) borrow aspects of John Searle's idea of *original intentionality* that the mental states are somehow reducible to causal arrangements provided by the physiology of the brain (see, for example, Searle, 1984, 2004).[11] For Searle, original intentionality is a property of the whole system, it is the whole organism that has thoughts about cups of coffee not parts of it. Unlike Searle, AA seem to allow that the non-derived content condition applies to parts of systems rather than the whole system (2001, p. 51). How else can they rule that part of a system, namely Otto's notebook, is not cognitive? The signs in Otto's notebook do not possess intrinsic intentionality because they rely on a social system of conventions and practices; they need to be interpreted before they have meaning for the organism.

By appealing to intrinsic content AA hope to establish a disanalogy between Otto and Inga. Otto's access to the content of the notebook requires perception of the relevant page of the notebook and then interpretation via a system of social conventions about language and addresses in NYC. In the case of Inga, no such interpretation is necessary – the contents of her memory are directly available. AA conclude that Parity Principle does not work in securing the HEC. The extra-cranial parts of the extended system are not cognitive because the content of their representations is fixed through what we might simply gloss as 'social practices'. They do not possess non-derived content.

[11] The question of whether intentionality is a mark of the mental is the subject of long-lasting and deep debate in the philosophy of mind which I shall sidestep for now (see, for example, Burge, 1986; Fodor, 1983, 1985, 2000; Putnam, 1973, 1988, 2004).

AA's second condition that any cognitive process should possess the fine-grained causal profile of human cognition appeals to the idea that scientific kinds are distinguished by the causal structure of their processes. "Science tries to get beneath observable phenomena to find the real causal processes underlying them; science tries to partition the phenomenal world into causally homogeneous states and processes. Thus, as sciences develop a greater understanding of reality, they develop better partitions of the 'phenomenological'" (2001, p. 51). The cognitive should be distinguished, like other natural domains, on the basis of underlying causal processes, and since it is a straightforward fact that familiar processes such as reasoning and perceiving are cognitive in the human case, other cognitive systems should have a similar causal structure.

There is certainly a worry that AA's second condition builds in anthropocentrism by setting the benchmark for cognition to be the fine-grained causal details of *human* cognition. While AA recognise that the burden of proof rests on them to show that their criteria are not just question-begging, there is doubt in the literature whether they have done this (Clark, 2010b; Menary, 2006, 2010a; Piredda, 2017).

There is a more general worry that AA have hitched their strategy to the bigger project of providing an adequate theory of non-derived content in general. That this makes AA's argument hostage to representationalist fortune has not gone unnoticed in the literature (see Hutto and Myin, 2013, Chapter 4).

> (…) (T)he concept of content, let alone underived content, is contested and incomplete. Furthermore, such definitions may be too narrow or restrictive, or they may rule out more exotic cases of cognition - for example artificial cognition and cognition in organisms without complex nervous systems (Menary, 2018, p. 192)[12].

[12] This comment is topical – I have already mentioned the interesting work being done in the field of *plant cognition* where the suggestion is that despite conventional wisdom plants *do* have

Earlier I wrote that AA interpreted Searle's holistic notion of original intentionality in a way that seemed to make intentionality and cognition local properties of parts of systems. This seems to get them into some deep water.

Here are AA supposedly making a joke at Clark's expense: "Question: Why did the pencil think that 2 + 2 = 4? Clark's answer: Because it was coupled to the mathematician". (Adams, F. and Aizawa, 2010b, p. 67). To which Clark's retort is another exchange: "Question: Why did the V4 neuron think that there was a spiral pattern in the stimulus? Answer: Because it was coupled to the monkey". (Clark, 2010b, p. 82).

Clark's reply invokes the mereological fallacy against AA in which a property of the whole system (i.e. being cognitive) is ascribed to a proper part (see Bennett and Hacker, 2003). I interpret CC to be saying that it is the whole system that is cognitive, and that the notebook belongs to this system because of its functional contribution. But AA seem to roll back their claim because they do not rule out the possibility that there can be parts of a cognitive system that do not involve non-derived representations. They concede "(…) not every component of an X system does X" (2009, p. 85), and "having argued that, in general, there must be non-derived content in cognitive processes, it must be admitted that it is unclear to what extent every cognitive state of each cognitive process must involve non-derived content" (2001, p. 50)[13]. Clark's response to this last concession seems to be the last word: "But this concession, I submit, removes the entire force of the appeal to intrinsic content as a reason for rejecting [HEM]. For it was no part of

cognitive capabilities (Calvo, 2016; see Frazier *et al.*, 2020; Hiernaux, 2021a; Linson and Calvo, 2020; Segundo-Ortin ad Calvo, 2019).

[13] This radical reductive argument applied to cognitive systems is what I call the *cognitive shrinkage* problem and is parallel and opposite to cognitive bloat. Clark makes a similar point in the section on hippo-world in *Supersizing the Mind* (2011a, pp. 109–110).

[HEM] to claim that one could build an entire cogniser out of Otto-style notebooks!" (2005, p. 6).

CC are concerned with "when is a physical object or process part of a larger cognitive system?" while AA seem to be asking the altogether more dubious question "when should we say, of some candidate part, that it is *itself* cognitive?" (Clark, 2010b, p. 84). It is not clear that this question is even intelligible. Shaun Gallagher writes: "in functionalist terms, no part, process or element is intrinsically cognitive. It is only cognitive in terms of the role that it plays in the system as a whole" (2018a, p. 426). As Mark Bickhard stresses, and I mentioned above, function only makes sense relative to the rest of the system and, as I shall add later, the system's larger goals (2000). It is a category error to speak of function independently of the rest of the system. A carburettor only has the function it does because of its linkages to other parts with other functions such as the fuel pump. Similarly, it might be that it is a category error to speak of cognition except as a property of the whole system. The central question is better framed as whether a component belongs to a cognitive system rather than whether it is itself cognitive.

1.3.3 The Otto two-step and the ultimate control argument

The final set of objections I shall examine here directly concern the boundary between the internal and external processes in CC.

CC emphasise that theirs is an 'active externalist' thesis – indeed the term is used to acknowledge its roots in the work of Hilary Putnam and Tyler Burge while, at the same time, drawing a contrast with their 'passive' content externalism (Burge, 1986, p. see; Putnam, 1973). Putnam and Burge were externalists because they thought that the world external to the brain/mind played a role in fixing the content of linguistic terms like 'water'. Semantic content involved actual states of the world and was not simply 'in the head', as illustrated in his twin earth thought experiment. "When I believe that water is wet and my twin believes that twin water is wet, the external features responsible for the difference in our beliefs are distal and historical, at the other end of a lengthy causal chain" (Clark and Chalmers, 1998, p. 9). The causal chain is lengthy enough for CC to claim that the relevant external features are passive because they play no role "in driving the cognitive process in the here-and-now". This is what is meant by the word 'active' in the functionalism claim of the HEC (see 1.2). The Tetris pieces are manipulated in real time and Otto consults his notebook shortly before setting out for MOMA. The causal contributions are happening as part of current action and are not distal.

CC remark that even if Putnam and Burge are right that representational content may have an external component, it is not clear that these external aspects make any difference causally in the here and now. This can be illustrated with a simple behavioural test: what happens to behaviour if the 'internal' part of a system is held constant, but the 'external' part is changed? The answer seems to be, in the case of Putnam and Burge, behaviour looks just the same. The Twin Earth twin behaves around water in much the same way as the Earth twin. The test leads one to conclude that, at least in terms of the production of action, internal structure is doing all the work. On the other hand, active externalism of the CC variety passes this test. Change the content of Otto's notebook by altering the string '53rd street' to '52nd street' and Otto's subsequent behaviour changes.

CC's insistence on the active nature of the externalism they espouse can be used against them, however. It raises a question about whether, in the Otto case, the external structure is actually doing the work or whether it is an internal belief such as: 'the notebook contains the address of MOMA'. This move is called the *Otto two-step*: that Otto believes that he will find the address of MOMA in the notebook, and this leads him to the action of consulting the notebook where he finds the correct address that leads to the action of going to the museum. Once the initial belief is acted upon and the notebook consulted, the action of going to the museum is generated by an internal interpretation of the inscription in the notebook – the salient belief is the one Otto has is about the notebook not anything contained in it and that the contents of the notebook are just a causal or informational input into the system.

CC's response to this move highlights two features of their approach: their emphasis on the role of explanation and an appeal to simplicity.

> If we must resort to explaining Otto's action this way, then we must also do so for the countless other actions in which his notebook is involved; in each of the explanations, there will be an extra term involving the notebook. We submit that to explain things this way is to take *one step too many*. It is pointlessly complex, in the same way that it would be pointlessly complex to explain Inga's actions in terms of beliefs about her memory. (1998, p. 13 emphasis original).

Again, arguing from the Parity Principle they point out that Inga does not first have a belief that the answer to the question of the location of MOMA is to be found in her biological memory and then she consults it. Memory, for Inga, is transparent and poised for action. Similarly, then, Otto is so completely habituated in using the notebook that it is second nature, and he automatically consults it without the intermediary of a belief about the notebook.

These objections fall into a familiar pattern; one where the locus of control and action is the brain and that the external structure supports the internal system but does not initiate action by itself. Under these circumstances CC's detractors argue that the external structures are not proper constituents of the system. I call this the *Ultimate Control* argument. It is an argument that concedes that an external component functions as a scaffold but that the control of the system happens internally.

Clark deals with criticisms of this type by pointing out that if the problem is that a cognitive component is identified by its capacity to control action then there are internal submodules that are not cognitive because they are not involved in control. "The worry is interesting because it again highlights the deceptive ease with which critics treat the inner realm itself as scientifically unified" (2010b, p. 55). "Suppose only my frontal lobes have the final say – does that shrink the real mind to just the frontal lobes? What if (as Dan Dennett sometimes suggests, most recently in 2003) no subsystem has the 'final say'? Have the mind and self just disappeared?" (Clark, 2010b, p. 56; Dennett, 2003). The control question is interesting, not least because it is trying a 'mark of the cognitive' manoeuvre by the back door by identifying a cognitive system by a control structure.

In my view, Clark's argument is convincing, and that insisting on control is too strong to be a necessary and sufficient condition for cognition.

However, the story should not stop here. There is something quite attractive about a using a control-type argument as a mark of the cognitive. What if the argument were weakened? What if 'control' were to be a sufficient but not necessary condition for cognition, and what if the notion of control were weakened so that it did not mean an actual causal producer of action but rather the coordination of action producing processes? This is exactly what the CSA is – a weakened control-type argument that serves as a sufficient but not necessary condition for membership of a cognitive system. By not being a necessary condition for cognition it would not succumb to Clark's argument above. The condition would not promote bloat since the functional profile of part of a control subsystem is narrower and more distinctive compared to the massively diverse kinds of function played by general causal connections in the system[14]. Controllers do not normally seep out gratuitously. It is distinctly possible that some of the external features of the examples we have been examining so far

[14] I want to highlight the difference between 'restraining bloat' and 'not promoting bloat' that is central to my approach. The PP promotes bloat (at least in the eyes of some writers) because it gives reasons to believe that many systems are cognitive purely on their functional similarities with systems that are conceivably cognitive (as we shall see in the next section). The principle that I shall be pursuing in this thesis does not (I shall argue) promote bloat in the same way. However, it does not restrain it since is it is not a necessary condition.

might exhibit control, or rather, perform coordination roles and therefore stake a claim to be part of the system. Moreover, it avoids the mereological fallacy because the kind of control/coordination roles I have in mind are emergent features of systems rather than being purely componential. This is the kind of argument that will be central in this thesis and will be taken up in chapter 3.

1.3.4 Grain parameter problems

Mark Sprevak notes that all varieties of functionalism come equipped with a grain parameter which expresses the degree of tolerance of a cognitive kind to functional variation (2009). Imagine a parity principle for corkscrews for bottles of wine. Let us suppose that an object is a corkscrew if it is functionally equivalent to the classical Archimedean screw design. By this I mean that it does the same kind of thing as the classical corkscrew, it removes a cork from a bottle, and that the way that it does it, broadly or narrowly, is the same. The grain parameter specifies what we are prepared to take as being the same way. Set the grain parameter for the comparison too coarse admits any device that can remove the cork: from a gold-plated diamond encrusted design from the Grosvenor Park Hotel to the pincer used in fancy restaurants and includes the wall that you bang the bottle on at a party because someone forgot the regular corkscrew (it works but requires patience!). Set the grain parameter too fine and only devices that are Archimedean screws are admitted. In the first case there is corkscrew ‘bloat’ - the functional kind is just too large. Taking a too fine-grained comparison would omit many devices from the category that are commonly regarded as corkscrews. The art is to find a Goldilocks point at which the functional grain parameter captures all and only the set of ‘true’ corkscrews.

The role of the grain parameter in the Parity Principle of CC faces a similarly tricky balancing act. The parameter needs to be coarse enough for the Parity Principle to be blind to functional particularities of human cognitive processes. But at the same time, it must be fine enough not to produce functional equivalence with processes that are intuitively non-cognitive - that is it must be fine enough to avoid cognitive bloat. The worry is that satisfying the latter condition will also rule out the Otto-plus-notebook system because as Rupert points out, it does not exhibit typical human memory traits such as the tendency to exhibit negative transfer effects, that is the interference of previous knowledge with new learning, or generation effects, that the effectiveness of memory is correlated with the effort of learning (2004). Set the grain parameter too fine and the processes in Inga and Otto's cases are not functionally equivalent: Inga's biological memory exhibits negative transfer and generation effects, while Otto's notebook does not. Hence, they would fail the parity principle's own test and CC's argument for the HEC would fail. I call this 'Rupert's fork': a fine-grained functional comparison is unlikely to be met by extended structures, a coarse-grained functional comparison is unlikely, according to Rupert, to support any interesting inductive generalisations and do serious explanatory work in psychology because it will consist in a motley of diverse processes that have little structure in common.

The only way that the Parity Principle can work is if the grain parameter is set at a relatively coarse level. But Sprevak points out that this is just too permissive (2009). He starts with what he calls the 'Martian Intuition': that it is possible for creatures with mental states to exist even if they differ physically and biologically to ourselves. He argues that the Martian intuition resists anthropocentrism in cognitive theorising and therefore should be preserved. There is no value of the grain parameter that makes the Martian intuition true but the HEC false, because the HEC can be couched in terms of the Martian intuition. For example, let us take the Martian facing the task of navigating to MOMA as storing bit-mapped images in its brain and then decoding them. Such a Martian would be functionally equivalent to Otto's notebook. So, AA and Rupert will need to ditch the Martian intuition if they want to show that the HEC is false showing that their theories are indeed anthropocentric.

All well and good one might think – first point to the externalists. Unfortunately, Sprevak shows convincingly that a functional grain parameter that preserves the Martian intuition also leads to cognitive bloat. It is conceivable that a Martian uses, say, a bit-mapped version of printed text for memory and this includes the entire Encyclopaedia Britannica. But the Parity Principle then seems to licence the inference that the Encyclopaedia Britannica is part of Otto's cognitive system – the result: radical bloat[15].

For these reasons, the Parity Principle is under threat. There is no value of the functional grain parameter for which all of the following are true: the Martian intuition is preserved, HEC is supported, and bloat is avoided.

1.3.5 Skewed Benchmarks for Parity

The final argument against the Parity Principle comes from HEC friend Mike Wheeler. There are three dimensions to the functional comparison on which the PP is based: what features are relevant in the comparison, what size of grain parameter, and what systems should be used as a benchmark. The PP seems to assume that the answer to the third question is 'the human cranium' hence the phrase in the PP "were it done in the head". We examined the second dimension in the previous section, but Wheeler asks about the first and the third.

> Here is the wrong way to [apply the PP]. First we fix the benchmarks for what it is to count as a proper part of a cognitive system by identifying all the details of the causal contribution made by (say) the brain. Then we look to see if any external elements meet those benchmarks. Why is this the wrong way to go? Because it opens the door to the following style of anti [HEC] argument: we identify some features of, say, internal memory that are not shared by external memory, and we conclude that since the Parity Principle is not satisfied, [HEC] is false (2012, p. 31).

[15] There is an interesting argument by Tim Fuller along similar lines concerning using the parity principle with respect to the system consisting of human beings and scientific instruments (2016).

Wheeler then goes on to show how Rupert does just exactly this (2004, p. 407) and concludes that such differences tell against any attempt to see the external process as being of the same explanatory kind as the internal. The fan of HEC must "simply refuse to accept that one should allow extant details of internal memory to set up the benchmark for what counts as memory in general" (2012, p. 31, see also 2010, p. 418). He claims that Rupert begs the question against the HEC by claiming that what counts as cognitive should be fixed by the fine-grained profile of the inner. What is at stake here are the benchmarks used in the comparison once the grain size is fixed. On the other side Mark Rowlands claims that the same charge can be levelled against friends of HEC. "If Rupert's arguments against the extended mind are question-begging because they presuppose a chauvinistic form of functionalism, it is difficult to see why arguments for the extended mind are not question-begging given their predication on a liberal form of functionalism" (Rowlands unpublished quoted in Wheeler (2010, p. 255)).

Wheeler's solution is that parity should not be conceived as a comparison with the inner *simpliciter* but rather "as parity with the inner *with respect to a scientifically informed, theory-loaded, locationally uncommitted account of the cognitive.* In effect, the Parity Principle, as I have interpreted it, is an appeal for equal treatment against an unbiased and theoretically motivated standard of what counts as cognitive." (2010, pp. 419–420). Wheeler is embracing something like the mark of the cognitive.

Sven Walter argues in a parallel manner that parity arguments do not get off the ground without a mark of the cognitive (2010). What standards guide our intuitions when applying the phrase; "(…) *were it done in the head*, we would have no hesitation in recognising as part of the cognitive process"? However, and this applies to Wheeler's argument too, once you have a mark of the cognitive, then the parity argument becomes superfluous. Surely, argues Walter, one can decide on a case-by-case basis by reference to the independent mark of the cognitive - no comparison is needed.

Clark’s response to these objections is that the parity principle “was meant to command rational assent as a means of freeing ourselves from mere bio-chauvinistic prejudices” (2005, p. 515). It was intended for identifying the material vehicles of cognitive processes by ignoring the old metabolic boundaries of skin and skull and attending to the “computational and functional organisation of the problem-solving whole” (2005, p. 515).

I take Clark’s point, but it seems to me that the PP does considerably more than what is modestly claimed for it. It appears, at least in CC, but also in some subsequent publications, that it plays a central role in establishing the HEC. In other words, it constitutes an argument rather than being merely an intuition pump. If this is right, then the objections discussed in this section towards PP actually expose difficulties with the argumentative strategy adopted by CC. It is difficult to see how the parity argument can resolve these issues.

1.4 Systems arguments

It is conceivable that the HEC can be defended without appealing to PP. But more worrying to the friend of HEC are objections that target the very notion of an extended cognitive system itself. In the coupling-constitution fallacy argument AA raise difficult questions about the very metaphysics of cognitive systems and ask under what conditions CC can distinguish them, without a clear mark of the cognitive, from mere causally coupled systems.

1.4.1 Coupling-constitution fallacy

The coupling-constitution fallacy is perhaps one of the hardest objections to overcome because it requires engaging with the metaphysics of systems. For this reason the whole of chapter 2 is devoted to a specific programme for trying to solve it. The fallacy consists in confusing an element's causal contributions to a system with it being a component of the system. The accepted wisdom is that causal relations take place over time, causes precede effects, while constitution is a synchronic relation; a part belongs to a whole instantaneously. AA attempt to show that many of the examples given in support of HEC in the literature make the mistake of slipping from causal relations to constitutive relations. One of their targets is Robert Wilson's description of the children's game *Rush Hour* which involves plastic rectangles representing cars and lorries packed horizontally and vertically on a 6X6 grid surrounded by a fence. One of the cars is red and has to be extricated through a hole in the fence by moving the other vehicles into vacant squares. You are not allowed to remove the vehicles from the playing surface.

> The way in which most of us go about solving even a relatively simple Rush Hour problem involves a sustained perceptual and cognitive interaction with a highly structured environment. The board of fixed dimensions, the rules for the movements of the cars and trucks, and the objective of the game all structure and constrain what we can do in playing the game. But in playing it we do not simply encode all of this and then solve the problem. (Go ahead be my guest!) For most of us, at any rate, that is not possible. Rather, we solve the problem by continually looking back to the board and trying to figure our sequences of moves that will get us closer to our goal, all the time exploiting the structure of the environment through continual interaction with it. We look, we think, we move. But the thinking, the cognitive part of solving the problem, is not squirrelled away inside us, wedged between the looking and the moving, but developed and made possible through these interactions with the board. (Wilson 2004: 194).

AA accuse Wilson of committing the coupling constitution fallacy in this example. "What we might expect that Wilson means (…) is that, in this case, cognitive processing is not squirrelled away in the brain but extends into the interactions with the board. If this is what he means, although he does not literally say it then he appears to be guilty of the coupling-constitution fallacy". (AA 2009: 82).

The problem here, it seems to me, is that the notion of system seems to be doing significant work in this argument. If a system is, minimally, a set of causally coupled components, then what does it mean for something to belong to a system over and above that it is causally coupled to other parts of it? Such a notion of system surely permits the move from a component being causally coupled to part of a system to the conclusion that it belongs to the system as a constitutive part. But the problem with this is that it produces system bloat on a massive scale. Without extra conditions on what it means to be a system or on the nature of coupling, a causal notion of system constitution will not be sufficient to bound a system. Systems have a habit of being causally linked to their environments. AA are right to insist that systemhood (and particularly cognitive systemhood) requires extra conditions.

Of course, AA do supplement their notion with an extra condition for the system being cognitive which is that it satisfies the further conditions of the mark of the cognitive[16]. But as we saw in section 1.3.2 there might be difficulties fashioning them into workable demarcation criteria.

These extra conditions can be epistemological – a system is what is required to explain a given phenomenon, or ontological – a system possesses certain metaphysical features independent of whether it is being observed. In their mark of the cognitive AA are going down an ontological route. They are individuating cognitive systems through their involvement of intrinsic content and the particularities of the causal relations (yet to be set out) that produce cognition. Rob Rupert also suggests an ontological condition. “We should expect cognitive psychology to deliver a functional criterion distinguishing the genuinely cognitive

[16] Ross and Ladyman (2010) suggest that AA’s concept of the coupling-constitution fallacy is based on a false application of a container metaphor. As I suggest in part II this metaphor underlies many positions in the debate and is problematic.

from the merely causal contributors to cognitively driven forms of behaviour" (2010a, p. 118).

The epistemological route is somewhat familiar to us from the natural sciences or engineering and seems perfectly plausible. In a car what components do we need to consider if we are interested in explaining its braking capacity? Here we would identify the system consisting of the brake pedal, the hydraulics, and the brake pad assembly. It is integrated into the whole system of the car but is picked out by its role in making sense of this particular capacity. Part of this explanation is causal, but the causal picture is married to a functional explanation of the subcomponents of the system - such as the role of the braking fluid that transmits force from the brake pedal to the brake pads (at least it did on the cars around when I learned to drive). The braking fluid plays the functional role of moving, for the want of a better term, 'braking information' from the driver to the brake pad. The functions in the explanation are linked to the tasks of these systems in the overall project of getting from A to B. Perhaps we do not include the bolts that hold the pad assembly to the brake disc because they do not play an active role in the explanation even though the brakes would not work without them.

But both AA and Rupert interpret an epistemic criterion as incipient operationalism. "Another way that [friends of HEC] attempt to provide a mark of the cognitive is by a tacit operationalism. They may well reject this characterisation of their project, but one can see it in their tacit assumption that whatever process or mechanism accomplishes a given task must be a cognitive process or mechanism" (Adams, F. and Aizawa, 2009, p. 88). "If Clark holds the view that anything contributing to the production of cognitive explananda is thereby cognitive, then I too reject his view as excessively operationalist" (Rupert, 2010a, p. 122). But maybe there is some nuance that is lost in AA's dismissal that deserves a second reading. I agree that just equating the causal system responsible for a token performance of a task with a cognitive system is being too quick. However, if the evidence shows that the causal system is responsible for a set of related tasks and that this pattern has some stability then surely this is the basis for scientific advancement? Rupert himself agrees: “A science gets going because there appear to be distinctive phenomena in need of explanation… [sometimes] there turns out to be enough unity to ground a distinct science” (Rupert, 2010a, p. 122). Taking everything that is causally responsible for the relevant behaviour as part of the system is too liberal but perhaps there is a core set of processes that play a special functional role, whose functional description can be added to the minimal conception of a system to do the job of demarcation. Rupert’s recommendation is something that I shall follow up in part II of this thesis.

1.4.2 Embedded Cognition

The notion of system is central to a set of powerful arguments against the HEC by Rupert (2004, 2009a, 2009b, 2010a, 2010b, 2011, 2013, 2014, 2016)[17] Their potency derives from the fact that they share many of the basic assumptions and insights of the HEC. It is difficult to see how arguments can pull them apart. For example, both Rupert and CC embrace the idea that, for many types of cognition to happen, the environment external to the agent must play a crucial cognitive role and that the system cannot function without it. Despite agreeing then in the role played by environmental structures they differ on whether these external structures should be rightly considered part of the cognitive system. For CC, parity considerations may be used to justify the inclusion of organism-external

[17] Rupert goes some way to providing an extra condition for systemhood (see the previous section). He appeals to an informational/mechanistic condition which amounts to something like a density of interaction argument - like clustered networks. Such arguments have precedents in the literature (see for example Grush, 2003; Haugeland, 1998, p. 215; Simon, 1996). But the problem with this type of argument for system demarcation is that modules within a system are often less tightly bound according to measures of information than they are to the environment outside the system (see Clark, 2011a; Kaplan, 2012, p. 553). A well-designed modular system will have low bandwidth connections between modules in an ideal situation. Edwin Hutchins writes “(…) the normally assumed boundaries of the individual are not the boundaries of the unit described by steep gradients in the density of interaction among media” (1995a, p. 157).

components in the cognitive system. Rupert, on the other hand, treats these environmental components as an essential but external scaffold for the internal cognitive system. This is the *Hypothesis of Embedded Cognition* HEMC - the second of the 4Es[18]. Cognitive systems are internal to the organism but reliant upon rich environmental resources.

Rupert's first argument is based on explanatory parsimony. If embedded cognition and extended cognition make substantially the same predictions and invoke many of the same mechanisms then parsimony leads to the acceptance of embedded cognition because it makes fewer ontological commitments – it is, according to Rupert, the 'conservative' option. Rupert claims that for every HEC-type explanation we can always provide an empirically equivalent but theoretically more conservative analysis in terms of Embedded Cognition.

> In what follows, then, I treat HEMC and HEC as offering distinct, competing explanations of various cognitive phenomena. Of great dialectical importance will be the question whether we can make do with HEMC, or whether HEC offers superior explanations of the phenomena of interest to cognitive scientists. If HEC does not, then all other things being equal, we should endorse HEMC over HEC, by dint of the methodological principle of conservatism (2004, p. 395).
>
> If the cases canvassed here are any indication, adopting [extended cognition] … at the very best, yields only an unmotivated reinterpretation of results that can, at little cost, be systematically accounted for within a more conservative framework. (Rupert, 2004, p. 390).

But is HEMC really the more conservative option and where exactly does the burden of proof lie? Sprevak points out that what Rupert means here by conservativeness is an appeal to conceptions of the mind involving internal folk-psychological categories such as beliefs and desires and it would be a mistake

[18] Actually, as Mark Sprevak remarks (2010, p. 357) the version of embedded cognition defended by Rupert (2004) does not deny that cognition extends beyond the skull it just does not affirm it. In this thesis I shall take embedded cognition to mean the stronger internalist thesis.

to treat this as the output of a mature theory (2010, p. 360). Why, he asks, should folk-psychological notions should hold privileged status in the debate? Conservatism is an explanatory virtue only if the older more conservative theories are tried and trusted scientific theories that have stood the test of time. Appeal to folk-psychology is also a theoretical commitment that needs to be justified[19].

Rupert's second argument is more compelling. He claims that cognitive systems must be persistent. In fact, in later work he builds persistence into the definition of cognition: "a state is cognitive if and only if it is the state of a mechanism that is part of a relatively persisting, integrated set of mechanisms that produces intelligent behaviour" (2014, p. 107). He observes that there are many practices in the cognitive sciences premised on the persistence of the systems they study. Developmental psychology requires the existence of a system that persists over time to map its developing cognitive characteristics. Edward Tolman's experiments on the hippocampus of rats in mazes (Tolman, 1948) presupposed a robust persisting system in which changes could be induced through experimental interventions. "In order to tell whether a given system cognitively developed in this or that direction we must be sure to have the same system in view" (Schlicht, 2018, p. 232). But CC allow so-called soft assembly: systems opportunistically put together on the fly to perform a certain task and then dissolved when done. For example, rotation of a particular Tetris piece is part of the whole system only while it is being placed. Once attached permanently in the

[19] Of course, there is much written about the explanatory status of common-sense psychological concepts, and I shall not enter the debate here. From the point of view of this thesis, the key question is how far one needs to commit to a particular theory of cognition, whether folk-psychology or another theory, in order to say something useful about demarcating cognitive systems. The aim is to commit to the bare minimum in order not to beg the question. I shall add that I am sympathetic to the view of Adam Toon that folk psychological categories are useful fictions supporting our everyday social actions (Toon, 2016, 2021).

game the piece plays no further part in the system. Rupert's persistence requirement would rule out this kind of system.

In order to deflect Rupert's second argument, the friend of extended cognition would have to take a different view of what a system is. The received view in cognitive science, perhaps informed by its origins in computer science, is that a system is a persisting infrastructure through which something ephemeral (like information) flows. Systems are usually taken as being self-contained items, screened from their environments and interacting with them only through well-defined input and output channels. This thesis argues that these assumptions arise out of an over-reliance on the container or machine metaphor for systems and rejects them to mount a challenge to Rupert's persistence criterion[20].

Next, Rupert argues that extended cognition positions must earn their keep by supplying explanations that are not available to internalists. This objection trades on the view stated above that HEC explanations can be translated into HEMC explanations without loss. The embedded theorist, on this account, would take an externalist story such as the use of pencil and paper in performing a long multiplication task and describe pencil and paper as 'cognitive scaffolding' rather than 'functional system components'. She would quite happily agree with the extended cognition theorist that, in some cases, the system cannot function without this scaffolding and yet insist that scaffolding is all it is[21]. But this argument can cut both ways. If the friend of the HEC can show that the extended view is actually more parsimonious or simpler in some way, then this would amount to

[20] See the 'process' view developed in chapter 3.

[21] This is the argument that prevents the psychologist Lev Vygotsky from grasping the nettle and becoming an early extended cognition theorist. He realised the importance of scaffolding but he thought that ultimately external scaffolding had to be internalised (see Vygotsky, 1978).

an argument for the HEC. Then, it is internalist positions that must earn their keep. Again this is a position that is adopted in this thesis.

Finally, Rupert makes some positive claims for embedded cognition over extended cognition. He argues that embedded cognition gives a privileged causal-explanatory role for the persisting cognitive architecture over less permanent structures (2011, p. 433). He argues that the causal details of Inga's persisting neural architecture play more of a role in explanation than the causal coupling of Otto with his notebook. This seems to be an argument about the style of explanation appropriate to cognitive systems, and again the container metaphor seems to be in the background, cognition being implemented on something like a machine through which data flows. He is consequently reluctant to grant that the architecture of the system can change radically in the short term. Of course, all players in the debate accept that brains change over time. The question is ultimately about the timescale over which such changes occur. Friends of the machine metaphor will insist that data changes on a much shorter timescale than the architecture of the system.

An example or two might help here. Black cab drivers in London must pass a test called 'the knowledge' in which they must be able to name and navigate the 25000 streets within six miles of Charing Cross station (Productions, 2014). There are physical changes to the hippocampus involved in this learning exercise (see Maguire *et al.*, 2000, 2006). Similar changes in the hippocampus are observed in rats learning a water maze where the activation of specific cells ('place cells') matches the position of the rat at particular locations in the maze (Wilson, M. A. and McNaughton, 1993). Are these examples of neuroplasticity to be thought of as changes to system or data? The short-term view is that the hippocampus is part of the infrastructure of the system and the various neural phenomena within it are the mechanisms responsible for cognitive processes. But on a longer scale learning produces new infrastructure and the cognitive process of learning precipitates changes in what might be regarded as the fixed system in the short term. In other words, certain forms of learning might be associated with long term structural change. The conclusion might be that there is no clear distinction between fixed system architecture and the data passing through it, but rather just interlocking processes acting on different timescales (see Gallagher, 2018a) . The lesson of these examples is that it is possible that the fixed architecture computer metaphor plays too strong an implicit role in guiding intuitions in this area[22].

1.4.3 Simplicity arguments

Rupert argues that Clark's Scrabble tiles can be equally thought of as props or as part of the system. Neither position can be inferred from the success of the explanation because, so the argument goes, they are both equally successful. To decide between them one either needs an independent condition such as a mark of the cognitive, or one needs to appeal to characteristics of the explanation such as simplicity or the understanding that the explanation provides. "Ultimately, the proof is in the pudding. The deepest support for [externalism] comes from the explanatory insights that the extended mind perspective yields" (Chalmers, 2011, p. xvi).

The simplicity argument rests on the claim that treating the 'external component' as part of the system simplifies the explanation, hence making it better. For example, CC claim that it is more natural to treat the manipulation of the Scrabble pieces on the tray by the player as an epistemic action as part of an extended cognitive process. "Of course we could always explain my action [manipulating scrabble pieces] in terms of internal processes and a long series of 'inputs' and 'actions', but this explanation would be needlessly complex" (1998, pp. 9–10).

[22] I would go as far as saying that the basis of conventional computer science and 'good old-fashioned artificial intelligence', GOFAI (see Haugeland, 1998), is a Cartesian assumption that the system is screened off from the world.

However, it is entirely conceivable that Rupert could question whether this is indeed simpler without there being an agreed metric for simplicity.

This is the stalemate alluded to in the introduction. But given that both sides employ a simplicity criterion the stalemate would be broken if one or other side could show that their explanation was in fact simpler. Sprevak acknowledges that his analysis of this stalemate would not go through if there were a ‘thick’ notion of system that played an explanatory role (private communication). In this case a notion of system that did some explanatory work would earn its keep and break the deadlock. For example, a system could be a set of processes that give rise to some higher-level emergent function that is relevant to the explanation. Invoking the system now makes the explanation simpler (or even possible) because without it the emergent function cannot be explained[23].

Let us spend a moment thinking about how to understand the term ‘needlessly complex’. What drives a wedge between two explanations of Scrabble? The first takes the rearrangement of the pieces to be part of the system, and the second to that the system is taking in perceptual inputs and producing behaviour. Perhaps CC are thinking in terms of the extra operations of transducing perception and action at a system boundary which are required in the second scenario but not the first. This view seems correct if one signs up to the representationalist camp (broadly speaking the view that cognition essentially involves the manipulation of representations of some kind) because transduction is a conversion between causal inputs or outputs and representations. In this case system fills the space between perceptual input and behavioural output.

[23] This argument for a thick system sees such a system as an *emergence base* for a crucial cognitive function – as we shall see coordination is such a function.

If we do not want to sign up to representationalism, there are other options available which are examined later in the thesis. In any case the argument depends on a (hidden) premise involving a thicker notion of system than just a causal nexus to account for the difference between CC and Rupert.

Let us put it to one side for now and return to CC's second argument. The second kind of simplicity argument concerns cognitive kinds. CC argue that treating Otto's extended state as a belief helps to unify both the use of the notebook and Inga's biological memory under one explanation. "By using the 'belief' notion in a wider way, it picks out something more akin to a natural kind. The notion becomes deeper and more unified, and it is more useful in explanation" (1998, p. 14). It "unifies the psychological explanation of Otto and Inga in a valuable way. It allows one to see a common psychological action at work, irrespective of whether the agent relies only on internal resources or off-loads work onto the environment" (Sprevak, 2010, p. 357). CC effectively characterise belief as a functional kind that can be realised in many different ways whose details (including location) do not matter. There is nothing strange going on here, this is a broad explanatory strategy used in psychology, good old-fashioned functionalism.

The problem with the natural-kinds argument, according to Sprevak, is that it does not select between extended cognition and embedded cognition. CC's argument relies on natural kinds being cognitive, yet the benefits of the natural-kinds argument (unity, depth and black-boxing) can be had without this assumption. Making the natural kind cognitive has no effect on explanatory benefit.

The situation according to Sprevak is the following. Extended cognition invokes a more complex ontology because it posits that many different kinds of thing can play roles in cognitive systems. Once the ontology is established the explanation is simple and avoids the need for, say, extra inputs and outputs in the Scrabble example. A more complex ontology buys simplicity in the explanation. By complexity here I mean that there are more ontological categories and relations between them. For embedded cognition it is the other way round. The ontology is simpler, but the resulting explanation is more complex, by which I mean that the explanation has more elements to it. It has to account for the extra inputs and outputs between the system and the ‘external’ resource. Sprevak argues that there is a trade-off and in the result is an explanatory tie[24].

In the chapters that follow I shall argue that this is not correct and that in some cases the extended explanation is the only one that is intelligible, in others it is one that is simplest.

1.5 Agent-centredness and distributed cognition

So far in the exposition we have taken for granted that extension in HEC means what CC take it to mean – literally that cognitive system of an agent extends out beyond the boundary of the agent’s cranium. Ed Hutchins calls this approach *agent-centred cognition.* In this case the agent is privileged in the system and is responsible for the assembly for the extended parts and is the central owner of the system. In contrast, the CSA developed in this thesis avoids agent-centredness by adopting a systems approach which does not assume at the outset that there is a privileged ‘centre’ to the system. I want to argue that the rejection of agent-centredness takes the sting out of some of AA’s and Rupert’s criticisms and reduces the risk of a question-begging anthropocentric or neurocentric approach. One of the inspirations for the CSA is the theory of distributed cognition developed amongst others by David Kirsh and Ed Hutchins.

[24] Mark Sprevak has acknowledged (private correspondence) that he was inspired in these arguments by Alex Oliver (1996, pp. 1–12). Oliver points out that there is more to this problem than an inverse linear relationship between ontological complexity and explanatory simplicity (what he calls ‘ideology’). These ideas involve heavy metaphysical machinery such as trope theory – but the main point is the same. There is some kind of offsetting between ontology and explanatory complexity even if it is non-linear and subtle.

In this section I want to review Hutchins' arguments against Clark's agent-centredness in order to clear the ground for a new approach.

Many, but not all, writings of Clark suggest such an agent-centred perspective. Take his principle of ecological assembly (PEA) for example: "According to the PEA, *the canny cogniser tends to recruit, on the spot, whatever mix of problem-solving resources will yield an acceptable result with a minimum of effort*" (2011a, p. 13 emphasis original). The 'canny cogniser' is the agent in this case. There is a similar hint later in the book: "Let us make the (surely uncontroversial) assumption that the biological brain is, currently at least, the essential core element in all episodes of individual human cognitive activity" (2011a, p. 118).

> But where we find dense inter-animation and that inter-animation looks to be serving recognizably cognitive (for example, broadly speaking epistemic or knowledge-oriented) ends, then (assuming, see below, that we can also assign 'ownership' of the relevant states or processes to a distinct agent) then there is—or so I argued—no good reason to carve the mental cake according to merely metabolic joints (2011b, p. 447).

The ownership issue is crucial for Clark; this central distinct agent is responsible for the assembly of the system on the fly and for being 'impartial' (Clark's word) about which processes go on internally and externally[25]. In other words, it is possible to interpret CC as being an *Extension Thesis* for agent-centred cognition; that is, the physical realisers of the mind sometimes extend beyond the brain and skull – but the brain remains the central locus of control (see section 1.3.3). This is what Richard Menary refers to as *artefact extension* (2018, p. 202) and what I call in this thesis 'materially extended systems' (see Chapter 6). It is precisely this ownership issue - linked with the question of which part of the

[25] Clark discusses the work of Wayne Gray and Wai-Tat Fu in terms of the cognitive economics of distributing cognitive processes across internal and external structures (see Gray and Fu, 2004).

system is responsible for assembling the extended system that I reject in this thesis.

Ed Hutchins regards Clark's position as falling short of a general theory of extended cognition (1995a, 1995b, 2008, 2011, 2014). In his article *The Cultural Ecosystem of Human Cognition* he writes: "[CC-style] extended mind assumes a *centre* in the cognitive system: the organism (or the organism's brain), which is the normal mind container with respect to which cognition can be said to extend" (2014, p. 36).

If the HEC is conceived as an Extension Thesis: cognition extends out from, and depends essentially upon, a central cognitive agent, then it is open to attack from a number of directions such as AA's causal-constitution objection or Rupert's parsimony argument. AA or Rupert can assert that the cognitive properties of the system piggy-back on the cognitive features of the central agent. For inscription to be part of a cognitive process then it needs to decoded and interpreted by the central agent, therefore it is not *truly* cognitive. All the inscription does is provide a causal input to internal cognitive processes. I am not saying that these kinds of argument are correct as they stand, but merely that an agent-centred view makes them more plausible and places the spotlight on the boundary between the central agent and the environment.

Hutchins suggests another perspective. Suppose we simply observe that there is intelligent behaviour in the world – that Tetris is being solved, a piece of music is being created, and MOMA is being navigated to. We can ask what parts of the world are playing essential roles in the production of these performances. Perhaps the system responsible for the performance consists of an individual in harness with a number of environmental structures such as a notebook, or we might find that many individuals are involved in an orchestrated way. This is what Hutchins calls looking at the world from a *distributed cognition* perspective "all instances of cognition *can be seen* as emerging from distributed processes" (2014, p. 36 emphasis original).

Hutchins (2011, p. 439) points out that, at times, Clark does acknowledge the possibility of a non-centred cognitive system:

> The flow of control is itself fragmented and distributed, allowing different inner resources to interact with, or call upon, different external resources without such activity being routed via the bottleneck of conscious deliberation or the intervention of am all-seeing, all-orchestrating inner executive (Clark, 2011a, pp. 136–137).

Both Hutchins, and Clark with his distributed hat on, emphasise that a system approach will need to grapple with the thorny problem of how the system assembles and coordinates itself, given that there is no central organising agent to perform this task. Clark quotes Dennett's solution to this coordination problem with approval: "the manipulanda have to manipulate themselves" (Dennett, 1998; quoted in Clark, 2011a, p. 133). There is no central homunculus able to go about manipulating the artefacts in the system, rather bits of the system manipulate other bits of the system. Clark suggests that the HEC plays a useful role in countering homuncular views of the mind:

> The HEC reminds us that there is no single, all-powerful, hidden agent inside the *brain* whose job is to do all the real thinking, and which is able to intelligently organise all those teams of internal and external supporting structure. Indeed, on the most radical model that we have scouted, it is (as it were) supporting structure "all the way down," with mind and reason the emergent products of a well-functioning swirl of (mostly) self-organising complexity. (2011a, p. 136 emphasis added).

This is the approach taken in part II of this thesis. If we replace the word 'brain' with the word 'system' we get the CSA being proposed here.

1.6 A very brief sketch of the way forward

There is a tendency to portray the debate over the HEC as between two well-defined sides: the defenders of HEC and those who reject it. But things are a bit more complicated than this. I have only had space to present the parity driven arguments of CC and the major lines of argument against them. I have not said anything about the second and third waves of extended cognition or about enactivist and ecological approaches. Some of these positions reject the statement of the HEC presented in 1.2. In fact, the CSA developed in this document, also defends a slightly different version of the HEC that is not agent-centred, as the arguments in the last section suggest.

There is plenty to learn from such a plethora of views and while each contribution to the debate casts light on some aspect, none of them are entirely free of problems. Of the problems discussed in this chapter some of them such as how to fix the grain size parameter are associated with the specifics of CC's parity-based argument. Others such as cognitive bloat and the coupling-constitution worry are more general and deeper questions about the nature of cognition and the theoretical resources used to investigate it. These questions can be distilled into five general areas of contention in which the principal protagonists disagree:

1. What is the appropriate style of explanation in cognitive science?
2. Should cognitive systems be conceived of as agent-centred or distributed?
3. What theory fixes the reference of cognitive terms - how should we understand cognition?
4. What role does functionalism play in the theory?
5. What is the role of representation in the theory?

These areas bear on more specific problems listed in the text:

- How to respond to the HEC specifically its system, functionalism, manipulability and location claims?
- How to deal with the problems identified in the PP argument: specifically cognitive bloat, mark of the cognitive, Otto two-step, grain parameter, and skewed benchmark problems?
- How to solve the coupling-constitution problem and the challenge of embedded cognition.

In this section I outline, very sketchily, the way forward with these issues.

The story I want to tell has three main parts to it. The first part in chapter 3 establishes some somewhat thicker notions of system. Cognition emerges from the coordination of a set of more or less autonomous processes. The system consists in precisely these processes responsible for the emergence of cognition. Moreover, the system possesses emergent properties that themselves constrain the processes that make up the system.

The second part of the story in chapter 4 takes the standpoint that there is a commitment to cognition primarily as the set of processes responsible for intelligent behaviour understood as goal-directedness. Systems possess distal goals (perhaps linked to self-maintenance) which translate into tasks facing the system. The coordination of system processes needs to be sensitive to these tasks in order to produce behaviour directed at performing them.

The third part of the story in chapter 5 tells how the appropriate theoretical framework for understanding such a system is a kind of wide, non-representational extended functionalism. Cognitive processes, I shall argue, are individuated by their functional contributions to the performance of a task. These contributions are not described in terms manipulations of symbols but rather in terms of their operations on the world. The coordination functions in the system in many cases, themselves emerge from the operation of system processes.

The argument structure in the thesis is analogous to that of CC's simplicity and natural kinds argument in section 1.4.3. I shall take a system to be the minimal set of processes required to support an adequate explanation of the phenomenon, in this case a given goal-directed behaviour see as a performance of a set of tasks. I show in chapter 6 that this avoids Rupert's 'operationalism' objection.

The system itself may be diffuse and its boundaries difficult to draw, but my hope is that the core of the system, responsible for coordinating and pulling together all the different elements that constitute it, is easier to demarcate, because of the special functional role it plays.

1.7 Concluding remarks

The analysis of CC yielded five major areas in which philosophers in the field disagree. While there are undoubtedly difficult technical problems to solve, some of the work in this project is directed at the higher-level questions to do with the type of explanation required for understanding cognitive systems. Cognitive systems will be taken to be constrained, at the bare minimum, by the explanatory resources required to explain the essential features of the production of intelligent (goal-directed) behaviour. Starting from a systems view of cognition does not load the dice regarding the benchmarks for a parity style comparison, and it builds on some features of existing approaches, while avoiding their pitfalls.

Rupert wrote somewhat ironically at the end of his 2004 riposte to CC:

> (a)t this juncture, we might just consider abandoning the attempt at uniquely cognitivist theoretical research, moving instead to the study of complex systems in general: individual human systems, ant colonies, whirlpools, and extended systems that include individual human organisms together with external elements, among other possibilities. This might be a viable route for science to take, but it is not consistent with HEC: within such an eliminativist framework, mind and cognition -

> extended or otherwise - no longer appear as causal explanatory kinds (2004, p. 428).

One might see the chapters that follow as taking Rupert at his word. My hope is that *pace* Rupert cognition emerges not only as an explanatory kind, but one whose inner structure is available for investigation. Amongst its examples might number, not whirlpools, but certainly ant colonies, Mexican waves, parents clearing classrooms, and other extended systems.

The systems approach in the following chapters takes its inspiration from many sources. I have mentioned the cyberneticists and complex system theorists in the preface. A more recent inspiration is David Kaplan and his attempt to produce a general criterion for system membership - for any system - using the meagre resources of causal intervention. It is a bold and heroic effort, and even if it is not clear whether it succeeds entirely, it suggests a general kind of approach to the demarcation problem. It is to this that I now turn.

Chapter 2
Mechanisms and mutual manipulability: throwing the cognitive baby out with the bathwater.

2.1 Introductory remarks - Kaplan’s programme for demarcating cognitive systems

Chapter 1 detailed the original arguments for the HEC and listed some problems with them. One of them was the accusation that the HEC conflated components of a cognitive system with causal contributors to the system thereby committing the coupling-constitution fallacy. As Davidson (1987, p. 453) pointed out, a phenomenon can be causally dependent on something, even essentially so (for example sunburn to UV radiation) yet still not be constituted by it. Yet, friends of the HEC want to say that various extra-cerebral objects and processes are not just causally essential for cognition but are partly constitutive of it. There are a host of examples of this in the literature: body action in vision (O’Regan and Noe, 2001), eye-movements in memory tasks (Kaplan, 2012, p. 563; Wilson, R. A., 2004, p. 194), notebooks in navigation tasks (Clark and Chalmers, 1998), rotating blocks on the screen in a Tetris game (Kirsh and Maglio, 1994), inscriptions on paper in a long multiplication task (Rumelhart and McClelland, 1986), the use of gestures in problem-solving (Chu and Kita, 2011; Clark, 2011a), and transactive memory processes between long-standing couples (Sutton *et al.*, 2010; Theiner, 2009, 2018). CC argued that what distinguished constitution from mere causal contribution was the functional role played by the extra-cerebral component. The Parity Principle provided a sufficient condition for constitution: if we would unhesitatingly accept a process with the same functional role as cognitive if it occurred intracranially, then the extra cerebral component was partially constitutive of cognition. Unfortunately, we saw in the last chapter that the Parity Principle faced a number of problems, not least that there was no principled way of setting the grain size for the comparison. Furthermore, a liberal notion of functional equivalence required to secure the HEC, overlooking the smaller functional differences between notebooks and neural memory say, turns out to be too permissive and includes too much of the world in cognition, leading to cognitive bloat.

I propose abandoning the Parity Principle – though, as we shall see in the chapters to follow, not its functionalist pedigree. Later we shall see that the distinction between causal coupling and constitution can be marked by function, not comparatively, but by supplying a functional criterion for drawing boundaries around cognitive systems. First, I want to examine whether a causal criterion could work.

There is an existing account in the literature that attempts to supply a mechanistic criterion for system membership. David Kaplan (2012) uses mechanistic methods to distinguish the relations of constitution and causation. The first is synchronic and holds between parts and wholes, while the second is diachronic and is generally taken to hold between distinct objects or events. What makes this approach interesting is that it attempts to make this distinction for *any* system whatsoever and sidesteps the awkward problem of characterising cognition. It elegantly avoids a problem that Kaplan noticed in the debate over HEC, namely that each side appealed to what he called ‘proprietary demarcation criteria’ in deciding what was cognitive (2012, p. 548). He claims, correctly, that the debate regarding HEC is sensitive to the theories that each side appeals to in fixing the reference of cognitive terms and these theories are different, as we observed in chapter 1. The hope is that by adopting a general systems approach, a general constitution criterion can be found that is largely independent of theories of cognition. This chapter is devoted to a discussion of this method.

Kaplan proposes a mutual manipulability criterion (MUMA) for constitution that derives from the new mechanisms literature, in particular Carl Craver (2007). MUMA aims to identify when a given mechanism is a component of a phenomenon (rather than being say a causal contributor). It does this by manipulating the supposed component and observing changes to the phenomenon, and then manipulating the phenomenon and observing the effect on the putative component. If differences are observed in both cases, then the given component is part of the mechanism responsible for the phenomenon.

For cognitive systems, Kaplan still needs to characterise cognitive phenomena on which his demarcation criteria can be set to work. Eschewing a ‘mark of the cognitive’ he prefers to list a set of representative behaviours that are considered to involve cognition in a relevant way. These are behaviours performed by human subjects such as solving a maths problem, learning to play an instrument, playing Tetris or copying a pattern of coloured blocks. For a given behaviour on the list Kaplan uses MUMA to demarcate the responsible mechanism. If the mechanism thus identified lies outside the boundaries of the brain, then some version of extended cognition is involved, and the extended cognition hypothesis is true. MUMA therefore purports to provide a principled criterion for identifying the mechanism responsible for the cognition-involving behaviour. Kaplan insists that any successful argument to demarcation criteria should be descriptively adequate in the sense that it accurately reflects actual practice in cognitive science and neuroscience and that he asserts that MUMA satisfies this requirement.

This chapter will focus on Kaplan’s strategy, not because it is ultimately successful, indeed I shall argue that it is not, but because it gives valuable insights into some of the problems that any account of cognitive systems will have to face. The Coordinated Systems Approach (CSA) developed in part II of this thesis is, in part, a response to Kaplan's strategy and incorporates some of his insights.

I shall argue that there are problems both with the general strategy adopted by Kaplan and the specifics of the method. MUMA might be able to function as a heuristic guiding research and discovery, but it cannot function as a device to draw a line between the cognitive and the non-cognitive. In doing this I am guided by some other criticisms of Kaplan such as that of Beate Krickel (2018a, 2019a) and Michael Kirchhoff (2014, 2015, 2017), although I offer some new criticisms and come to substantially different conclusions to these authors regarding the way to take the investigation forward.

The chapter is organised in the following manner. The first two sections sketch the main features of the new mechanisms framework and MUMA needed to state Kaplan's argument and involve some technical detail. This is unavoidable since the most worrying objections to MUMA concern the details of the method. This applies particularly to section 2.3 where MUMA is set out. Section 2.4 sets out my reading of Kaplan's position. Section 2.5 discusses problems tied specifically to the central use of MUMA in the argument. Section 2.6 discusses more general problems of Kaplan's argument which may turn out to be problems for any purely mechanistic-style approach and provide some guidance for the direction of the investigation. Ultimately it may be that a purely causal criterion for distinguishing constitution from causation encounters obstacles.

2.2 Mechanisms and mechanistic levels

Mechanisms, dualistically, involve things and their activities. Their purpose is to explain phenomena in the world[26]. Mechanisms are, in the words of Romero, “bundles of structure and activity” (2015, p. 3755). They consist of “entities and their activities organised such that these are productive of regular changes from start or set-up to finish or termination conditions” (Machamer *et al.*, 2000, p. 3)[27]. A phenomenon is produced by the entities and activities in the mechanism.

Although there is some disagreement in the literature, a phenomenon is an occurrence in the world, and is what is either constituted or caused by a mechanism. Examples of phenomena are the behaviour of DNA as it replicates, the process of neurotransmission (Machamer *et al.*, 2000, p. 3), the maintenance of human blood temperature at 37C, or growth of an organism (Illari and Williamson, 2012, p. 124). Stuart Glennan is clear that the phenomenon is behaviour (2017, p. 24). I shall take the phenomenon then to be behaviour in a general sense that includes these examples.

The mechanism is responsible for the production of a phenomenon, a behaviour, but is not to be confused with it. As Glennan notes immediately following the quotation above:

> (…) we also speak of the patterns of phenomena/behaviour for which a mechanism is responsible. (…) When we say that a mechanism is

[26] Although mechanisms come in different forms in the vast literature, my starting point will be the seminal work of Stuart Glennan (1996, 2010, 2017), Carl Craver (2001, 2006, 2007, 2015; Craver and Kaplan, 2014), and William Bechtel (2007a, 2008, 2009a, 2009b, 2011, 2019; Bechtel and Abrahamsen, 2005; Bechtel and Richardson, 2010).

[27] Bechtel and Abrahamsen (2005) are much closer to the approach taken in this thesis in that they take mechanisms to be implicitly functional: “A mechanism is a structure performing a function in virtue of its component parts, component operations, and their organisation”.

> responsible for a phenomenon (or pattern or behaviour) we are saying that there is something behind that phenomenon. This distinction between the phenomenon and the mechanism that is responsible for it, or the pattern and the process that produces it, is central to understanding the nature of mechanisms and their importance for the scientific enterprise. (2017, p. 14).

Almost all the authors in this field emphasise that mechanisms are phenomenon relative. Craver refers to this as ‘Glennan’s law’: a mechanism is always for a particular phenomenon. There are two senses in which a mechanism connects to the phenomenon: by causing it or by constituting it. For example, dehydration is part of the causal mechanism for thirst, but it does not constitute it. On the other hand, the hippocampus is part of the constitutive mechanism for the navigational capacity of a mouse. In both cases, the literature emphasises that the relation between the mechanism and the phenomenon possesses some kind of regularity: “[m]echanisms are regular in that they work always or for the most part in the same way under the same conditions” (Machamer *et al.*, 2000, p. 3). This chapter is mostly concerned with the second kind of mechanism given the interest in system constitution.

I should add parenthetically at this point that there are different interpretations of mechanistic explanation within the philosophy of science. Wesley Salmon, amongst others, distinguished between *ontic* and *epistemic* conceptions of explanation (1994). Roughly speaking, the ontic view seeks to reveal the causal structure of the world, while the epistemic view takes explanation to be directed towards providing information, understanding, prediction, unification and so on. According to the latter view explanations involve epistemic constructs and representations such as models and text. Beate Krickel identifies two targets towards which the mechanist can take either an ontic or an epistemic view: the relata in the explanation, that is the phenomenon to be explained, and the mechanism that does the explaining, and the relation itself (2018b, pp. 19–21). A new mechanist may be strongly ontic in that she interprets the phenomenon and the mechanism, as well as the relation between them, being out there in the world. There are two types of ontic relation: constitution and causation. An example of a strongly ontic mechanist is Carl Craver (2007).

William Bechtel takes a different view:

> “(…) mechanisms do not explain themselves. They are operative in the world whether or not there are any scientists engaged in offering explanations. Explanation is an activity of scientists who must contribute mental labour in advancing explanations. Even the advocates of the ontic perspective are unable to avoid invoking epistemic notions, although they try to minimise them” (2008, p. 18).

I think that Bechtel is broadly correct here in advancing an epistemic view of mechanistic explanation where mechanisms exist in the world but that the explanatorily important relations require the machinery of representations and so on. I shall return to these questions in section 2.6.

It is through the organisation of entities and activities that the phenomenon is produced. Typical examples of entities in the literature include *place cells* in the hippocampus of a mouse running a Morris water maze, a protein ion channel in the membrane of a neuron, the double helix of a DNA molecule or parents faced with a disturbingly messy room after a children’s party. Typical activities are hippocampal cells being *activated*, cell wall proteins *swivelling*, DNA strands *separating*, and neurons *polarising*. The arrangement or organisation of entities and their activities gives the mechanism a distinctive behaviour that is likely to be different from the activities of its components. This is the *systematic* quality of many mechanisms - complex behaviours are accounted for by the organisation of simpler parts. This organisation-dependence seems to be a feature of many of the systems that we consider in this thesis, and in many of them, higher-level phenomena are brought about through the organisation of a small number of kinds of parts. It is the organisation of the entities and activities in a mechanism does much of the work in producing complexity and variety of behaviour rather than the complexity of the individual parts. Bechtel and Richardson state that: “A system with complex capacities must have a complex structure” (2010, p. 63). This emphasis on complex organisation of simple entities and their activities to form complex behaviours is a central theme of this thesis and we shall come back to it. A clock is composed of relatively simple components but because of the sophistication of their arrangement is able to produce a relatively complex and unlikely behaviour.

Entities can be mechanisms in their own right, which is one of new mechanisms’ most powerful features and makes mechanistic ‘levels talk’ possible. A complex mechanism is an organisation of entities and activities and each of these entities itself is a mechanism that is an organisation of further entities and activities and so on. Levels of mechanistic organisation are local to a phenomenon – there are no general grounds for inter-mechanism comparison. It is the existence of mechanistic levels that allow a mechanism to constitute a phenomenon and at the same time to be separate from it. On this story constitution is not identity.

Before I discuss mechanistic levels in detail let us look at an example. Krickel gives the example of a mouse running a Morris water maze (2018b, p. 103). This is a circular water pool (to prevent scent playing a role in the experiment) containing a hidden platform on which food is available or which offers a way out of the water. The mechanism that gives the mouse the capacity to navigate the maze seems to be a spatial map that is generated in the hippocampus. This spatial map consists in neurons inducing long term potentiation (LTP) (see Craver, 2007). LTP in neurons requires a mechanism involving the activation of NMDA receptors. The entities and activities at each of these levels belong to a mechanism that constitutes entities (and their activities) at a higher level. Thus, activation of NMDA receptors (partially) constitutes the mechanism for long term potentiation in neurons. This in turn constitutes the production of spatial maps in the hippocampus and so on. The phenomenon is the navigation behaviour of the mouse. The highest level is the spatial map, followed by neurons undergoing LTP, followed by NMDA receptors activating.

Mechanistic levels go some of the way to understanding the division of scientific labour into specialisms that target phenomena at different scales. The importance of mechanistic levels could, potentially, be felt both within scientific disciplines and across them. Neurophysiology and biochemistry studies both the LTP in the neurons of the mouse and the activation of NMDA receptors. The latter is part of the explanation for the former. Experimental psychology is interested in the navigational abilities of the mouse. The psychological explanation of the navigational capacities of the mouse in the maze could be viewed as just a different level of explanation to the neurophysiological explanation of the activation of NMDA receptors, the two levels being unified by the relation of mechanistic constitution. The NMDA receptor mechanism is a sub-mechanism of the hippocampal mechanisms responsible for the navigation of the mouse. The activity of entities involved in the navigational capacities of the mouse, place cells and so on, are produced themselves by mechanisms involving the activity of NMDA receptors. Thus, the idea of mechanistic levels gives substance to the idea that neurophysiological and psychological investigations are just different ways of looking at the same phenomenon. Whether this can be generalised to support an argument for the unity of the sciences looms large in the literature and is highly controversial (see Fodor, 1974, 1997). This controversy is relevant to the research in this thesis because of what it has to say about the adequacy of functional versus mechanistic accounts and whether one needs to provide a causal explanation at the lowest mechanistic level in as much detail as possible. I return to this issue in chapter 5.

The constitutive account of mechanistic levels is promising from the point of view of this chapter since it promises a clear distinction between constitution and causation. The inter-level relation is constitutive while relations between entities at the same mechanistic level are causal (see Craver and Bechtel, 2007). The relation between the NMDA receptor and LTP is constitutive while that between items on the same mechanistic level as the NMDA receptor such as sodium and magnesium ions is causal. As we shall see in the next section there are some difficulties with this account. Indeed, Bechtel has partially rejected some of the conclusions of his earlier paper with Craver (Bechtel, 2017).

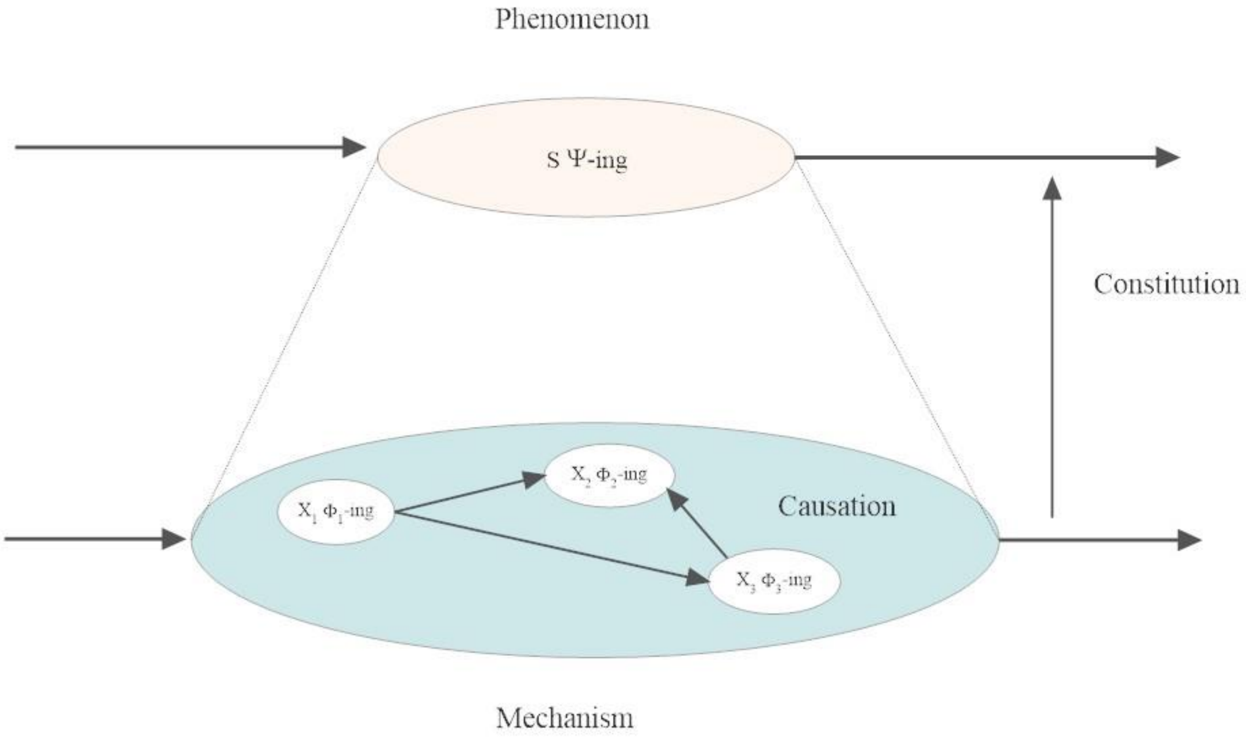

Fig. 2.1 The relation between a mechanism and a phenomenon. Since the phenomenon is the activity of an entity it can play a role in a larger mechanism hence this diagram also illustrates the relation of mechanistic levels. The large arrows represent causes in relation to external entities while the smaller arrows represent causal relations between constituent entities of the mechanism. Constitution is the vertical relation between mechanistic levels while causation is the relation between entities and activities at the same level. Diagram based on Craver (2007, p. 7).

2.3 The Mutual manipulability (MUMA) conditions for constitutive relevance

In many cases the experimental investigator is interested in the constitution relation between levels in a mechanism. A key question is often whether a certain structure X is a component of the mechanism S responsible for a certain behaviour. Craver in his (2007) discusses a causal criterion, *mutual manipulability* for X being a constituent of S or in the jargon: whether X is *constitutively relevant* to the behaviour of S.

Craver tells us that “one cannot delimit the boundaries of mechanisms – that is, determine what is in the mechanism and what is not – without an account of constitutive relevance” (2007, p. 141). Mutual manipulability (MUMA) is Craver’s test for constitutive relevance, that is, whether a putative component X really belongs to a mechanism S responsible for phenomenon Ψ or not[28].

> … (A) component is relevant to the behaviour of a mechanism as a whole when one can wiggle the behaviour of the whole by wiggling the behaviour of the component and one can wiggle the behaviour of the component by wiggling the behaviour as a whole. (Craver, 2007, p. 153 emphasis original).

Componenthood is stronger than parthood; something can be a part of a mechanism without being a component of it. The hubcaps are part of the car but not a component of the mechanism for its motion. Craver’s MUMA are intended to provide jointly sufficient conditions for a component to belong to a given mechanism[29].

[28] In keeping with the conventions of this field I use Latin uppercase for entities and Greek uppercase for activities.

[29] Technically neither Craver nor Kaplan claims that MUMA supplies necessary conditions as well as sufficient ones. Krickel (private communication) follows Baumgartner and Gebharter in

Kaplan’s plan is twofold (of which only the first step is stated explicitly). Step 1: use MUMA to demarcate the mechanism of a cognition-involving behaviour under various conditions. Step 2: Combine the mechanisms identified in step 1 in some way to produce the system responsible for the cognitive capacity that is represented by these behaviours (Kaplan, 2012). For example, a cognition-involving behaviour might be a mouse navigating a maze. Step 1: MUMA picks out the mechanism for this behaviour, for example, involving certain place cells in the hippocampus of the mouse. Set the mouse a different navigation task. MUMA picks out another mechanism for this behaviour, involving say different place cells in the hippocampus. Step 2: Combine these mechanisms to come up with a system responsible for mouse navigation. I shall discuss these steps in turn in the sections that follow.

To motivate MUMA let us consider the case where X is a component of mechanism S - X’s Φ-ing is a component of S’s Ψ-ing. Then an intervention can be applied to the activity of X, X’s Φ-ing, and a difference observed in the behaviour of the whole mechanism, S’s Ψ-ing. Conversely, an intervention can be made on the behaviour of the mechanism S, S’s Ψ-ing, and a resulting change observed in the activity of X, X’s Φ-ing. These two interventions establish a bi-directional relation between X’s Φ-ing and S’s Ψ-ing. Take an example of the blood circulation system. Suppose we want to know whether the heart and its activity is a constituent of this system. We can intervene on the heart (for example by electrically stimulating it) and observe the resultant changes in the behaviour of the whole system. Conversely, we can cause the circulation system to behave differently by, for example, tasking the subject to run up the stairs and then observe changes to the heart’s activity. This bidirectionality separates the case where X is a component from that where X is a background causal condition for

assuming that it does (2016). Necessary conditions are required to rule out putative components from being constituents of a mechanism.

the behaviour of S. If X is a component of S (*constitutively relevant*) one would expect a bi-directional relation, whereas if X is just a causal background condition (*casually relevant*) one would expect a uni-directional relation. For example, the force of gravity plays a causal role in running up the stairs, but it cannot be considered a constituent of the locomotion system since it is not affected by changes in the locomotion task. Moreover, these conditions spawn a practical experimental research programme. An experiment can be made to alter the behaviour of the putative component and the behaviour of the whole mechanism observed. Then a second experiment can be carried out changing the behaviour of the whole and observing the behaviour of the component.

The interventions required in the MUMA conditions are Woodwardian *ideal* interventions. I shall spend a little time unpacking this idea, despite it being a little technical, because the success of the MUMA account depends on the interventions being ideal (and one of the objections is that they are not).

Woodward considers the *relata* in causal relations to be variables, that is, items that can take a number of values. Craver, Kaplan and indeed Woodward himself emphasise that the causal relation does not however just hold between *abstracta* (see Craver, 2007, pp. 94–95; Kaplan, 2012; Woodward, 2003, p. 14). Variables are taken to be the sort of things that appear in ordinary causal vocabulary such as ‘electrical discharges cause thunder’. Consider the question whether C causes E. Roughly speaking, Woodward requires that we intervene on C in order that it takes on a specific value C_1 say, and then note that E takes on a value E_1. If this relation is regular in the sense that whenever C is intervened upon to be C_1 then E takes on the value E_1 then we can say that C causes E. It is important that the intervention I that fixes C does not also affect E, otherwise it is not C that is doing the causal work but I itself. Similarly, I should not affect any other variable that causes E or interfere with any causal intermediary between C and E. Such an intervention that changes E only through the change in C is called an *ideal* *intervention.* Adapting Woodward and Hitchcock (2003), Craver adopts the following definition:

I is an ideal intervention if the following conditions are satisfied:

1. I does not change E directly
2. I does not change the value of some causal intermediate Z between C and E except by changing the value of C
3. I is not correlated with some other variable W that is a cause of E.
4. I acts as a 'switch' that controls the value of C irrespective of C's other causes (Craver, 2007, p. 96).

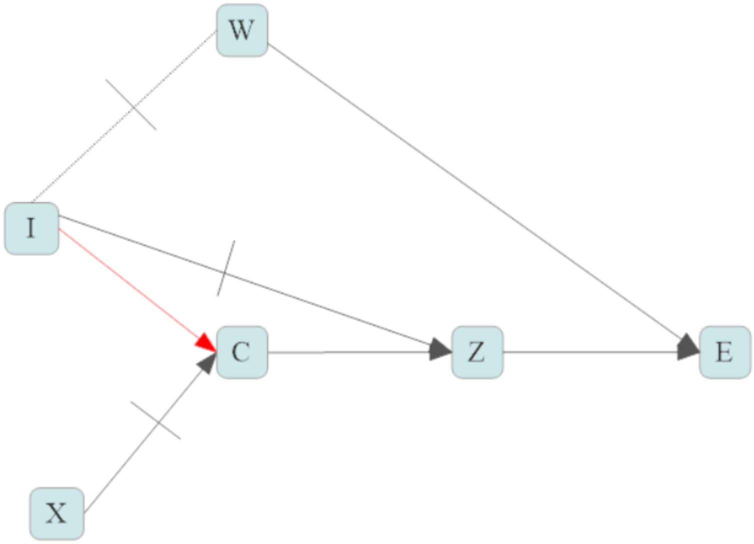

Fig. 2.2 A diagram of the causal structure of an ideal intervention I to establish a causal relation between C and E. Arrows indicate causal relations. A line through an arrow indicates that the causal connection is rendered inoperative by the intervention. A dotted line indicates a causal correlation. I changes C and switches off X's influence on C, is not correlated with W and does not act on Z. Investigators can observe the changes to E because of I.

Now we can state the MUMA conditions. Craver invites us to consider the following situation. X and its activity Φ-ing is a candidate component in the mechanism S and its behaviour Ψ-ing. What are the conditions upon X and its Φ-ing for it to be a genuine component? MUMA posits the following conditions:

X and its Φ-ing are part of S's Ψ-ing mereologically as a part to a whole[30].

M1 (bottom-up condition): there is an ideal intervention on X's Φ-ing with respect to S's Ψ-ing that changes S's Ψ-ing.

M2 (top-down condition): there is an ideal intervention on S's Ψ-ing with respect to X's Φ-ing that changes X's Φ-ing (Craver, 2007, p. 153; adapted by Krickel, 2018b).

The claim is that these conditions are sufficient for constitutive relevance. The heart is a mereological part of the blood circulation system and, as we saw above, it satisfies the conditions M1 and M2, therefore it is a constitutive part of the mechanism for the phenomenon of blood circulation. On the other hand, consider a situation in which an experimental subject is tasked with completing an incomplete English sentence grammatically. The heart is a mereological part of the subject engaged in the task and intervening on the heart (for example by stopping it) certainly affects the completion task, so it satisfies M1. But changing the completion task, I take it, has no effect on the heart's behaviour so it does not satisfy M2. In this case MUMA fails. Since MUMA are not necessary conditions this does not immediately imply that the heart is not a part of the mechanism responsible for sentence-completing behaviour but of course the suspicion must be that it is at best a causal background condition and not a constituent part of it. All we can say is that MUMA does not demonstrate constitutive relevance in this

[30] Carl Craver omits this in his original text but Krickel corrects him (2017, 2018b, p. 98).

case. Failure to satisfy M2 is not a defeater for the claim that X is a component of S.

2.4 Kaplan's argument

Kaplan claims that "(t)he mutual manipulability criterion (…) supplies an objective basis from which to distinguish background conditions from components in a cognitive mechanism" (2012, p. 562). As discussed in the introduction the advantage of this approach is that it is cognitively agnostic – it does not require one to sign up, in advance, to a particular characterisation of cognition[31]. The account is a generic criterion that applies to all mechanisms, it does not appeal to features specific to cognitive systems avoiding proprietary demarcation criteria.

> Because intervention-delimited boundaries are resilient to challenges arising from these assumptions, the [extended cognition] debate can thus be resolved without first settling more controversial debates about the nature of cognition. As indicated above, this is valuable because as long as proposed criteria depend upon prior adoption of some controversial assumption about cognition, the bump in the rug is simply shifted to that domain (Kaplan, 2012, p. 557).

Kaplan does not spell out his argument explicitly, but I shall take it to be something like the following.

1. HEC is the thesis that part of the system responsible for cognition can lie outside the physical boundaries of the brain.
2. Certain behaviours involve cognition.

[31] For problems with cognitive agnosticism see Walter and Kaestner (2012). I discuss this at the end of the chapter.

3. In order to find the extent of the cognitive system it is therefore necessary to find the extent of the mechanism responsible for these cognition-involving behaviours.
4. MUMA sets out sufficient conditions for finding the extent of the mechanism responsible for a given behaviour.
5. Apply MUMA to the external (environmental or non-brain involving) putative component in a cognition-involving behaviour.
6. If it satisfies the MUMA conditions then the external component is part of the mechanism responsible for the cognition-involving behaviour, therefore the system containing this mechanism possesses extra-cerebral components and HEC is true.

To check whether this reading is faithful to Kaplan's intention, let us try out the argument on Otto and his notebook described in chapter 1.

(1) 'Extended cognition is the thesis that part of the system responsible for cognition can lie outside the physical boundaries of the brain'.

(2) 'Certain behaviours involve cognition'. Following Clark and Chalmers let us take navigation to MOMA to be cognition-involving.

(3 - 5) 'Application of MUMA'. The notebook can be taken to be a mereological part of the Otto-notebook mechanism. For M1 let us apply an ideal intervention to the notebook – say the substitution of a different address for the museum, say 56th street instead of 53rd street. Given the terms of the story, Otto would turn up at 56th street instead so system behaviour would be changed. Now apply an ideal top-down intervention to the Otto-notebook mechanism for M2. Suppose that Otto keeps all the addresses of places he liked to go to in his notebook. We intervene at the mechanism level by changing the task – so instead of Otto wanting to go to MOMA we (omniscient and omnipotent authors of the thought experiment that we are) make him want to go to the Birdland jazz club. Then Otto

consulting his notebook would find, on a different page, that the address was W44th street and set off there instead. Otto's interaction with the notebook was different. He turned to a different page. In other words, intervening to produce a different system behaviour produces a different component behaviour. Therefore, the notebook and its activities are part of the mechanism for producing the required cognition-involving behaviour.

(6) The notebook component is external to the brain of Otto so HEC is true.

A second example is an experiment that has become a staple of the Extended Mind literature. Dana Ballard and his team conducted a series of experiments that involved subjects in a colour copying memory task (Ballard *et al.*, 1995). The subjects had to copy a model set of coloured blocks on a screen divided into a model and a workspace. They typically did this by physically manipulating blocks in the workspace to check their colour against the model, involving hand movements to control a mouse and eye movements relative to the screen. Both hand movements and eye-movements were recorded. The interesting question from an extended mind perspective is whether the saccadic eye movements are part of the cognitive machinery for performing the colour-matching task. One version of the experiment required subjects to stare at a fixed point on the screen. This is effectively an intervention on the putative component of the mechanism – a bottom-up intervention (M1). Subjects typically completed the task successfully, but it took three times as long. Moreover, when the nature of the task was changed the saccadic eye-movements changed as well. An intervention on the mechanism through changing the task caused a change in the activity of a putative component. Thus, saccadic eye-movements satisfy the conditions for mutual manipulability. By Kaplan's criterion then, they are a component of the mechanism responsible for the performance of the cognitive task of colour matching.

Finally, let us test MUMA on a classic problem in the extended cognition literature – the problem of separating causally necessary background conditions from causally coupled system components – a key issue to be solved if cognitive bloat is to be avoided. Suppose that a computer is engaged in a game of chess. Consider the problem of demarcating the system responsible for the chess-playing behaviour. How can we distinguish genuine parts of the system from causal background conditions? Suppose that there is a miner in South Wales (there used to be!) called Gareth Jones. Mr Jones mines coal that feeds the generator that generates electricity that powers the computer the runs the chess playing program. We can apply a bottom-up intervention and cause Mr Jones to be absent from his mining job. Let us assume, for the sake of the argument, that Jones was critical for the production of electricity and his absence meant that there was a power cut and therefore the chess program could not be run[32]. Under these circumstances then an intervention to a putative component of the system causes a change to system behaviour. So M1 is satisfied. Let us now try a top-down intervention. Returning to the normal situation where Mr Jones is working at the mine. Let us change the task for the computer and suppose that it is set to play draughts instead of chess. But we observe no change in Jones' behaviour. He is just mining coal as before. This means that M2 is not satisfied.

This is the indeterminate case: M1 is satisfied but M2 is not. Craver says that in these cases more information is needed: "[t]he complexities in the componency relationship make it difficult to say more about the intermediate cases in which only one half of the mutual manipulability account is satisfied. What to say in such cases, I suspect, depends on details peculiar to given experiments that admit of no general formulation" (2007, pp. 159–160).

[32] The fact that this scenario is unlikely because Jones' workmates could cover for him, illustrates that systems may be able to compensate for disruption is a problem for a MUMA account.

In this case MUMA does not do what Kaplan wants it to do. It does not distinguish constitution from causal relevance.

This is puzzling because Kaplan seems to think that it can cope with these cases:

> According to mutual manipulability, if it were functioning as a genuine component, then a top-down experimental intervention [on the system] should produce observable changes in [the component](as required by M2). (2012, p. 561 emphasis added).

This only makes sense if the MUMA conditions are necessary for componenthood. But this is not the case as Kaplan himself acknowledges. MUMA is sufficient for componenthood i.e. satisfaction of MUMA with respect to X implies that X is component, but X is component does not imply MUMA as is seen by considering the component of the blood purification system which is the left kidney. If a person has two healthy kidneys and the left kidney is removed there is no effect on the blood purification behaviour of the mechanism so violating M1. MUMA is not satisfied but X is still a component. Therefore, MUMA is not necessary contradicting the passage above. Nothing can be concluded from MUMA alone in these intermediate cases. The illicit use of MUMA as necessary conditions by Kaplan has not been commented upon in the literature as far as I can tell.

2.5 Problems with the MUMA account

The central plank of Kaplan's argument comprises the set of MUMA conditions (steps 4 and 5) so this will form the starting point of the critique.

It is well documented in the literature that there are problems with MUMA (Baumgartner *et al.*, 2018; Baumgartner and Casini, 2017; Baumgartner and Gebharter, 2016; Baumgartner and Wilutzky, 2017; Gallagher, 2018a; Kirchhoff, 2015, 2017; Kistler, 2009; Krickel, 2017, 2018b, 2018a, 2019a; Leuridan, 2012; Menary, 2018; Romero, 2015). I shall summarise the most pressing of these issues here. Although not all of these objections are fatal, their combined effect is seriously to undermine the account and point to the need for it to be modified.

2.5.1 Inconsistency of ideal interventions and top-down interventions

The first objection is serious because it casts doubt on the whole MUMA project by questioning whether ideal top-down interventions are possible or even intelligible. In Krickel's words:

> (…) the problem is that ideal interventions into phenomena (…) with respect to their constituents (…) are impossible. Interventions into phenomena are necessarily fat-handed, that is, they are necessarily common causes of the phenomenon and a constituent via two independent paths (2019b, p. 7).

Recall that Woodward's method for establishing that C causes E requires an ideal intervention on C. This is a causal intervention that fixes the value of C without affecting the value of E *or any other variable on the causal path from C to E.* The rationale for this condition is that we want to be sure that it is the variable C that does the causal work not a direct intervention on E itself or any other variable causally connected to E. The top-down intervention (M2) required by MUMA is that a causal intervention on S Ψ-ing will show up by changing a component part X that is Φ-ing. The problem with an intervention I on S's Ψ-ing is that there seems no guarantee that I will leave other parts of the system untouched including X's Φ-ing. After all X is a part of S by the mereological part-whole requirement, so doesn't an intervention on S also imply an intervention on X? But then I is not an ideal intervention and the antecendent condition for M2 is never satisfied. Ideal top-down interventions do not exist. And the irony is that they do not exist because of the *systemhood* of S!

An example will illustrate this problem. Consider the colour-matching task discussed in the previous objection. The component being investigated is the saccadic eye movement and the phenomenon is the colour-matching behaviour.

The top-down intervention is conceived as changing the task. Changing the task changes the colour-matching behaviour, and it also changes the saccadic eye movements. Let us grant for the time being that this can be couched in the terms of a 'causal' intervention setting a value for the phenomenon – the activity of the whole mechanism, I shall have a bit more to say about this in a moment. The question is whether this intervention is ideal – the intervention of changing the task does not change the saccadic eye-movements via another causal pathway, for example via other lower-level components in the system such as the motor-machinery. If the change in task also changes a component (say to do with the movement of the arm) that also changes the saccadic eye-movements (because it changes the focus of the agent's attention) then it does not seem to be a top-down intervention in the sense of a Woodwardian ideal intervention because it violates the rule about interfering with components on the causal pathway between C and E, namely the causal path I to Z is not blocked (see fig 2.2).

Consider a second example: intervening on the time-keeping capacities of a mechanical clock, say by forcing the hands, causes a change in almost all the mechanism. In all these cases it seems that there are changes across the board. An intervention at the system level produces a whole host of changes elsewhere. Baumgartner and Gebharter (2016) coined the phrase *fat-handed* to signify interventions that violate the ideal intervention criteria because they change more relevant variables than just the intended target of the intervention. It is as though a man with fat hands (it is always a man in these examples) is trying to play the piano and instead of picking out single notes he plays dissonant clusters of them. Interventions at a mechanism level produce a coordinated set of changes, they are fat-handed, ironically, because of the structure of the mechanistic system itself. Mechanisms are complex related ensembles and causal interventions on them are hardly likely to be clean surgical operations.

Maybe fat-handed interventions can still do the job required of them, Woodward (2017), Baumgartner and Gebharter (2016), and Krickel (2018a) suggest different modifications to MUMA to make it work, but then it loses its status as a one-stop shop for checking constitutive relevance to a phenomenon. Instead, there would need to be a recurrent programme where constitutive relevance must be inferred from a series of experiments by induction. A putative component X of a mechanism responsible for a phenomenon Ψ may respond to a single top-down fat-handed intervention but further testing is still required to infer that it is really part of the mechanism responsible for the phenomenon. If it continues to be activated by differing top-down interventions, then the experimenter may wish to infer to the best explanation that X is indeed part of the mechanism responsible for Ψ. The evidence base is much weaker in this case – being inductive instead of deductive. There is still the problem of induction or aggregating the results (see Krickel, 2018a).

2.5.2 Redundancy and self-repair

Carl Craver himself points to situations in which MUMA fails to provide a necessary set of conditions for a component being a constitutive part of a mechanism responsible for a phenomenon, because systems might adopt compensatory strategies in the face of inhibitory interventions (2007, p. 144).

Let us consider inhibitory interventions. Craver states that "A complete characterisation of the phenomenon requires one to know its *inhibiting conditions*" (2007, p. 126 emphasis original). This means inhibiting the activity of a potential component to check whether the phenomenon still occurs, as a bottom-up (M1) intervention. The non-occurrence of the phenomenon in these conditions is evidence that the putative component may well be part of the mechanism although this intervention on its own cannot distinguish between a constitutive component and a causal background condition. The purpose of the top-down (M2) intervention is to separate these contributions.

That is the theory. However, there are circumstances where the inhibition of one component is compensated for by another component in such a manner as to leave the phenomenon unchanged. As we saw earlier the removal of a kidney will not affect the blood purification behaviour of a human being with two healthy kidneys. But the kidney is still a component of the blood purification system. A lesion in one brain area is sometimes compensated for by other brain areas. This does not however provide evidence against that brain area being a component of the system for the relevant behaviour. Indeed, living systems are characterised both by possessing some degree of redundancy and by a capacity for self-repair.

It is not only living systems that can possess redundancy or can be adaptive in the face of an inhibitory bottom-up intervention. If we attempt to perform an experiment to check whether a particular sector of a hard disc is part of the mechanism responsible for the phenomenon of computer storage by disabling part of the disc, we are likely to be unsuccessful. The hard disc manager software simply marks the disabled blocks as inoperative and dynamically re-allocates the data to operative sectors of the disc (assuming that we haven't inadvertently messed up the state tables used by the disc management system).

Furthermore, there may be systems where the phenomenon in question is an emergent property of the interaction of a large number of simpler components. Disabling a component might not change the phenomenon because it does not significantly weaken the emergence base. Removing an ant from a foraging trail will not change the capacity of the nest for foraging. This is a significant problem because, as I shall argue in part II, interesting cognitive systems are likely to be of this kind.

Therefore, with the sort of systems that we are interested in, not even bottom-up interventions provide reliable data for componenthood. Ironically from an extended cognition point of view, the very systems that we are most interested in are likely to employ such dynamic adaptive or compensatory strategies.

2.5.3 The problem of sterile effects.

This objection claims that the MUMA conditions are, in fact, not even sufficient for constitutive relevance. It is revealing because it challenges us to think about what it means to be a mechanism for producing a cognitive phenomenon. Consider the mouse in the Morris water maze. Let us consider the component X of the mouse's brain which is the blood flow to the hippocampus. Let us denote the activation of this blood flow by Φ and the navigational behaviour of the mouse by Ψ. Clearly X is a part of the mouse so satisfies the part-whole condition. Moreover, if we intervene on Φ, for example by restricting blood flow to neural regions containing place cells then the mouse's navigational ability is (plausibly) altered so Ψ changes. Similarly, if we change the navigational task, we intervene on Ψ, then the pattern of place cell activation changes and this results in a change in blood flow to the hippocampus so changes Φ. By MUMA, this means that blood flow to the hippocampus is part of the mechanism for the navigational behaviour of the mouse. But almost everyone in the debate would describe blood flow to the hippocampus as being a necessary causal background condition and not a constituent component of the cognitive system. Blood flow is closely correlated with the cognitive mechanism – this is the principle on which fMRI works – but it is not a component of the mechanism. It is, in the jargon of the field, a *sterile effect*.[39]

[39] Kaplan acknowledges this case but alleges that "because regional increase in blood flow temporally lags behind neural activation by some small amount, it is safe to assume that preventing those changes cannot, strictly speaking, alter neural activation or the task performance it supports" (2012, p. 560 footnote). This is a strange line of argument to take since it would also count against an example that Kaplan uses to support MUMA namely the monkey haptic discrimination task (Kaplan, 2012, p. 557; see also Talbot *et al.*, 1968). Other writers such as Craver and Krickel *do* regard the problem that MUMA takes sterile effects as being genuine mechanism components as being unsolved. In any case Kaplan's time-lapse argument does not apply to the example of the 'epiphenomenal' cog in a piece of clockwork discussed below.

This example is interesting because it casts doubt on the ability of a set of purely causal conditions for constitution (such MUMA) to distinguish a genuine system component from one that is causally correlated with it. As we shall see, this doubt might be well-founded, and we might need something extra on top of purely causal conditions to do the job of demarcation.

For a second example of this problem, consider the colour copying task discussed in 2.4. We saw that according to MUMA saccadic eye movements are part of the task. The problem is that the same can be said of arm movements. When the task is changed then the arm movements change as well. And if the arms are restricted, then the player’s performance of the task is seriously compromised. This means that MUMA delivers the counterintuitive result that the arm movements are part of the cognitive process. Yet Kaplan himself suggests that arm movements “naturally seem like background conditions and not working parts of the mechanism underlying task performance” (2012, p. 564).

There are two options available to the advocate of MUMA at this point. The first option is to bite the bullet and accept the counterintuitive result that the causally correlated background condition is, despite appearances and accepted wisdom, a component of the mechanism. There are few willing to do this in the case of the correlation of blood flow and hippocampal activation. Moreover, Beate Krickel points out that biting the bullet, in the case of sterile effects, produces what she calls the problem of trivial extendedness. This is the problem that the “mechanisms that constitute cognitive behaviours will be extended purely due to the fact that we are dealing with cognitive *behaviours”* (2019, p. 12). That something is a behaviour implies that there are processes involved in producing the behaviour that are what I refer to later in the thesis as *plant* processes – that is processes that perform tasks that are not obviously cognitive processes[33]. Otto's motor processes controlling the movement of his legs arguably satisfy the MUMA conditions, but these processes are not plausibly involved in the cognitive task of navigating to the Museum of Modern Art. Krickel's objection of trivial extendedness can be understood in my terms as mixing up plant or implementational processes with cognitive ones – and MUMA cannot distinguish them. As soon as the phenomenon to be explained is designated as a cognition-involving behaviour, then the mechanism responsible for it (determined as it might be by MUMA) becomes a cognitive mechanism resulting in severe cognitive bloat.

Another problem with biting the bullet is that it is not consonant with our explanatory practices. Consider the mechanism of a mechanical clock. At least part of the mechanism consists in a set of interlocked gears consisting of toothed cogs. Add an 'epiphenomenal' cog to the mechanism which engages with one of the other cogs but otherwise does nothing else and has no other physical effect on the mechanism. I claim that the epiphenomenal cog satisfies the MUMA conditions. If you intervene on the extra cog by applying pressure, it will cause the time-keeping behaviour of the clock to fail. If you intervene on the time-keeping mechanism in some way the extra cog will be affected in the same way as the other functional cogs (its behaviour is locked to theirs). So, by MUMA the epiphenomenal cog is a component of the mechanism for the clock's keeping-time behaviour. Yet the functioning of the clock is completely explainable without any reference to the extra cog and, given the task of explaining how a clock works, there is no one who would include an extra cog in such an explanation.

[33] To be more precise here I will make the distinction in part II of the thesis between processes involved in coordination and those that are not. The former are associated with cognition and the latter, which I call plant processes, are not. It turns out, in many of the systems of interest, coordination is an emergent function and the two cannot be separated. But this is not always the case – see the examples in this section.

The mechanism for the behaviour of the clock simply does not include the extra cog. Biting the bullet puts the theorist at odds with our ordinary conception of mechanistic explanation. It fails both Kaplan's and Craver's own norms of descriptive adequacy.

The second option is to retain the causal conditions in MUMA but question whether they are satisfied. One way of doing this is to show that at a sufficiently fine-grained resolution the causal background condition is not precisely correlated, in the sense of satisfying M1 and M2, with the behaviour of the system. The main problem with this approach is that changing the grain-size of the correlation does not separate the active components from the correlated components as we can see with the epiphenomenal cog. This cog has precisely the same causal correlation to the phenomenon as the cogs in the active mechanism. There is no reason why active components are more correlated than background correlated components. In fact, the problem arises because being an active component is a *functional* not a causal condition. As I explain in chapter 5 causal correlation does not guarantee equivalence of function – functions are normative while causes are not. But MUMA is a purely causal criterion[34].

There are other objections which could be levelled at MUMA[35]. Some authors point out the contradictions involved in employing a causal condition to establish constitution – apparently defying the edict of keeping causation and constitution

[34] This is also discussed in section 6.6.

[35] Systems that incorporate multiple realisation with respect to task performance - that there is more than one way in which a particular task is performed - or possess other forms of redundancy, may fail correlation conditions such as M1 and M2. They may also fail in systems in which there are components which perform more than one function. This might be the case in connectionist networks. An example here is the work that Jeffrey Elman did in the 1990's concerning simple recurrent neural networks trained to predict the word or word category that immediately follows a word in a given sentence (1991). In such a system there is not a direct correlation between the states of individual nodes in the network and the behaviour of the system. It is the patterns of overall activation that are relevant to its functioning.

separate (Gallagher, 2018a; Leuridan, 2012; Menary, 2018). Some new mechanists (including Craver himself) want to distance themselves from the possibility of inter-level causation in mechanisms (Craver and Bechtel, 2007). Yet the top-down condition M2 in MUMA is precisely such a case of top-down causation. Another objection occurring in the literature is that MUMA does not adequately capture the diachronic process-driven nature of systems (Hutto *et al.*, 2014; Kaiser and Krickel, 2017; Kirchhoff, 2012, 2014, 2015, 2017; Krickel, 2017) or tends to view systems as possessing parts with fixed properties (Kirchhoff, 2012). Some of these questions will be dealt with in part II of this thesis. So, while the MUMA conditions look promising at first blush, the objections detailed here seem to threaten the project of demarcating the cognitive system by using MUMA to set the boundary of mechanisms responsible for cognitive phenomena. The next section will examine general issues with Kaplan's account.

2.6 Mechanistic explanation and general problems with Kaplan's account

I now turn to the more general problem of establishing the extent of the system from a MUMA or mechanistic style account. I shall take up the thread started in 2.2 regarding the nature of mechanistic explanation.

A Bechtelian epistemic interpretation sees the aim of mechanistic explanation to be advancing understanding of the phenomenon through representations and other epistemic constructs. In these terms then mechanistic explanation comprises three key elements: a description of a phenomenon to be explained, a description of a causal mechanism and an 'elucidation' of the relation between the two. "Explanation involves revealing the productive relation [of the mechanism to the phenomenon]" (Machamer *et al.*, 2000, p. 22).

The first step then is a description of the phenomenon to be explained. The description of the phenomenon may be something like a description of a token event or behaviour such as a mouse running a maze. But it could be described as the more general navigation capacities of the mouse. Different mechanisms might be invoked by an explanation depending upon how the phenomenon is described[36].

Given a description of a phenomenon, the second stage involves picking out a mechanism responsible for it – possibly using the method of MUMA. This might be the place cells in the hippocampus of the mouse. For Bechtel the mechanism may be represented by a diagram or schema, for Craver or Kaplan it may be an actual token mechanism in the world.

The third part of the explanation is cognitive*[37], meaning a text or verbal description that links the description of the phenomenon to a description of the mechanism supposedly responsible for the phenomenon[38]. This could be a text linking the description of performance of the mouse in the maze with the description of the activation of place cells. The philosopher of biology Denis Walsh points out that "a good explanation answers to both metaphysical and [cognitive*] demands" (2013, p. 45). The cognitive* demand is met "by describing the system in a way that makes the productive relation intelligible". And further: "Intelligibility arises not from an explanation's correctness, but rather from an elucidative relation between the explanans and the explanandum" (2013, p. 45). To explain a phenomenon, it is not sufficient to exhibit a mechanism, but the mechanism needs to be linked by a text to the phenomenon as described. This is important because it introduces a new element into the story – intelligibility. It

[36] This is sometimes referred to as the 'constitution of the phenomenon' in the literature.

[37] Cognitive from the point of the view of the explanation rather than the target of the investigation. I shall write 'cognitive*' when I am talking about features of the explanation and 'cognitive' when talking about the target of the investigation.

[38] Craver writes that in the explanatory text "the *explanandum* is a description of the phenomenon and the *explanans* is a description or schema of a mechanism" (2007, p. 139 footnote).

is no good exhibiting a mechanism as an explanation if that mechanism is just too complicated to elucidate the phenomenon. It is precisely the intelligibility requirement that might separate CC and Rupert's approaches in terms of simplicity (see section 1.4.3). Seeing a system through Rupertian eyes as doing many transactions with an external structure might be less intelligible than conceiving the whole as a single system. Having a preliminary understanding of mechanistic explanation, we can move on to the problems.

2.6.1 The problem of goal-directedness

The first problem is a general worry that MUMA and mechanistic explanation in general is inadequate to capture the richness and complexity of cognition. This is best illustrated by examining the top-down interventions invoked as part of a MUMA procedure. These interventions are characterised in Woodwardian fashion as being causal, but I claim that this description does not really do justice to what is going on in practice. What is going on in the case of top-down interventions on putative cognitive mechanisms is that *the system task is changed*. By top-down interventions in MUMA almost all authors mean changing the task presented to the system and then observing the effects on a lower-level component of a causal intervention on the whole system. Let us review the examples in the literature: changing the sensory discrimination task (Romo *et al.*, 1998), changing the navigation task for the mouse (Krickel, 2018b; Tolman, 1948), changing the block colour-matching task (Ballard *et al.*, 1995; Krickel, 2019), changing the Tetris game rules - effectively changing the task for the player, changing the Otto's navigation task by changing the destination, initiating a tennis task/reaching task/running task (Craver and Bechtel, 2007; Kaplan, 2012), changing the long multiplication task (Rumelhart and McClelland, 1986). In each of the cases what is changed in a supposed top-down causal intervention is the task faced by the system.

Despite the importance of the normativity involved in the performance of tasks it goes unnoticed in the new mechanisms literature. Mechanistic explanation describes how a mechanism produces a phenomenon. It does not say anything about whether a given behaviour is successful in performing a task or not. Yet task performance seems to be essence of the production of intelligent behaviour. Indeed, it is the essence of the systems to which the new mechanists apply their methods. Changing the task presented to the system seems to be rather more than simply causing the system to do something. While Woodwardian interventionism takes a cause to be a variable, so it can be, it seems, anything at all (2003, p. 21), something important is missed out by treating task as cause.

In their defence, mechanists may reply that goal-directedness is implicit in their account, they speak of entities and activities in productive relations, or engaged in functional activity. If this remains implicit then it is difficult to see what work these implicit teleological aspects do in the explanation. It seems altogether wiser to make them a central feature of the account. These ideas are developed in part II.

2.6.2 From mechanisms to capacities: the type-token problem

There are two further problems in Kaplan's account that threaten any account that relies on a classical (Machamer *et al.*, 2000) perspective of mechanistic explanation. They both concern a gap between a mechanistic explanation of a phenomenon taken to be a cognition-involving behaviour, and the kinds of explanandum encountered in cognitive science – typically cognitive capacities. The former is a token such as a given instance of mouse M running maze Z. The latter is typically taken to be a type – a cognitive capacity such as memory, problem-solving or navigation by mice in general.

The first problem is that there is a type-token mismatch between the two. Another way of putting this problem is to ask how does one arrive at the cognitive system from a method that just delivers mechanisms for token cognitive behaviours? On Craver's view a mechanistic explanation exhibits a token mechanism which is a specific causal arrangement in the world which explains a token phenomenon (a cognition-involving behaviour). Explaining a capacity, understood as a set of behavioural dispositions, will need more work (see Krickel, 2018b, pp. 113–114). How do we proceed? The second step of Kaplan's argument (section 2.4) requires that a set of 'cognition-involving' behaviours are selected as representative of the capacity.

Once we have a set of representative behaviours, and let us pass over for now the problem of how they are selected, the second issue is how are the token mechanisms responsible for these behaviours to be aggregated to yield a mechanism for a capacity? How do we put these mechanisms together to get a general mechanism? This amounts to picking out capacity-relevant features of the mechanism *qua* the capacity rather than incidental factors involved in the behaviours. Suppose that we are interested in mouse navigational ability, and we observe a mouse running a maze to get the cheese on a number of occasions. How do we separate the navigation mechanisms from those responsible for running or doing other things like smelling the scent of cheese?[39]

Kaplan is unclear on these questions and vacillates in his paper between taking the phenomenon to be a behaviour and a capacity. He starts by describing a mechanism component responsible for a behaviour token. "(…) [I]ntervening on the putative component should produce a corresponding effect in the target behaviour or performance" (2012, p. 563). There is some ambiguity here. The word 'behaviour' seems to refer to a token event - a particular behaviour - which is indeed what MUMA delivers. However, 'performance' seems to imply

[39] The astute reader will recognise that this is again the problem of separating 'plant' from 'cognitive' mechanisms.

‘performance of a task’ which brings in a non-behavioural normative notion of task. Later he only refers to performance of a cognitive task: “Theorists wanting to make use of [the Ballard colour copying task] in support of [extended cognition] might invoke mutual manipulability to help clarify the claim that saccadic eye movements (and the oculomotor system) are genuine components underlying *performance of this cognitive task”* (2012, p. 564 emphasis added). Kaplan shifts to ‘capability’ in his discussion of Clark’s example of the extended swimming mechanism of the bluefin tuna. Clark wants to show that the tuna create vortices and whirlpools that are exploited to aid swimming: “the tuna find and exploit naturally occurring currents so as to gain speed, and use tail flaps to create additional vortices and pressure gradients, which are then used for rapid acceleration and turning” (1999, p. 345). Kaplan describes this as an “example of an environmentally-extended mechanism (…) involving the role of the local ocean environment in the swimming *capabilities* of certain fish species” (2012, p. 565 emphasis added). The ambiguity concerning the kind of phenomenon to be explained masks the problem of how to construct a mechanism for a capability from the mechanisms for token behaviours. There is I suppose nothing to prevent the mechanisms for token behaviours exhibiting a capacity being non-overlapping. Why should we assume that the mechanisms for the mouse running maze no.7 overlap with a mouse running behind the skirting board of no. 4 Privet Drive?[40]

Let us consider Kaplan’s options for answering the question. The two obvious choices for aggregating mechanisms to form a system are the mereological sum and the mereological product. The mereological sum simply collects all the mechanisms together and calls them the system for the capacity. This seems rather too inclusive and runs the risk of serious cognitive bloat. Moreover, it

[40] Perhaps there is an implicit modularity assumption that does some work in the argument that the mechanisms for token behaviours overlap.

comes up against the inability of MUMA to distinguish ‘plant’ and ‘cognitive’ components – so the resulting collection runs together every process satisfying the MUMA criteria in every cognition-involving behaviour.

On the other hand, the mereological product means only taking the common mechanisms (or parts of mechanisms) for all the representative behaviours in the collection as being the system for the capacity[41]. This seems rather too strong and is vulnerable to the same problems of redundancy and multiple realisation of tasks as the MUMA approach. The bigger the representative set of behaviours, the bigger the risk that the system identified by the taking only components active in every behaviour is empty - that is, there is no system identified. If Kaplan takes a capacity as the mechanistic phenomenon, then it is very difficult to see how MUMA can yield results in practice and more generally how mechanisms can furnish an explanation at all.

2.6.3 Cognition-involving behaviours

The final problem is a consequence of Kaplan’s strategy of thinking of cognition in terms of certain behaviours that are supposed to be cognition-involving. Remember he does this to sidestep making a commitment on the nature of cognition. However, the worry is that he will need to confront this issue anyway. The difficulty to be overcome is how to pick these representative behaviours without a circular reference to cognition. For the behaviour to yield a mechanism for a cognitive capacity these need to be cognitive behaviours that demonstrate the capacity but, *ex hypothesi* they need to be chosen without a criterion for cognition. The solution adopted by many writers reaching this point in the argument, and adopted by Kaplan, is to take ‘uncontroversial’ examples of

[41] If I have got her right, this is the route taken by Beate Krickel by espousing a criterion of behaviour non-specificity (2019b, p. 17).

cognition-involving behaviour. The problem here is that this strategy might involve implicitly appealing to some pre-theoretical notion of cognition, or even some anthropocentric prejudice, to fix the reference of cognitive concepts, so ushering in a theory of cognition by the back door.

The CSA makes a different move at this point, by taking goal-directed behaviour as the relevant behaviour to be explained by a cognitive system, as defined in part II of this thesis. This sets the bar quite low and might invite accusations of over-inclusiveness. But an advantage of this approach is a cognitive gradualism that I defend in part III.

2.7 Some remarks on cognitive agnosticism

Kaplan's MUMA proposal is interesting, not least because it asserts demarcation criteria for cognitive systems in the absence of a substantial notion of cognition. Before I bring this chapter to a close, I want to mention that there are arguments against such cognitive agnosticism in an approach to the demarcation problem. This is relevant because the approach taken in this thesis builds on Kaplan's general strategy.

Some authors cast doubt on the whole agnosticist strategy. Rupert writes “the author who asserts that cognition extends into the environment had better be prepared to tell the rest of us what it is that extends into the environment” (2010a, p. 114). Larry Shapiro concurs: “(…) the controversy over [extended cognition], if it is to avoid dwindling into linguistic insignificance, must confront questions over the meaning of mental, or cognitive, processes. Whether the body and world may be constituents of cognition or merely causally related to cognition depends, first, on what we mean by cognition and, second, on whether the body and world fall within or outside the circle we draw around the constituents of cognition” (2011, p. 161). Sven Walter and Lena Kaestner point out that examples will not do on their own, “examples must be *interpreted*, and any interpretation presupposes a theoretical background against which it is made” (2012, p. 16 emphasis original).

I think these points are broadly correct. There needs to be something more than just mechanistic constitution added to the mix to support the project of demarcation of cognitive systems. Although Kaplan’s programme is ingenious in its distinctive bid to ground a demarcation of cognitive systems upon nothing more than a criterion of mechanistic constitution, ultimately it fails. It does so because the machinery it appeals to is too meagre to do the job required. However, I disagree with Rupert that a fully-fledged theory of cognition is needed. Demarcation can be made on the basis of a shallow characterisation of the basic features of a cognitive system, in a relatively uncontroversial manner. That is the plan of this thesis.

2.8 Concluding remarks

This chapter examined Kaplan’s bold strategy of basing a cognitive system demarcation criterion on mechanistic constitution. We have seen that there are some serious issues with the strategy, and, if it is to be successful, work needs to be done to strengthen it or modify it.

Some of the problems stem from the MUMA technique itself. Serious questions arise concerning whether MUMA can, in fact, distinguish between the components of a mechanism and its causal background, and whether it can underwrite constitutive relevance or merely causal relevance. Principal amongst these are problems to do with the applicability and even intelligibility of top-down ideal interventions. But even if these could be solved, MUMA cannot deal with systems which can self-repair or which contain redundancy, it cannot separate sterile effects causally correlated with the phenomenon from active components.

These are documented problems. Undocumented is a second cluster of problems concerning the more general strategy of using behaviours as proxies for a cognitive capacity to identify a mechanism for the capacity. Even if MUMA, or some other criterion, can identify the mechanisms responsible for these behaviours there seems to be an explanatory gap between these mechanisms and the system responsible for the cognitive capacity. How these mechanisms are to be assembled into a cognitive system is far from clear.

MUMA has nothing to say about the normativity involved in cognitive behaviours. Interestingly, all the examples of top-down interventions in the literature are the setting or changing of tasks for an experimental subject. Tasks are normative in that they have success conditions. This will need to be addressed.

In the end then, the notion of mechanism does not seem to be substantial enough to get a grip on the demarcation problem for cognitive systems without invoking some kind of ‘mark of the cognitive’. Indeed, there is doubt in some quarters that such cognitive agnosticism can ground a criterion for demarcating cognitive systems. By remaining agnostic about the nature of cognition, Kaplan has thrown the cognitive baby out with the bathwater. Without a thick enough notion of system to which mechanism can be tied he is unable to do justice to the complexity and sophistication that is cognition.

Nonetheless, Kaplan’s approach is novel and creative and offers some useful pointers for the current investigation. His emphasis on the characteristics of the system as encompassing resources of different kinds rather than thinking in terms of extending outward from a prior cognitive agent is a fresh change of perspective. MUMA leads us to think more generally about cognitive systems in terms of their responsiveness to changes in environmental features and how features of the system are responsible for kinds of behaviour. Rather than treating these changes as causal interventions we should take them for what they are in the real and imagined experiments that play a role in these debates: changes in tasks.

Finally, Kaplan helps us understand the role that the concept of system plays in the argument. The argument at the end of chapter 1 showed how a system was more than just a collection of its parts. Components are able to interact with the system itself in such a way as to determine their properties in a non-trivial manner. Perhaps this is not best expressed in terms of mechanisms, although mechanistic approaches are right to highlight what Michael Kirchhoff calls 'organisation-dependence' (2014, p. 269). Given that systems are strongly or loosely bound coalitions of components the question is how they are coordinated to be productive of intelligent behaviour. Trying to answer this question will provide what we have been missing so far – a simple, but hopefully effective, mark of the cognitive.

Introduction to part II

In 1952 the cyberneticist Ross Ashby wrote “The fact that the stability of a system is a property of the system as a whole is related to the fact that the presence of stability always implies some *coordination* of the actions between the parts” (Ashby, 1960, p. 57 emphasis added). Stability for Ashby names the capacity of the system to respond appropriately to its environment. If he is right, then coordination is the key to the production of intelligent action. If we understand intelligent action as being goal-directed, then the suggestion seems to be that coordination is the crucial feature of cognitive systems that can further the project of demarcating such systems. Part II of this thesis explores how far this approach takes us by developing a theory of coordination.

It starts by characterising goal-directed behaviour. It then asks what constraints there would be on a system comprising a coalition of process to be able to produce such behaviour. These constraints will be necessary and sufficient conditions for coordination. The argument will be that coordinative processes are responsible for the goal-directedness of behaviour produced by the system. As such then they lie at the core of the system. If such processes were identified in the wild as ‘external’ to the organism, then this would be evidence to support the HEC.

An important step towards such a theory of coordination is finding the right level of description at which coordination manifests itself. The previous chapter suggests that perhaps the causal mechanical level of description is not appropriate for this task. I shall argue that a higher more abstract mode of description is appropriate – the level of function. Furthermore, coordination is a process rather than a structure and therefore sits more comfortably within a process ontology rather than a mechanistic ‘thing’ ontology.

A positive takeaway from the discussion of the mechanistic MUMA strategy in Chapter 2 is the possibility of cognitive agnosticism: in its soft form, being able to propose demarcation criteria in the absence of a fully-fledged, and possibly question-begging, understanding of cognition. The Coordinated System Approach (CSA) is predicated only upon the notion that a cognitive system is responsible for the production of goal-directed behaviour. This is hopefully something that many of the players in the debate can sign up to.

An advantage of this approach is that the CSA applies to a broad range of systems independent of the details of their implementation. This means that the approach can be used to analyse sub-personal systems, personal systems, super-personal or social or distributed cognition, robot swarms, and the like. There is no preferred scale or format built into the method.

Embracing a general characterisation of cognition (or a weak cognitive agnosticism) means that, unlike other approaches, a central role for representation is not baked into the project at the start (see Adams, F. and Aizawa, 2010b; Aizawa and Adams, 2005 for example). If it turns out that representation of a minimal kind drops out as a consequence of the approach, as we shall discuss in part III, then all the better.

These ideas do not occur in a vacuum. They are a synthesis of three major clusters of sources. The quote from Ashby above sets the scene for a systems approach inspired partly from the cybernetics and general system theory movements in the middle third of the Twentieth Century (see for example Ashby, 1960; Bateson, 2000; Carlson and Doyle, 2002; Conant and Ashby, 1970; Hooker, 2011, 2011; Ladyman and Wiesner, 2020; Pickering, 2010; Rapoport, 1986; Rosenblueth *et al.*, 2017; Thurner *et al.*, 2018; von Bertalanffy, 1969;

Walter, W. G., 1950; Wiener, 1961)[42]. The task of characterising goal-directed behaviour is guided by another important mid-Twentieth-Century movement through the work of the psychologist Edward Tolman (1948, 1967). Underlying some of the ideas in Part II is a third tradition spanning the Twentieth and Twenty-first Centuries concerned with constraints, normativity, and the dynamics of self-organising systems[43]. This research has uncovered connections between these clusters not visible on the surface.

The notion of system is central to our project and is thick enough to support some of the argumentative tissue. The basic challenge is the following: because systems are typically embedded in larger systems and contain smaller systems, demarcating them is a challenging task. Furthermore, there are a multiplicity of systems with a multiplicity of boundaries. What criteria can be used to pick out the relevant boundary? These challenges arise due a fundamental embeddedness; systems are never isolated. But it is this embeddedness that allows normative and teleological talk of goals and tasks fundamental to the CSA. Systems are embedded not only in causal systems but also in normative ones. Some systems face tasks because they are imposed from the outside by the systems in which they are themselves embedded. Others face tasks because their own self-maintenance mandates specific kinds of interaction with the environmental systems in which they are embedded.

[42] For a concise history see Hofkirchner and Shafranek (2011) and for an insightful in-depth investigation of the English cyberneticists see Pickering (2010).

[43]I implicitly draw upon an extraordinary body of work on constraints in thermodynamics, systems, normativity and autonomy especially Howard Pattee (1969, 1971, 1973, 1979, 1983, 2013) but see also Wayne Christensen (1996, 2004, 2012), Mark Bickhard (2009a, 2009b, 2011; Christensen and Bickhard, 2002), Terry Deacon (Deacon, 2013; Deacon *et al.*, 2014; Leijnen *et al.*, 2016), Alvaro Moreno and Matteo Mossio (Moreno *et al.*, 2008, 2011; Moreno and Mossio, 2015; see also Winning and Bechtel, 2016, 2018), and Xabier Barandiaran (2017; Barandiaran *et al.*, 2009). I shall only scratch the surface of this work in the thesis.

The system is an explanatorily virtuous item. At least part of the job of picking the appropriate boundaries in the demarcation programme will depend on our explanatory interests. A system is something that contains the essential explanatory resources needed to understand how behaviour is directed towards a set of tasks chosen by the investigator. It is explanatorily encapsulated in the sense that one does not need access to resources outside the system to explain the production of behaviour in respect of a set of tasks. Of course, this is not to say that systems so described are any less real. The braking system of a car is no less real because it is demarcated in relation to a particular set of speed-reduction tasks.

There are many questions raised by this account that will be discussed in the following chapters. How is goal-directed behaviour characterised? What does it take for a system to produce goal-directed behaviour so characterised? What do we mean by task? What exactly is coordination? What standard of explanatory adequacy is implied by the CSA? How can we distinguish system processes from behaviour?

There is a sense in which the chapters in this part are written backwards. Because of the complexity of the material, Chapter 3 gives an outline of the main claims regarding the nature of coordination processes and the role that coordination plays in establishing an argument to HEC. It does so without much justification or presentation of detail. Coordination is illustrated using simple mechanical examples even though the intention is to use the CSA to analyse complex systems.

The chapters that follow pick up the threads and provide the missing details and justification. Chapter 4 develops the main outlines of a theory of goal-directedness and makes links with the coordination conditions. Chapter 5 unpacks how we should understand function in the context of this thesis and argues that there is enough heft to functional explanation to identify systems.

Finally, I emphasise that the aim of this thesis is not to provide a general theory of cognition but rather to propose an outline of an organisational theory of goal-directed action.

Chapter 3

The Coordinated System Approach (CSA)

> The study of distributed cognition is very substantially the study of the variety and subtlety of *coordination*. One key question which the theory of distributed cognition endeavours to answer is how the elements and components in a distributed system – people, tools, forms, equipment, maps and less obvious resources – can be coordinated well enough to allow the system to accomplish its tasks ” (Kirsh, 2006, p. 258 emphasis added).

3.1 Introduction

The aim of this chapter is to articulate the main theoretical claim of this thesis: that systems responsible for goal-directed behaviour can be distinguished by their organisation. Specifically, they involve a functional unit responsible for responding appropriately to tasks in the environment and coordinating processes in the system as part of the performance of such tasks. I should warn the reader that I shall introduce this coordination function in a fairly sketchy manner at this point and skate lightly over some points that need further discussion. My aim here is not to give a thorough grounding for the argument. Rather it is to set out the main claims and illustrate them with some simple examples.

Before I set out the conditions that put some flesh on the notion of coordination, I want to give the notion of ‘system’ a real job to do in the explanation. To do this I shall place it in the context of an ontology that is rooted in processes rather than in things or substances. I shall start by unpacking process and then showing how we can think of systems as being coalitions of processes. I shall compare processes to mechanisms and show how the preferred notion of system in this thesis is amenable to top-down process-determination, something that, traditionally at least, has not been a feature of mechanisms.

Towards the end of the chapter, I set out the structure of the argument based on the coordination conditions listed here and expanded upon in the chapters that follow.

3.2 Process ontology

The introduction to this chapter asserted that systems are constituted by processes. In this section, I shall briefly sketch the characteristics of processes that will help us understand the nature of systems as conceived in this thesis.

Process is a series of activities over time that exhibits some degree of coherence or continuity (Rescher, 2000). Nicholas Rescher describes this in terms of a programmatic structure; something like "characteristic patterns of sequential occurrence" (2000, p. 26) or "[t]he concept of programmatic (rule following) developments is definitive of the idea of process". Perhaps most useful for our purposes is "[a] process is an actual or possible occurrence that consists of an integrated series of connected developments unfolding in programmatic coordination: an orchestrated series of occurrences that are systematically linked to one another either causally or functionally" (Rescher, 2000, p. 22). Here are John Dupre and Dan Nicholson in their *Manifesto for a Processual Philosophy of Biology*: "A series of activities constitute an individual process (…) [when] they come together in a coordinated fashion to bring about a particular end" (Dupre and Nicholson, 2018, p. 13). Built into this definition is a hint of teleological function - processes possess coherence because they are directed towards a function or an end. The next three chapters is an attempt to flesh out these notions regarding cognitive systems.

Temporal structure is the key to understanding the nature of a process. A process is not to be thought of as a succession of static time slices but rather as having a coherent temporal existence. There really is more to movement, say, than successive stills on a cine film. This means that two identical states at an instant may actually be spatial parts of two quite different processes. This is something that is quite familiar to us. Two bowls of water, identical at a molecular level (thus with the same temperature), may nonetheless be part of two different processes; the first part of a heating process because the ambient environment is warmer than the water, while the second is part of a cooling process because the ambient environment is cooler than the water. Identity conditions for processes are different to those of things. Whatever these conditions are, and it is a far from trivial matter (see for example Seibt, 2004, 2018; Seibt and Rescher, 2018), they must involve the temporal nature and unfolding of processes. Processes are not captured by their state at a given instant in time.

Moreover, it matters whether we regard something as a process or not. As Rescher, Seibt, Dupre and Nicholson point out, it matters because it fixes the burden of explanation (Dupre and Nicholson, 2018; Rescher, 2000; Seibt, 2018; Seibt and Rescher, 2018). Traditional substance ontology consisting of objects and the relations between them (properties being simply unary relations) takes for granted the continuation of existence of the objects and their relations. Substance ontologists are burdened with explaining change. Process ontologists, on the other hand, take processes to be the fundamental entities in the world and change to be primitive. For them what requires explanation is stability.

Another key difference for our project is the nature of the part-whole relation in processes compared to a substance ontology. Johanna Seibt has written that General Process Theory (GPT) sees processes as dynamic entities (Seibt, 2018, p. 133). GPT holds that there is one basic relation that holds between processes namely the relation of ‘being part of’ “in its most basic sense of ‘belonging with’” (2018: 117). She gives examples of ‘blogging is part of life’, ‘music is part of God’s universe', 'learning to negotiate is part of the advocacy process'. Importantly this part-whole relation is irreflexive, antisymmetric, and intransitive. Blogging is not part of blogging, and if blogging is part of life, then life is not part of blogging. The intransitive feature underlies some of the more surprising results of the later chapters of this thesis. Seibt gives the following example:

> (…) (I)onization is part of hydrolysis, which is part of proteolysis, which is part of the adaptive immune reaction by B-cells, which is part of the human immune system. Unless we read 'is part of' in a narrow sense as 'is a spatio-temporal part of', it is false to say that ionisation is part of the human immune system - the functional organisation that the latter term identifies normally does not 'reach that far' i.e. it leaves indeterminate how proteolysis occurs (Seibt, 2018, p. 117 fn8).

If I have got her right, Seibt seems to be saying that the functional organisation of a system might not reach down to processes on the smallest scales. The organisational reach of the highest level of the system is limited. This does not mean that chaos reigns at the lowest levels either. In a hierarchically organised system, there may be lower-level organisational elements that constrain these processes. This would be a bit like a manufacturing firm in which some of the lowest level processes that contribute to its functioning are outsourced.

It is perhaps surprising that process 'belonging' is not transitive. Afterall with things it is accepted that something that belongs to a part belongs to the whole but for processes this does not follow. If we think of processes and subprocesses as a family tree so that a subprocess belonging to a process is like a child of its parent process, then it is possible for grandchildren to be processually disconnected from their grandparents. This is important for the demarcation problem because it suggests that what processes belong to a system is not just a spatio-temporal question. Starting with a system and then drilling down through subprocesses of subprocesses one may end up with a subprocess that does not belong (in a process sense) to the system. A familiar refrain of systems theorists is that the processes in a system are not neatly separated from the rest of the world; the system may have vague boundaries and the world can leak in from the sides. But this view of processes suggests that the world can also leak in at the bottom.

The idea that functional organisation of a system might not reach down infinitely far is attractive for a number of reasons. It suggests for example that multiple realisation is a structural consequence of process thinking. The processes that are ‘outsourced’ by the system might be realised in a variety of different ways since the details are outside the control of the system, just like a firm that outsources component manufacture to an independent manufacturer might not worry about the details of the manufacturing process as long as the component satisfies the specifications (effectively constraints on the manufacturing process).

Additionally, this picture of processes contrasts with that of mechanisms portrayed in chapter 2, in which sub-mechanisms of a mechanism sat nicely within the spatio-temporal envelope of the whole. On this view once you are inside a system you stay inside the system however many mechanistic levels one descends. Taking a process view discourages ‘smallism’, the idea that systems are best understood by reducing them to properties of features at their smallest spatial level. While I am not claiming that smallism is an official doctrine of the new mechanists, it is implicit in the ‘more details are better’ argument promoted in some of the writings of Kaplan and Craver (Craver, 2006, 2007; Craver and Kaplan, 2020; Kaplan, 2011). In the process view carrying out a mechanism-like reduction ends up outside the process!

Johanna Seibt writes that the non-transitivity of the ‘belonging’ relation allows formal representation of “emergent parts of processes and feedback structures (…) emergent products of an interaction dynamics that causally influence the conditions under which the interaction dynamics occurs and is further propagated (…) as it occurs in self-maintaining systems such as organisms” (2018, p. 119, 2009 see also, 2015). In other words, the GPT provides the right conceptual space in which to frame the ideas I want to explore in this chapter: that systems are not cleanly differentiable from the environment like machines, that coordination of system processes may emerge from the mutual interaction of those same system processes, and that understanding them may not require a detailed investigation of their mechanistic or causal substrate, but rather an understanding of the overall dynamics of this interaction[44].

The sceptical reader might object that a view of systems as processes sits uneasily with the overall project of this thesis which is to arrive at a plausible method for demarcating cognitive systems. Processes evade the neat containment of a system within a boundary that is characteristic of machines, through their tendency to cross such boundaries. I entirely accept this as a limitation of the project. As Dupre and Nicholson (2018) rightly assert, processes possess fuzzy or indeterminate boundaries and are individuated not so much by where they are than what they do. This might mean that the project of determining the extent of cognitive systems should be thought of in terms of identifying key processes involved in such systems through their function and perhaps giving up on trying to draw a definite and determinate boundary. I shall explore this in the next section. The rest of this thesis will be an attempt to do this and to show how demarcation so conceived is supportive of the HEC.

3.3 Mechanical systems, processual systems, and emergence

In this section I want to sketch out a notion of system within a framework of process ontology. The aim is to be able to frame the question of cognitive extension in terms of coordinated processes. I argue that such a framework builds on the insights gained in the debate so far and avoids some of the pitfalls. In

[44] Not everyone agrees that the details of mechanistic implementation may not matter for the explanation (Krickel personal communication).

these terms the notion of system actually does work in the debate and hence membership of the system, or as we shall discover later in the chapter, the core of the system, matters.

3.3.1 Mechanistic systems and systems of processes

To start with a basic question, what is a system? The concept of system is often treated as a contrastive notion. Things are systematic when the properties of the whole depend on the organisation of the parts. Change these relations then the properties of the whole are changed. The contrast class I shall call ‘heap’. The properties of a heap do not depend critically on the relation of its parts – they are in some sense crudely aggregative. An arch made up of bricks can be thought of as a system. The properties of the arch depend on the relation of the bricks that compose it. Change the relation of these bricks, say by removing the keystone, and the properties of the whole change – the arch can no longer bear a load and collapses. A heap of bricks, on the other hand, is not systematic since we can change the relation of the bricks to each other without (substantially) changing the properties of the heap.

There are two rather different ways of expressing the idea of system. They are linked with more fundamental metaphysical questions about whether it is best to conceive the world in terms of relatively enduring things and their properties or in terms of intrinsically dynamic processes and their properties. This debate is often couched in terms of what ontology to appeal to: the ontology of ‘things’ (or substances), or the ontology of processes.

The mechanistic explanations discussed in chapter 2 seems to take a ‘thing’ ontology for granted. Systems are taken to be sets of enduring entities which exhibit productive activities. Mechanisms are made up of sub-mechanisms that are spatio-temporal mereological parts. The mechanism of a kitchen timer marking a duration of an hour is contained within the spatio-temporal extent of the casing. There is never a situation where a bit of mechanism sticks outside the space-time envelope of the whole. Because of this it makes sense to think of the system as being contained within a closed envelope in space and time. Transactions with the environment therefore take place through relatively well-defined input and output channels. Indeed, there is a clear-cut system-environment boundary.

Not only is the mechanistic system contained inside a clear boundary, but its parts and their relations are relatively persisting enough to make sense of the notion of ‘fixed architecture’. This is what constitutes the system independently of its input conditions. The system as a whole can be thought of as a relatively persisting unit of organisation through which something like energy, information, or activity flows. A clock is a wonderful example of such a stable system.

Relative persistence of a system allows theorists to transform talk about processes into talk about parts or components and their properties and arrangement. The process of removing exhaust gases from the cylinders of an engine can be translated into the activities of the piston, the tappets, and the exhaust manifold. There can be systems, such as the car engine, in which this kind of translation is appropriate. One might look to a mechanistic explanation when one is reasonably confident that the system is decomposable into parts, their local properties, and their relations.

The underlying metaphor of this kind of system is a machine. It is ubiquitous in cognitive science and philosophy of biology, perhaps to its detriment (see Moss and Nicholson, 2012; Nicholson, 2012, 2013, 2014 for criticism of this notion in biology). A well-worn example in the literature is a washing machine. The boundary of the machine is its white metal casing. There are inputs to the internal processing: water, electricity, soap, conditioner, programme selection and dirty clothes passing from the environment to the machine in clearly defined input channels. There are clearly defined outputs: dirty water and (hopefully) clean clothes. A consequence of thinking of systems in this way is that there is something different about the components that are internal to the system compared to the clothes, water, soap, and electricity that come in from the outside and cross the boundary, from the subsystems that control the various washing events. For example, inside the machine there are low voltage control circuits including pressure switches and temperature sensors. There are also subsystems that drive the various stages in the form of motors, brakes, heaters and pumps connected to high voltage supply. Neither the control circuits nor the higher voltage washing components cross the boundary of the machine. They remain entirely enclosed within it and perform a narrow repertoire of activities. The interactions amongst these components do not have to undergo the transformations characteristic of inputs or outputs of the system.

While this kind of machine-like system may be useful in some situations it would be wrong to think that its assumptions of a clear-cut boundary with the environment, its fixed architecture and fixed properties of its components are true of all systems. I argue that there are many systems, especially those that figure prominently in the extended cognition debate, that violate these clean-cut conditions. This motivates looking at a different kind of system.

The second kind of system is best described as a collection of processes that is tied together in some way. These constituent processes do not have fixed properties and can be constrained by high-level system features. By this I mean that what the processes do might be dependent in some sense on the working of the whole system. In a system consisting of ants foraging in the forest, the pheromone traces left by the actions of the ants themselves coordinate their further action; the system as a whole and its action determines the properties of its component processes – in this case pheromones determining ant movements. The system is not so easily screened-off from the world but is entangled with it. Rather than saying that the ant system takes an input and produces an output it is perhaps better to think of the system effecting a transformation of a set of world states or processes into another set of world states or processes[45]. The state that exists prior to the operation of the ant system involves separate ants' nest plus food source. The ants 'operate' on the world, being entangled with it, or incorporating it into their operations. The final state is ants' nest having assimilated the food source. The system acts on the world to transform its state. In the transformation process the system may become entangled with it and involve parts of the world in system processes. This is a way of thinking about input and output when the system does not have clear boundaries. Of course, it is often better to think about the inputs and outputs as being processes rather than states. In this kind of situation, it is of limited use to think of something flowing through fixed architecture like current through wires or water through pipes.

This way of thinking of the system as a transformer of processes or states in the world is more general than the input-output picture since the latter is just a special case of an entangled system in which the manner of the system's interaction with the world is limited to clearly defined input-output channels. Therefore, the washing machine example is also amenable to analysis as an entangled system.

What is at stake here is the 'container metaphor' as described by cognitive linguists George Lakoff and Mark Johnson (1980, pp. 30–32). If systems are processes and processes do not have clear-cut boundaries, then systems don't either. I propose that the container metaphor is not useful in many of the cases that I examine in this thesis and that in some cases it leads us astray in our investigation. In the remainder of this thesis, I shall frame systems concepts in terms of the second kind of system that may be entangled with the environment and consists in interlocking processes and treat the first kind of system as being a special case.

[45] In accordance with a process view I think of a state as being a stable kind of process. I shall abuse the term consistently in what follows.

It is tempting to think of these two kinds of system as being quite distinct, but further thought reveals that they are extreme points on a continuum. Even in paradigm mechanical systems like the washing machine the parts do not possess exactly fixed properties because they wear down – it is just that these processes are on a much longer timescale than that of the washing processes.

Similarly, we should not think that a system in a processual view is entirely bereft of rigidity or infrastructure. More persisting elements just have a slower ‘turnover’ to use Mark Bickhard’s phrase[46] and may not be the main determinant of system behaviour. A given process will have a repertoire of behaviours which can be constrained by the behaviour of the whole system, in the sense that system-level features can dynamically affect its properties.

At this point, a key objection to a process view of system emerges. If we accept the vagueness of system boundaries, especially when systems are entangled with the world, then the question arises as to how it is even possible, in principle, to demarcate cognitive systems. Does the adoption of processual systems force us to abandon the whole project? I shall argue later that a system responsible for goal-directed behaviour must perform certain core coordination functions. These functions help identify a core functional unit in the system which will, as far as is possible, identify the core of the system. System-environment entanglement is not a problem for establishing the HEC provided we can identify the coordination core.

[46] Comment at online Interactivist Summer Institute organised by Itay Shani, Jedediah Allen et al June 2021.

3.3.2 Emergence

The kind of feature that we shall use to characterise cognition, at least in the interesting cases, emerges from the interaction or organisation of the relevant system processes. As Wayne Christensen puts it:

> The ontology of this type of system crucially includes *relational* properties which emerge through the interaction of the component physical parts of the system. These properties cannot be reduced to those possessed by the system's physical parts considered individually, because they are contingent on the *organisation* of those parts. (1996, p. 302 emphasis original)

Before I try to capture more precisely what I mean by emergence, it might be helpful to visit some examples. There are both inanimate and biological examples in the literature of systems that exhibit some kind of emergent coordination, from spontaneous symmetry breaking in crystals and magnetism (Gillett, 2010; Laughlin, 2005), the formation of paths in the snow (Goldstone and Roberts, 2006; Goldstone and Theiner, 2017; Helbing *et al.*, 2001; Helbing, Keltsch, *et al.*, 1997; Helbing, Schweitzer, *et al.*, 1997), the behaviour of certain Eumenid wasps, social insects and flocks of birds (Heylighen, 2016; Marsh and Onof, 2008; Resnick, 1997; Theraulaz and Bonabeau, 1999), swarm robotics (Spezzano, 2019), to the phenomenon of Mexican waves at sporting venues (Sims unpublished).

Processual coordination is not limited to high-level biological or social systems. Robert Laughlin describes the process of symmetry-breaking where matter collectively and spontaneously acquires a property absent in the underlying rules by which this matter interacts. This can be seen, for example, in the formation of crystals.

> (O)rganizational principles can give primitive matter a mind of its own and empower it to make decisions. Symmetry breaking provides a simple convincing example of how nature can become richly complex all on its

> own despite having underlying rules that are simple (Laughlin, 2005, p. 44).

Symmetry-breaking refers to the propensity of systems of interacting physical entities to ‘choose’ a certain option such that it correlates with the other entities in the system, such as a spatial alignment of the magnetic fields of iron particles or molecules in crystals. Because of emergent ordering of matter that is not ‘put in’ to the system at a lower level, it is important to recognise that organisation at a high-level cannot necessarily be explained by adverting to a lower level, as another Nobel prize winner P.W. Anderson points out.

> The behaviour of large and complex aggregates of elementary particles, it turns out, is not to be understood in terms of simple extrapolation of the properties of a few particles. Instead, at each level of complexity entirely new properties appear, and the understanding of the new behaviours requires research which I think is as fundamental in its nature as any other. (Anderson, P. W., 1972, p. 393).

The organisation of the whole system feeds into the features of its parts. This might be described in terms of ‘levels’: that system-level properties constrain the properties of processes that make up the system, in the sense that they reduce the number of degrees of freedom available to them.[47]

Now I turn to characterising emergence in the sense used in this thesis. There is a long history of controversy surrounding this notion in the literature and I shall not spend time summarising the debate (see for example Bickhard, 2004, 2011; Broad, 1925; Corradini and O’Connor, 2010; Gibb *et al.*, 2019; Humphreys, 1997, 2016; Powell and Dupré, 2009; Silberstein and McGeever, 1999; Walsh, 2013; Winning and Bechtel, 2019). All I require is that there is an intelligible notion of non-epiphenomenal system properties: properties of the system as a process that

[47] I do not speak of top-down causation specifically, preferring to stay agnostic as to what sort of top-down determining relation is at work here.

can exert some kind of determining relation on constituent processes in the system. Evan Thompson’s definition captures the sense in which this thesis understands emergence:

> A network, N, of interrelated components exhibits an emergent process, E with emergent properties, P, if and only if: (1) E is a global process that instantiates P and arises from the coupling of N’s components and the nonlinear dynamics, D of their local interactions. (2) E and P have a global-to-local (”downward”) determinative influence on the dynamics D of the components of N (Thompson, 2007, p. 418).

The literature yields examples in nature of the kind of thing Evan Thompson has in mind in his definition. For example, Michael Anderson discusses starburst amacrine cells (SAC) in the retina of mammals in as *enabling constraints*.

> (A)n *enabling constraint* is a relationship between entities and/or mechanisms at a particular level of description and a functional systems at the same or a different level, such that the entities/mechanisms bias (change the relative probabilities of) the outcomes of processing by the systems (Anderson, M. L., 2015, p. 11 emphasis original).

I interpret this to be a kind of emergence in Thompson’s sense. It is the higher-level organisation of the cells that bestows functions (in this case directional sensitivity) upon their parts through constraining them in some way. In another paper with Vicente Raja, Anderson raises the intriguing possibility that behaviour of a system itself may act on its own components as an enabling constraint (Raja and Anderson, 2020). This is profoundly interesting from the point of view of this thesis and is something that I explore in terms of stygmergic systems in part III[48]. The point here is that the functional properties of something like the neuron is, in the mechanistic picture, explained by the way in which it is constituted by its parts (Craver and Bechtel, 2007), which is not the case for the SAC, according to

[48] Anderson tends to refer to systems in terms of things and relations, but Vincente Raja thinks that there is scope for framing this approach in terms of processes (personal communication).

Anderson. Therefore the “[mechanistic formulation] is not wide enough to capture the variety of mechanisms in the brain” (Anderson, M. L., 2015, p. 12). This is because there can be system level features that constrain or determine its constitutive processes – Thompson’s emergent processes. Such processes will play an important role in our account and hence give another reason for moving beyond a purely mechanistic model of systems.

Beate Krickel (personal communication) argues that there is nothing in a mechanistic worldview that rules out such top-down determination or loopy systems. My response is that in the standard formulation of mechanism of, say, Craver (2007), the central idea is that a mechanism accounts for the phenomenon by constituting it. But the entities and activities that make up the mechanism can be thought of as phenomena that are themselves mechanisms whose properties are accounted for, in turn, by the entities and activities that constitute them. Therefore, the arrow of explanation is always upwards from lower-level organisation to higher-level phenomena; it is a style of explanation that is based on nested localisation.

Another good set of examples that satisfy Thompson’s loopy emergence condition are dynamical systems that exhibit continuous reciprocal causation (CRC) (see Clark, 1997). This occurs when a system component A is coupled to another component B in a feedback loop such that A’s behaviour causally influences B, but that, in turn, B’s behaviour causally influences A. Such a causal loop typically engenders complex dynamics and is the centrepiece of embodied or enactive accounts such as that of Anthony Chemero (2009), Dan Hutto and Erik Myin (2013, 2018) and dynamical systems approaches such as that of Randall Beer (1995, 1998, 2000; Beer and Williams, 2015; Psujek *et al.*, 2006; Williams and Beer, 2010), Tim van Gelder (1998, 1999), and Orestis Palermos (2014). Palermos realises the central role played by such causal feedback loops by making them the defining feature of his account of cognitive systems. He indicates its central role by appealing to the non-locality of explanations involving CRC:

> (…) in such cases of continuous mutual interaction, the postulation of a single coupled system brings explanatory value. That is, the postulation of coupled systems is necessary with respect to the explanation of certain systemic properties, which we would otherwise be at a loss how to account for. Accordingly, coupled systems are not open to the common eliminativist line that Xs do not exist because our best explanations are not committed to the existence of Xs (i.e. that positing Xs does no explanatory work. Coupled systems (…) must therefore be taken as real. (2014, p. 32).

Explanations essentially involving CRC, and processual systems in general, are stylistically different to mechanistic ones. CRC units are explanatorily indivisible in the sense that the properties (say stability) of the whole cannot be determined by just looking at the properties of a single unit in the system. The individual dynamic units of a CRC system give rise to a system-level dynamics which determines the dynamic properties of the individual units. The classic example here is the phenomenon of Bernard cells. Oil under the right thermal conditions self-organises into hexagonal patterns of convection. Heating causes movement of oil particles which produces cells which constrain the further movement of oil particles hence the whole exhibits continuous reciprocal causation. There is nothing about the local properties of the oil (which is taken to be undifferentiated) that will give information about where the cells will form. It is the emergent organisation of the whole that places constraints on individual particles and therefore causally determines the geometry of the pattern.   In an ant’s nest the activity of a single ant seems baffling. The ant seems to be behaving randomly. It is the reciprocal causal coupling with the whole system of ants that makes them an explanatory unit and therefore suggesting that explanation at the system level is simpler than describing the nature of the interactions between components. Even in the simplest case of a two-component reciprocally coupled dynamical system (where the state variables of one are the parameters of the other), it is not really possible to describe the dynamics in terms of components' inputs and outputs. Restricting attention to one component does not give you half an explanation – it gives you no explanation at all. The system must comprise both dynamical components.

3.4 Groundwork for the coordination function

Having prepared some of the conceptual machinery for the approach let us begin to put it together. I shall do this largely without discussion for simplicity and clarity. The rest of part II will expand and develop these ideas.

The starting point for the Coordinated System Approach (CSA) can be summarised in six commitments:

(1) Cognition is taken to be a feature of a system responsible for intelligent (goal-directed) behaviour.
(2) A system should be understood in the sense of a set of interconnected processes operating on different timescales.
(3) These processes are loosely or tightly bound in a functional whole.
(4) To produce goal-directed behaviour, or do 'useful work', the processes of the system need to be coordinated.
(5) The coordination function is a core function of the system and is likely to emerge from the interaction of multiple system processes.
(6) The processes that realise the coordination function are core components of the system.

A basic motivation for these six starting points is a fuller version of the Dupre and Nicholson quotation from Section 3.2:

> Processes are individuated not so much be where they are as by what they do. A series of activities constitute an individual process when they

> are causally connected or when they come together in a *coordinated* fashion to bring about a particular end. Many of processes found in the living world, moreover, exhibit a degree of cohesion that demarcates them from their environment and thereby allows us to identify them as distinct, integrated systems - as entities in their own right. (Dupre and Nicholson, 2018, pp. 13–14 emphasis added).

The key point is that processes are individuated functionally, "by what they do". What binds them together into an integrated whole is that they are *coordinated*. I propose to explore in more detail what this means and to use it to demarcate cognitive processes from their background environment. The aim is to derive functional conditions for coordination that will help in describing the system in the wild. In the cases where the system is integrated in the way that Dupre and Nicholson describe above, identification of the coordination processes will also identify or demarcate the system, insofar as processes responsible for functions can be demarcated (see Seibt's point in the previous section).

For these reasons, this part of the thesis is focused on investigating what it means for processes in a system to be coordinated. Coordination will itself be a process, perhaps an implicit or emergent one, that is not necessarily separate from the other processes that make up the system. Whatever it is in a system that performs this coordination function, it is responsible for the intelligent, or goal-directed property of system behaviour. It might not be the same as the processes involved directly in the production of the behaviour, but it is responsible for *triggering* these processes in a time-critical manner in response to the environmental, and system conditions. If one conceives of the system as providing a repertoire of performances each of which is relatively simple, then the coordinator takes care of the scheduling of these performances in such a manner as to produce a coherent response to environmental conditions.

Coordination, so conceived, is constraining in the sense that it is implemented through placing constraints on other processes in the system. This amounts to reducing their degrees of freedom to act in the world or, as Anderson put it, changing the probabilities of their producing certain outcomes. The banks of a river constrain the flow into a channel meaning that the path of a given water molecule possesses one extended dimension along the direction of flow and two rather restricted dimensions laterally and vertically. Given the existence of the banks it is likely that a given water particle will, over a period of time, travel downstream. This would not be the case if the banks were not there. Of course, the banks themselves are processual. What makes them constraints is that the timescale on which the bank process occurs is much longer than that of the movement of the water in the river.

A river is a system that nicely illustrates some of the features of processual systems that we shall need for our account. A river creates banks and gorges and the like for itself which constrain its constituent processes. The creation of such constraints is a nice illustration of the loopiness of continuous reciprocal causation. The river system has rather fuzzy boundaries (the banks are created by the river and constrain it, but do they belong to it?), and does not possess clear-cut input and output channels with its environment. Moreover, rivers possess a kind of dynamic stability (what I shall call robustness) regarding their tendency to 'want' to flow downhill. Place a large rock in the path of a stream and eventually the stream will flow round or over it.

What rivers do not have (to any great extent at least), and the kinds of system that concern us in this thesis *do* have, is the ability to respond to changes of goal – what I might call adaptiveness in the service of a larger distal goal such as maintaining the conditions for continued existence of the system. Rivers do not avoid situations in which their existence is threatened, for instance by skirting areas marked out for future hydroelectric schemes! For simplicity I shall take all these tendencies to be examples of goal-directedness with the understanding that I am primarily more interested in systems that are adaptive. Cognition requires something of a balance between stability and sensitivity to the environment.

Goal-directedness in this more substantial sense is a feature of the CSA account and endows it teleological normativity. The system faces a set of tasks in the world as part of its pursuit of a goal which fits nicely into the observation by Rescher, or Dupre and Nicholson that processes possess a fundamental directedness. A basic goal-directedness in biological systems may derive from the fact that such systems exist as open systems in a far-from-thermodynamic-equilibrium state, a state that is unlikely to occur by chance flows of energy. This means that the system must actively maintain itself to survive. Indeed, it also needs to maintain the kind of relationship it has with its environment in order to survive perhaps by changing its environmental niche in a way that makes its survival more likely. Mark Bickhard calls this *recursive self-maintenance* (Bickhard, 2004, 2009b; Christensen and Bickhard, 2002; Deacon, 2013; Deacon *et al.*, 2014). Bickhard points to a candle flame being self-maintenant because it stabilises itself at exactly the point where the temperature at the bottom of the flame is enough to melt the styrene to produce a combustible vapour, the current of air produced by the temperature differential sucks in oxygen to support the burning and the movement of the hot gases upward from the flame ejects waste products from interfering with the reaction. But after some time, the candle burns itself out. It cannot maintain its environment to continue its self-maintenance by, for example, adding new sources of styrene to its bulk. For a complex system like a living organism, recursive self-maintenance operates as a distal goal for the system[49].

[49] Of course, organisms die eventually, but this is not necessarily because they there are no recursively self-maintenant processes as in the candle example.

Given a particular environmental situation the basic goal translates into a set of tasks. A task is the transformation from the current state of the world to a goal state or process. Maintenance of the human body temperature at 37C in an ambient temperature of 40C is a task. Serving a tennis ball within the baseline is a task. Removing cola cans in the lab by a robot is a task. Tasks derive their normative force from a normative framework in which the system is embedded. In the case of the body temperature this normative framework derives from the goal of basic survival of the organism. In the tennis example the norms for playing tennis operate. In the case of the robot the norms are imposed through its human programmers. This thesis is less concerned with the origin of the normative framework than its existence; noting that some philosophers propose that they are derivable from the basic requirements of organisms' recursive self-maintenance (see Christensen and Bickhard, 2002; Piccinini, 2015; Satne, 2015). I shall say more about tasks in the next chapter.

The coordination process will need to be sensitive to the tasks facing the system. This means being adaptive to changes in task. A task might change as a result of environmental conditions changing. A robot clearing up cola cans in the lab faces a change in task when an experimenter drops another can on the floor. The human body faces a different task if the temperature drops to -5C. The tennis player changes the serving task when playing doubles. But a task can also change if the environment stays the same but the distal goal changes. If the goal of the robot is changed to seeking a power socket, then the task changes. This is important with complex systems with goals linked perhaps to social environments where goals are not the direct result of system self-preservation but rather of complex cultural frameworks of norms and practices.

3.4.1 Coordination conditions

Given that the coordination function is central to this account we will need to unpack precisely what we mean by coordination. In the literature coordination is variously described as: non-accidental correlation, or coupling of systems, leading to the emergence of new features or behaviour (De Jaegher *et al.*, 2010), the orchestrating functioning of mechanisms (Bechtel and Abrahamsen, 2005), “the management of the interdependencies of activities” (Consoli, 2016, p. 141) or competitive or collaborative ‘transactions’ (Huebner, 2014). The cyberneticist Francis Heylighen defines coordination in terms of the effect it has on the participating processes: alignment, division of labour, workflow, and aggregation (2013). These notions, while helpful, might not be sufficiently precise to function as demarcation criteria. Instead, I propose the following functional criteria with respect to processes and tasks:

A subsystem (possibly the whole system) performs a coordination function relative to a set of tasks τ if the following conditions are satisfied:

(1) It *tracks* and *triggers* processes that are responsible for performing tasks in τ.

(2) It is sensitive to changes of tasks in τ. This means that there is a ‘regular’ relation between the task faced by the system and the functional role played by the coordinator[50].

The rest of this part of the thesis will be devoted to developing and understanding these conditions.

As a simple example let us consider the electrical ignition system in a car engine. The distributor consists in a rotor arm assembly attached by a gear to the camshaft. As the shaft rotates it makes a contact that sends a short electrical impulse to the spark plug in each cylinder of the engine with the correct timing so that the spark combusts the petrol-air mixture in the cylinders at the appropriate moments. As input then it takes the camshaft rotation and as an output it provides

[50] I shall be more precise about the nature of this relation in the chapters that follow.

timed electrical impulses. Checking the coordinator conditions is straightforward. The distribution process triggers the combustion process in the cylinders by supplying the impulse that gives rise to the spark. The position of the rotor arm tracks the combustion process in the sense that it corresponds to the correct point in the combustion cycle for each cylinder. Condition (1) is therefore satisfied. For condition (2) we must think about how the task faced by the distributor changes. It faces a one-dimensional set of tasks - namely providing the correct timing for the current rpm of the drive shaft. But task sensitivity is satisfied because the casual structure of the distributor is such that the rotor is attached to the drive shaft by a gear. If the drive shaft spins faster, then the timing is faster. The distributor of a petrol engine therefore performs a coordination function for the task of timing the combustion in the cylinders.

Note that in this example the distributor coordinates a subsystem of the whole engine. The task to which it is sensitive is subordinate to the control of the engine itself which depends on the driver depressing the accelerator pedal. Coordination exists at different levels within a system and some systems such as a car engine can be thought of as a well-defined pyramid of processes and subprocesses with a corresponding cascade of linked coordination functions.

In this case the coordination function is performed by a dedicated module. But this is neither a required nor a desired condition. Indeed, the idea that the coordinator is a stable persisting thing is, in many cases, what Whitehead describes as the ‘fallacy of misplaced concreteness’ which is a danger inherent in taking as paradigm mechanical examples like this (Whitehead, 1978, p. 7). A much better and more likely situation is when coordination is distributed throughout the system, as evidenced by many of the examples of interest in this thesis. From the point of view of getting a grip on the extent of the system, these distributed coordination functions are much to be preferred. Living systems are often like this. Here is the philosopher of biology Denis Walsh writing about higher-level task coordination in organisms:

> Insofar as any single entity can be said to ‘control’, ‘regulate’ or ‘orchestrate’ this widely distributed plexus of causes, it is the organism as a whole. As a self-organising, self-synthesising, self-regulating, purposive system, the organism has the capacity to co-ordinate and marshal these causal influences toward the attainment of a stable, viable, adaptive system. (Walsh, 2015b, p. 158).

Andy Clark also admits of the possibility of distributed coordination although he puts it in terms of control which as I point out in the next section is a word I want to avoid. “(C)ontrol is itself fragmented and distributed allowing different inner resources to interact with, or call upon, different external resources without such activity being routed via the bottleneck of conscious deliberation or the intervention of an all-seeing, all-orchestrating inner executive” (2011a, pp. 136–137).

However, there will be cases, in more modular systems, where coordinative and non-coordinative components are distinct. I have already hinted at this in the distinction between cognitive processes and those which are purely implementational, which I propose correspond to coordination processes and *plant*, inspired by Wayne Christensen (1996, 2007).

I want to make a further couple of general remarks about the coordination conditions before I turn to a more detailed discussion of them.

The first is that coordination as set out in the conditions is a task-relative function. More precisely, whether something is a coordinator or not depends on a set of tasks being investigated. A given subsystem may play a coordination role with respect to one set of tasks but not with respect to another. The distributor in a car engine coordinates spark plug timings but does not coordinate fuel injection. Using the coordination function as part of a strategy for demarcating systems will have the consequence that what counts as a system depends on the set of tasks that is of interest to the investigator. This is a perfectly normal situation in the natural sciences and engineering (and parallels the phenomenon-dependence of mechanisms that we encountered in the previous chapter). The previous example of the braking system of a car is quite different to the electrical system – they face different tasks. Moreover, neither system is less real for being functionally individuated. Surely this is the lesson of process thinking - processes and the systems they are part of are individuated with the help of function[51].

The second point is that the coordination conditions are functionally integrated and co-dependent. It is because of the tracking function that the triggering function can operate. To achieve the right coordination dynamics triggering is conditional upon the right sort of tracking of other processes in the system. This is also true of the task sensitivity condition. The whole functional organisation of tracking and triggering depends on the details of the task. If the drive shaft is spinning at speed R_1 then the tracking and triggering functions face a task of providing timing T_1. But if the drive shaft is spinning at R_2 then the task faced is to provide timing T_2. In a relatively simple system such as a car engine where the tasks are indexed by a single dimension such as the speed of the drive shaft, the mechanical structure produces the right relation to the task through causal organisation.

3.4.2 Coordinative tendencies in the literature

In the previous section we saw that there were references to coordination in a wide swathe of literature. Notable, given the discussion in the previous chapter, coordination is central to Bechtel and Abrahamsen's definition of a mechanism: "The *orchestrated* functioning of the mechanism is responsible for one or more phenomena" (Bechtel and Abrahamsen, 2005, p. 423 emphasis added). Despite this hint, the new mechanisms literature does not take the idea of coordination

[51] I am keen to avoid the conflation of the two phrases 'individuated with the help of function' and 'individuated by function'. The connection between a process and its function is a complex one. Functions may be multiply realised, while a single process might be multifunction.

further. Indeed the idea of emergent coordination functions does not fit well into the ‘classical’ new mechanisms picture of Craver and Kaplan (Craver, 2001, 2007; Craver and Kaplan, 2014, 2020; Kaplan, 2012; Machamer *et al.*, 2000).

The interactivist turn in social and distributed cognition research takes coordination of interactions between autonomous agents as a key idea. David Kirsh is one of an increasingly influential group of psychologists and cognitive scientists to point up the importance of coordination: “coordination is the glue of distributed cognition and it occurs at all levels of analysis” (Kirsh, 2006, p. 250; quoted in Sutton, 2006, p. 237 see also the quote at the head of this chapter). Bryce Huebner anticipates the coordination conditions (2014, Chapter 4) in his discussion of social cognition, although he limits himself to a computationalist view of cognition and is doubtful whether there are many genuine cases of what he calls *macrocognition* by which he means materially or socially extended cognitive systems.

> (…) goal directed behaviour is typically implemented by distributed networks of specialised subsystems, each of which produces local idiosyncratic (…) representations. Competitive and collaborative “transactions” are then employed to coordinate the outputs of these subsystems in ways that allow the system as a whole to cope with salient changes in the world (…). These claims recommend the possibility of a kind of collective mentality that arises through the coordination and integration of computations that are carried out by specialised individuals (or perhaps smaller groups). (2014, p. 90).

What is interesting is that Huebner allows the transactions between system processes to be competitive as well as collaborative. This is surely correct. There is no reason why higher order system features cannot emerge from competitive interactions. Moreover, Huebner also points to the requirement that these interactions make the system sensitive to salient changes in the world which in our account would be couched as changes in task.

Perhaps some of the closest precedents to these ideas can be found in the work of enactivist or interactivist philosophers. Simon Høffding and Glenda Satne write:

> (…) interactionism claims that the processes constituting cognition are themselves interactive. The paradigm has produced evidence that spontaneous coordination unfolds in stable patterns among agents in interaction and describes these interactive processes as unique dynamical systems themselves constituted by autonomous systems in interaction. (Høffding and Satne, 2019, p. 5428).

De Jaegher *et al* build upon the interactivist turn: “interactive processes are more than a context for social cognition: they can complement and even replace individual mechanisms” (2010, p. 441). They point out that more work needs to be done in scaling up interactive constitution in a dynamical system to account for more sophisticated mental acts of self-reflection and planning. There is much in common here, especially the idea that coordination functions emerge from interaction rather than being imposed on it by a control system ‘piggy-backing’ on top of other modules. However, none of these authors develops a detailed account of coordination or links it to the HEC.

I want to end this section by pointing to some interesting work in plant biology by John Dupre and Ozlem Yilmaz that identifies coordination as a means of identifying plant individuals:

> “(c)ognition is located in a specific cogniser which is capable of producing action as a coordinated whole. Given the modular nature of plants, it is not easy to say what exactly the cogniser is. Still, since plants have complex net of processes that produce coordination and appropriate action in response to environmental changes (even producing systemic responses), they are clearly cognisers but in their own way” (forthcoming).

The general idea that a complex precarious system possesses a time-critical coordination function is nothing new in the literature. However, as far as I know, its central role in an explanatory framework for goal-directed systems, and its functional characterisation as envisaged here, has not previously been investigated.

3.5 The causal structure of coordination processes

The tracking and triggering functions of coordinators place constraints on the kinds of causal system implementing them. This section discusses what the causal structure of these implementors might be.

3.5.1 Process stages

First a word about process stages. It is customary to think about processes in terms of stages. Indeed, for Rescher, a sequence of events without stages is not a process. “The basic idea of process involves the unfolding of a characterizing program through determinate stages” (2000, p. 26). He goes on to write:

> A *process* is a sequentially structured sequence of successive stages or phases that themselves are types of events or occurrences (in the case of an abstract process) or definite realisation of such types (in the case of a concrete process). The structureless sequence – just one darn thing after another – is not a process. (2000, pp. 26–27).

For example, the process of building a house consists of a number of stages. First the foundations are laid down. Then the structural features of the house such as walls and floors are constructed. Later the roof is added, and then the electrical system is put in and so on. Each stage of the process can be identified by characteristic properties. In the first stage there is no house to speak of – just a lot of work going on in laying foundations and the pipes and other structures that will be under the finished house. In the second stage, the house gradually takes shape and there are walls and floors in various stages of completion but no roof. The third stage sees the construction of beams and rafters and ends with the laying of roof tiles and so on.

The existence of process stages is important because they both track and trigger other processes. For example, only when the foundation is complete can builders start work on the walls. The state of the foundation tracks the foundation laying task and only when it is complete (something that expert builders recognise) does it trigger the next stage in the process.

3.5.2 Tracking and triggering

The diagram below shows this at work. Let us take process P_1 to represent the house-in-the-making during the first stages of construction. These stages are indicated by the labels S_1, S_2 and so on. We can assume that there are initial processes such as planning and so on (stage S_1). When these are complete, process P_2 is triggered, say building the foundations, initiating stage S_2 of the overall process. When this is complete it marks the completion of the stage and initiates the next stage S_3, for example letting the concrete in the foundations dry. When the concrete is dry it triggers the next process P_2, say building the walls, on so on.

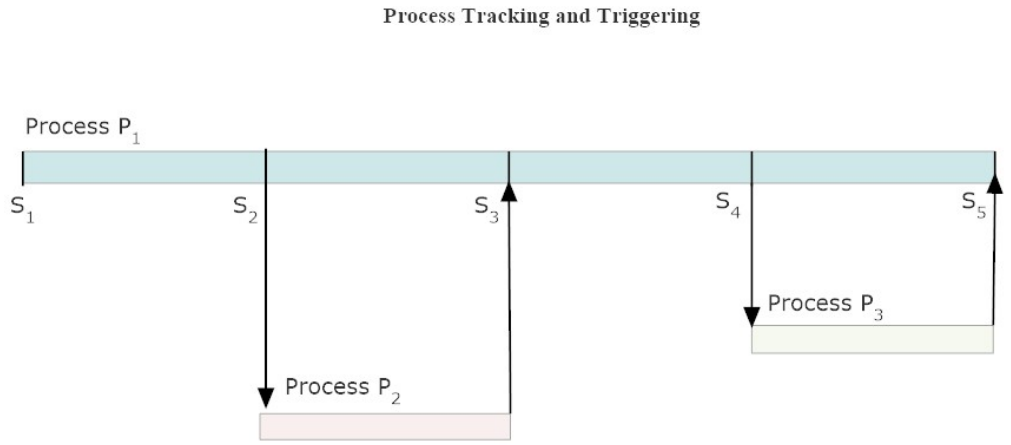

Fig. 3.1 Diagram showing how different stages of process P_1 trigger processes P_2 and P_3.

The process that is the house-in-the-making P_1 can be divided into stages which simultaneously track and trigger the various building processes. The tracking and triggering are functions that are implemented through causal relations that are indicated by arrows on the diagram. These causal relations are between processes (or process stages). We can think of this kind of process causation in terms of examples like the wearing-away of a riverbank by a river or the tipping up of a balance between a solid weight and a dish of water as the water evaporates. There are formal theories of process causation such as that of ‘activity causation’ of Krickel (2018b, pp. 81–90), or those defended by Anscombe (1981) and Salmon (1994; 1998). A discussion of this topic is a book in itself – it is sufficient that there is a defendable account of such a notion.

The key idea here then is that the functions of tracking – marking the stages of a process – and triggering – the initiation of a new process – can be implemented by a set of causal relations with the right structure. Moreover, that tracking and triggering, when they occur together as part of the coordination function, are functions that are intimately bound up with each other. In a coordinator these functions are often implemented by the same kinds of process as we saw with the building of the house. The different stages of the construction track the construction processes but at the same time trigger new construction processes. I shall say more about this example in chapter 7.

3.5.3 Information and control

In the example of the washing machine, it might be legitimate to speak of coordination in terms of control. There is a (low voltage) subsystem that takes care of the coordination of the various washing processes. Such a subsystem is functionally and physically distinct from the (high voltage) circuits implementing these processes. However, this thesis avoids framing coordination in terms of control for two main reasons. Firstly, I want to highlight the fact that in most of the systems of interest the coordination function emerges out of the interactions of the processes that make up the system. There is no distinct control subsystem. Examples can be had aplenty where the system consists of a coordinated collection of agents such as human beings in a football team where the coordination function emerges from the interaction of the players. Secondly, I want to draw a distinction between coordinators and *commanders*. A commander is simply a system element that issues commands but is not itself sensitive to the state of the system's processes. Commanders possess triggering functions without attendant tracking functions – they are units that provide a one-way line of determination. I discuss commanders in more detail in the next section.

Similarly, it is tempting to frame the tracking and triggering functions involved with coordination in informational terms by thinking of them as logic gates. Some readers may immediately think about the role that representations could play in such a coordinating structure for instance in the tracking function. But the current account does not make any commitments to either an informational analysis or representations. The kinds of systems to which we shall apply these notions may be representation-heavy or not. The account remains agnostic. All that is necessary is that the causal structure implements the functional structure.

Having said this there is something broadly informational that we can learn from the cybernetics literature. The first insight is that the system must have at its disposal a behavioural repertoire adequate to the variety of salient environmental situations facing it. This is known as Ashby's law of requisite variety (Ashby, 1956, p. 206). In terms to be discussed in the next chapter, this means that the coordination function should possess at least as many degrees of freedom as the task space it faces. If you want complex behaviour (to perform complex tasks) then, roughly speaking, you need to have complex control dynamics. But Conant and Ashby take this further and write that the control system should effectively be a model of the situation controlled (1970). This is highly suggestive for what follows because it raises the possibility that the control system can, as Clark hinted earlier, be situated out there in the 'environment', perhaps part of what is being controlled. Rodney Brooks (1991) puts this in terms of letting the world be its best model. I have more to say about this in section III.

As we saw tracking and triggering are intimately and reciprocally linked. This is a reminder of a central theme of homeostatic loops in the cybernetics literature and the importance of causal loops, ‘coupling’, and CRC in more recent cognitive science texts. One way of interpreting the coordination conditions is as a way of fleshing out the intuition that such loops are characteristic of cognitive systems.

3.6 Logical independence of the coordination conditions: would-be coordinators

In the previous section I emphasised that the coordination conditions are functionally integrated which means either that they are emergent properties of a non-modular system or, if the system is modular, then there are two-way causal links between the modules responsible for the different functions.

But the careful reader might wonder, then, whether the coordination conditions are conceptually distinct and whether we need all of them. I shall deal with the first question and defer the second to subsequent chapters where I offer a more detailed justification of the conditions. Are there examples that satisfy some of them and not others? In this brief subsection I shall show that there are indeed cases where a system aspires to be coordinative but does not quite make it.

Satisfaction of the triggering but not the tracking condition can be found in phototropic plants such as Evening Primrose *Oenothera biennis* or Sunflower *Helianthus annuus*. Like many other plant species these plants possess mechanisms that are regulated by the sun. The key in both cases lies in the name. The sunflower inclines its flowers towards the sun while the flowers of the evening primrose open rapidly when the sun sets. The movement of the sun across the sky certainly triggers behaviours in these plants. However, the tracking condition is not satisfied. The causal arrow only points in one direction. The sun does not record the progress of the opening of the primrose nor the movement of the sunflower. In the previous section I called such processes that trigger but do not track ‘commanders’ since the direction of functional influence is from commander to the system that obeys the command. The sun is a trigger but not a tracker in terms of its role in the relevant processes in these examples. It is a commander.

Likewise, it is easy to find examples of trackers: structures that track system processes but that do not trigger them. Normally speaking the vapour trail of a jet airliner tracks its progress but does not trigger any other processes pertinent to the plane.

Regarding the task sensitivity condition, we can ask whether we can find examples of structures that track and trigger system processes but are not task sensitive. This suggests a system that displays a basic responsiveness to the environment but is not adaptive. This could be a system like a robotic lawnmower that mows the lawn independently of the state of the grass. Consider a machine that simply works on a time interval, let us say that it does a full round of every part of the lawn every week. It is triggered by a clock, and it cleverly tracks its progress by computing GPS coordinates of the areas covered. But suppose that there is a big drought, and the grass simply dies. Then there is no need for the lawnmowing task, yet the machine carries on regardless. Here the task-sensitivity condition is missing.

The coordination conditions are therefore logically and conceptually distinct but come together within the integrated functionality of a coordination process.

3.7 The structure of the argument

Having seen a sketch of the machinery of co-ordinators, this is a good moment to take stock and make a first stab at the rough shape of the argument that we shall use to answer the research question of this thesis regarding the demarcation of cognitive systems.

The coordination argument to HEC.

(1) Cognitive systems produce goal-directed behaviour.

(2) Coordination is responsible for goal-directness.

(3) In order to understand the goal-directedness of the system the explanation must include how the system processes implement the coordination function (because the system is explanatorily encapsulating or (stronger) the coordination function is emergent from the system)

(4) Therefore, if the system is thought of as being a core explanatory unit, then the process realising the coordination function must be part of the system.

(5) Therefore, if an ‘external’ process plays a coordinating function then it is part of the system.

(6) Therefore, if a situation can be found in which this is the case, then HEC is true.

I write ‘external’ in scare quotes to indicate that the component in question is controversial with respect to debates about HEC.

Finally, I want to suggest that the argument is essentially pluralistic. Dupre and Nicholson talk of a *promiscuous individualism* as being a consequence of view of biology that proposes that there are in fact many boundaries that can be of interest when dealing with processes. I am inclined to take a parallel view of systems here which we might call, for the want of a better term, *promiscuous systemhood*. The investigator may, in the right circumstances, draw a line around the coordinator of a system and identify this as the core of the cognitive system in the sense that it delivers goal-directedness. But this condition will not, in general, identify the whole system responsible for delivering the cognitive behaviour. Processes involved in the production of behaviour are just too diverse to fit under a simple criterion such as coordination.

3.8 Conclusion: what needs to be done?

The previous sections set out the coordination conditions and discuss how the conditions should be understood in the context of a thick notion of system based on a process ontology. The conditions were presented with little, or no justification save for the observation that something like coordination figures in the literature of enactive cognitive science, artificial life, and autonomy. This feels rather unsatisfactory and leaves us with important questions to be answered. It is the job of the rest of part II to address them.

There are two main clusters of issues.

1. What is meant by goal-directed behaviour and a goal-directed system? What are goals and tasks and how are they related? Given that we want to open the black box of behaviourism, what can we say about the functional organisation of a system that produces goal-directed behaviour in the sense described? How do the coordination conditions follow from an understanding of the functional organisation of such a system? This is addressed in chapter 4.

2. What is meant by ‘function’ in the argument? Given that a system is, in this thesis, part of the furniture of an explanation, what kind of explanation is an explanation of goal-directedness? What role does (the appropriate kind of) function play in this kind of explanation? How does such a kind of explanation fit with mechanistic explanation? This is addressed in chapter 5.

Chapter 4
Goal-directedness and tasks

> “Intelligent action is for the sake of an end; therefore the nature of things also is so”, Aristotle Physics II Ch 8, 199a (1941, p. 250)

4.1 Introduction

In this chapter we pick up the most urgent threads left hanging in the previous one. I can imagine that the reader is intrigued but disturbed in equal measure by all this talk of goals, tasks, actions, and performances - with the idea of norms and normativity underlying it all. She can be forgiven for thinking that we have helped ourselves a lot of teleological language for free. During the current chapter I hope to repay this debt by developing these notions and sketching a way in which they can be made natural and secure.

Recall that in this thesis a system is cognitive if it produces intelligent behaviour, that is goal-directed behaviour. The aim of this chapter is to characterise goal-directedness in behavioural terms and then, given a notion of system as being a coordinated ensemble of more or less autonomous processes, show that this characterisation places constraints on the functional organisation of the system which mandate the coordination conditions.

The previous chapter addressed the idea that goal-directed behaviour is explained by the system being subject to a coordination function satisfying the coordination conditions. This chapter sets out to perform three tasks: firstly, to describe what is meant by goal-directed behaviour, secondly, to show that the capacity for this kind of behaviour sets constraints on the organisation of system processes, and thirdly, to show that these constraints are satisfied by the coordination conditions. I am broadly guided in the first task by the work of psychologist Edward Tolman and his work on goal-directed behaviour in rats, and in the second by the cybernetics movement in the middle years of the Twentieth Century. My hope is to provide justification for the claim that the coordination conditions on system organisation are sufficient to account for general goal-directed behaviour. It follows that if cognition is the production of goal-directed behaviour then I will have established that the coordination conditions are sufficient for cognition broadly construed.

Historically speaking, the aim is literally to open the black box of behaviourism. Starting with a characterisation of goal-directed behaviour we shall show how the organisation of the system follows as framed in broad terms by the coordination conditions.

How does the project of this chapter differ from the new mechanistic approach discussed in chapter 2? The main difference, as in many debates in philosophy of mind and cognitive science, concerns the appropriate level of analysis. While the new mechanists argue that a detailed mechanistic account is necessary for the explanation of cognitive phenomena, the line taken here is that functional description is adequate, at least for the purposes of demarcating systems. This is not to say that causal details are not important. Clearly, we will need to point to causal processes to be able to make assertions about their functional relations and therefore their belonging to the core of the system responsible for cognition. The focus on the functional level produces an analysis that gets the balance right between causal detail and generality regarding the demarcation problem. I say more to justify this choice in the next chapter.

Another difference between the project of this chapter and the methods of the new mechanists is the deliberate introduction of teleological terms. This account will need to make precise how it understands normative terms such as task, performance, and goal. These notions are contestable, of course, and it is not the intention of this thesis to provide a detailed naturalisation argument for prescriptively normative concepts. Instead, I shall sketch an outline to an argument that these are emergent properties of complex systems and point the reader to existing literature.

The contribution the analysis in this chapter makes to this literature is to identify a key characteristic of goal-directed systems being their sensitivity to changes of task. If tasks are understood as being the relatively local ‘canalisation’ of a distal system goal by environmental constraints (and those of the system), then tasks change when these constraints change. For, example the task of a mouse seeking cheese in a maze changes if the mouse enters the maze from a different point. A goal-directed system is one that can take account of a changed task in some way. It is a key aim of this chapter to establish the importance of this task-sensitivity[52].

Key to many of the following chapters in this thesis is the idea that systems differ in the set of tasks to which they are appropriately sensitive. Later I shall equate the size of this *task space* to cognitive capacity. For now, I shall just alert the reader to the fact that cognition, as conceived here, depends on the set of tasks under investigation. A system may be cognitive, that is, can produce goal-directed behaviour, in respect of some tasks but not others.

To summarise then, there are three aspects of goal-directedness that concern us here: that a system can possess a goal, that it can behave in a goal-directed manner relative to a set of tasks and that it can be organised in such a way as to produce such behaviour. Behaviour is treated first before I move on to system organisation. I leave the question of what it means for a system to possess a goal until last. This is a huge question and one that I do not intend to settle now.

[52] I would like to add parenthetically that despite the difficulties in theorising tasks it is well worth the effort of reaching even a partial understanding of them given the prevalence of this term in the experimental literature in psychology and neuroscience.

Instead, I am content to sketch an argument to support this claim and refer the reader to the copious literature.

The second half of the chapter develops the notion of task. I examine how tasks come in families and how the coordination function of a system responds to changes in environmental conditions that amount to changes in task.

4.2 Goal-directed behaviour

We tend to recognise goal-directed behaviour when we see it. The website of University of New South Wales embryology department shows an extraordinary video of a microscopic pursuit of a bacterium by a neutrophil – a white blood cell (Hill, n.d.)[53]. The neutrophil appears as a large, squashy, cushion-like entity squeezing between the red blood cells, while the bacterium is a small dark agitated dumbbell. When the bacterium changes direction the white blood cell does too. Eventually the neutrophil catches up with the bacterium and overwhelms it. To us observers, the behaviour of the neutrophil is explained by phrases like ‘it is chasing the bacterium’, ‘it is trying to catch it’, or even ‘it wants to catch it’. These statements seem to capture some essential feature of the situation. The bacterium is an object of pursuit. It engages in what are reasonably interpreted as evasion tactics. The neutrophil responds to these evasions by changing direction. It persists with its ‘goal’ despite changing environmental conditions; arguably it uses these conditions to its advantage, and it displays versatility in switching between behavioural regimes[54]. The behaviour of the neutrophil makes sense as a whole, rather than being a sequence of random moves. It exhibits persistence in the face of obstacles and can adapt to

[53] This example appears in *Organism, Agency, and Evolution* by Denis Walsh (2015a).

[54] I use scare-quotes around ‘goal’ to indicate that I do not take the basis for a teleological description for granted but rather want to provide some justification for it. This chapter can be seen as a process of removing them.

unexpected changes in conditions. Pre-reflectively, it is natural to describe the behaviour of the neutrophil as goal directed.

The philosopher of biology Dennis Walsh supports this view in his discussion of Aristotle's *Aitea* in Physics II. He puts it in terms of robustness, something that I shall develop below.

> …[T]here is all the difference in the world between chance events and purposive events. And a good theory of explanation ought to be able to tell them apart. Crucially, purposive events and chance occurrences have different modal profiles. Purposive occurrences *are robust across a range of alternate initial conditions and mechanisms*, chance occurrences are not. (…) For instance, if this were a purposive encounter, we would expect it to be somewhat insensitive to initial conditions like location (Walsh, 2015a, p. 193 emphasis added).

I add here, parenthetically, that it is the robustness of the process across a range of initial conditions that is captured by the coordination conditions – especially that of task-sensitivity. It is the link between the directedness of behaviour and the coordination conditions that I want to highlight in this chapter.

The example of the neutrophil suggests that goal-directed behaviour is something that observers can know. Some writers such as Joulia Smortchkova claim that one can perceive goal-directedness directly and not as a result of an inferential process (2020). Goal-directed action is defined as "efficient motoric means by which the agent achieves its outcome given the situational context" (Smortchkova, 2020, p. 857). She does not need that the agent intends such actions - in fact she does not need that the perceived agent possess mental states at all, attribution of goal-directedness is something that, for example, infants are able to do. They can distinguish such actions from inefficient, random or impossible ones (see Southgate *et al.*, 2008). In this respect there is a famous set of studies by Fritz Heider and Marianne Simmel (1944) in which subjects, presented with animated displays of geometrical shapes, attributed goal directedness to them and described them as 'chasing each other'. Further evidence that these attributions of goal directedness are independent of ideas about a theory of mind can be had in the form of a study of subjects on the autism spectrum – who are traditionally regarded as having difficulty attributing intentions to others. It showed that these subjects exhibited the same tracking characteristics when presented with seemingly goal-directed geometrical figures (Rutherford *et al.*, 2006). Smortchkova uses these studies to make a case for direct social perception including direct perception of goal-directed actions (see Krueger, 2012, 2018).

It should be pointed out that this work is controversial and there are those, such as Shannon Spaulding, who argue that perception of goal-directed action involves inference (2013, 2015). However attractive direct perception of goal-directed action is, I remain agnostic. It is sufficient that it can be inferred – that we can identify some behaviours as goal-directed and others not, however we do this. The interesting question is what is it that characterises such behaviours and distinguishes them non-goal-directed movement? To answer this, let's return to a set of exciting advances in psychology in the middle part of the Twentieth Century.

4.2.1 Edward Tolman and goal-directed behaviour

More familiar to some, perhaps, than the zig-zagging of a neutrophil is the work of the psychologist Edward Tolman regarding the behaviour of a rat in a maze (1948). Tolman considered the rat's behaviour purposeful. His experiments provided evidence that would support a move away from the dominant Watsonian behavioural paradigm of the day. While still calling his theoretical posture 'behaviourism', Tolman embraced the then unfashionable idea that *cognitive abilities* played a role in the rat's response to the varied tasks that it faced. His concern was with a theme that is central to this thesis: what I describe as the response of a goal-directed system to variations in task.

Tolman’s approach contrasted with that of Watson in that the latter took a ‘molecular’ approach to behaviour analysing it into its atomic microphysical components, for example, breaking down a whole grasping movement into a series of micro-physiological components such as individual muscle reflexes[55]. On the other hand Tolman conceived behaviour as being “a type of commerce with the environment” (Tolman, 1967, p. 19). A behaviour like a grasping movement was best understood as extended in time - an integrated ‘molar’ whole (1967, p. 7). This is not just a terminological difference. An act as single behavioural item is processual rather than a series of molecular snapshots. Like Smortchkova’s directed actions, the end point of a smooth movement is its goal.

Tolman’s criticisms of psychological molecularism could be a summary of the shortcomings of some new mechanistic accounts of cognition. He quotes his contemporaries, notably Weiss: “the inability to trace the ramifications of the micro-level is not a restriction on the study of the macro”; and Kantor “psychologists [should be] attempting to express facts about the whole organism rather than its parts” (1967, p. 9). In Tolman’s eyes, behaviour possesses coherence over time, it is a process rather than a series of atomic events.

What is it about behaviour that gives it this coherence? The answer lies in a specific pattern of sub-articulations that gives behaviour a goal-directed character (Tolman, 1967, p. 11). “(B)ehaviour (…) always seems to have the character of getting to or getting from a specific goal object, or goal situation” (1967, p. 10). There are “in-lying purposes and cognitions immanent in any behaviour” (1967, p. 19). Nothing is hidden about these purposes – they are manifest in two features of the pattern of behaviour of the rat. Firstly, the behaviour is coherent - the smaller units of behaviour are integrated to give a larger coherent arc producing

[55] Perhaps this makes the contrast appear too black and white. As Tolman himself points out Watson is at least ambivalent on this matter (1967, p. 7).

the molarity highlighted above. Secondly, the rat's behaviour is persistent - that it aligns with environmental contingencies, and, if they change, then the behaviour changes with it.

One way to think of this is in terms of constraints. As indicated, constraint can be thought of as something that reduces the degree of freedom of a behaviour (the description of the behaviour requires fewer variables), or something that changes the probability of certain occurrences. Let us focus solely on the locomotion behaviour of the rat, getting from its point on the outside of a maze to the cheese at its centre. There are 'internal' constraints that govern the pattern of locomotion of the rat due to its physiology. For example, the skeleton and musculature of the animal play a role in constraining movement possibilities. But there are also external constraints on its behaviour such as the barriers between the passages in the maze and the netting on top of it. Persistence requires that the rat select from its repertoire a behaviour that meshes with the environmental conditions.

I do not want to couch persistence in traditional terms of overcoming environmental difficulties since this way of thinking suggests that the environment contains pre-existing 'difficulties' or 'problems' that exist independently of the abilities of the rat – or that environmental constraints are always somehow limiting – the opposite might well be the case – i.e. that they are enabling. Rather, I want to emphasise that persistent behaviour (as a component of goal-directed behaviour) is an alignment between the constraints inherent in the system and those in its environment.

The characteristic goal-directed behaviour of the system can be summarised, similarly, as the alignment of environmental and system constraints with the goals of the system. This may result in performance of a set of tasks of the same kind but differing in small details or perhaps in the adoption of a quite different behaviour. Thus, a rat facing a gap in its path may jump it. As the gap is gradually increased there may become a point when the gap becomes too large, where the rat adopts a climbing behaviour instead. The shift in behaviour type we might describe as versatility rather than persistence but both kinds are what we mean by goal-directed behaviour – the aligning of constraints with goals or what this amounts to at the highest level of abstraction: an insensitivity of final outcome to initial conditions.

Such behavioural switching requires that elements are added to a behavioural repertoire when a new situation is encountered. Psychologists and cognitive scientists are interested in how this happens. The psychologist Edward Thorndike, a near contemporary of Tolman's called the process by which a behavioural element became available as part of an animal's repertoire 'docility'. A docile behaviour is one in which the environmental (and system) constraints are made to align with the goal of the animal. Thorndike created an experimental situation where a cat is placed inside a cleverly designed puzzle box ('Thorndike's box'). The only way to escape the box was to press a lever. Thorndike placed the cat repeatedly in the box and timed how long it took to escape. "When put in the box the cat would show evident signs of discomfort and of an impulse to escape from confinement. It tries to squeeze through any opening, it claws and bites at the bars of wire (…) The vigour with which it struggles is extraordinary(…) And gradually all the other non-successful impulses will be stamped out and the particular impulse leading to the successful act will be stamped in (…) until after many trials, the cat will, when put in the box, immediately claw the button or loop in a definite way" (Thorndike and Bruce, 2017, p. 35). In the beginning its movements were erratic and aggressive but after a number of trials the cat was able to exploit aspects of its environment to reach the goal of escaping from the box in a smoother series of moves.

Thorndike's box was seen by some psychologists as evidence for operant conditioning (the association of an act and its consequences)[56]. But Tolman saw it more as evidence of a process by which the behavioural repertoire of an animal is enlarged in a manner to make certain features in the environment exploitable or framed in the terms of this thesis as a way creating new tasks that align environmental constraints with the goals of the system.

In Tolman's terms, this process of adaptation endows these features with a 'means-end readiness' for the rat (1967, p. 30). Over repeated trials, the rat gains efficiency in the way that it co-opts parts of the environment in the service of its goal which is to satisfy its appetite for food that is placed in the middle of the maze. The means-end readiness of an object depends on its being exploitable relative to the goal. Blind alleys in the maze that do not lead to satisfaction of the primary goal are left off the list of 'ready' objects. Adaptivity is a long-term convergence in the rat's behaviour leading to smooth switching of different elements of a behavioural repertoire that, in conjunction with environmental features, producing efficient paths to an outcome.

> We note two significant features in this description [of Thorndike's cat in a box experiment] (a) the fact of the behaving organism's readiness to persist through trial and error, and (b) the fact of his tendency on successive occasions to select sooner and sooner the act which gets him out easily and quickly - i.e. the fact of docility (Tolman, 1967, p. 14).

At this point I want to interject a comment regarding the significance of Tolman's work in the larger sweep of thinking about cognition. Some readers may have latched on to the modern themes in the paragraph above. It could easily have been written by a contemporary thinker in the enactivist tradition (see Chemero,

[56] Thorndike used this work to establish the Law of Effect – that actions that lead to desirable/rewarding consequences are likely to be repeated. Daniel Dennett comments that the Law of Effect must be involved in *any* possible adequate explanation of behaviour (1981, p. 72). Although I do not put in quite these terms I agree with the thrust of Dennett's comment.

2009; Di Paolo, 2018; Hutto and Myin, 2018; Kiverstein, 2015, 2020a; Stewart *et al.*, 2010; Thompson, 2007). Enactivists see the relation of ‘affordance’ between an agent and objects in the environment as being central to their account. This stems from the work of J.J. Gibson (2014) who was himself a contemporary of Tolman. These general themes are also central (though expressed rather differently) in Part I of Heidegger’s *Being and Time* (1962; see also Dreyfus, 1992, 2014; Kiverstein, 2012). In Tolman, the seeds of an enactivist approach are already visible[57].

Tolman emphasises a further point that is significant to the current concerns of this thesis. Not only is the behaviour of the rat coherent, persistent and adaptive, but that as behaviour “(…) it is always an affair of the organism as a whole and not of individual sensory and motor segments going off *in situ*, exclusively and by themselves” (1967, p. 17). *Eo ipso* goal-directedness in behavioural terms can be understood as a feature of the whole system not of its individual parts. Later he quotes Perry (1921, p. 85): “The organism as a whole is for a time pre-occupied with a certain task which absorbs its energy and appropriates its mechanisms”. This idea is important for two reasons. Firstly, if we are trying to demarcate systems, then the whole system had better do some explanatory work. Secondly it the whole system that faces a task.

Tolman was profoundly aware of the heterodoxy of these passages of the effect of his teleological and intentional language on his contemporaries.

> (…) [S]urely any “tough minded” reader will now be up in arms. For it is clear that thus to identify behaviours in terms of goal-objects, and patterns of commerces with means-objects, is to imply something perilously like

[57] There is a large literature on affordances in addition to the works cited here (see for example Anderson, M. L. and Chemero, 2009, 2019; Estany and Martínez, 2014; Rietveld *et al.*, 2018; Rietveld and Kiverstein, 2014; Walsh, 2015a, Chapter 8). I discuss the links between the CSA and enactivist accounts in chapter 8.

> purposes and cognitions. And this will surely be offensive to any hard-headed, well brought up psychologist of the present day. And yet, there seems no other way out. Behaviour as behaviour, that is, as molar, *is* purposive and *is* cognitive. These purposes and cognitions are its immediate descriptive warp and woof. (1967, p. 12 emphasis original).

For Tolman goal-directedness is an objective feature of behaviour, available for discovery by a skilled observer. "[These purposes and cognitions] are objective and it is we, the outside observers, who discover (…) them as immanent in, and determining, behaviour" (1967, p. 19). Drawing these threads together then we can present the conditions that Tolman requires for behaviour to be goal-directed:

1. Goal-directed behaviour possesses a coherence and a wholeness that distinguishes it from random sequences of movements. It is *process-like.* It makes sense then to think of systems producing goal-directed behaviour as being collections of processes too.
2. Environmental constraints present possibilities for exploitation in goal-directed action. The system can align selections from its behavioural repertoire to these constraints to conduce to this action. This can be characterised as *persistence*. It can also be described as the willingness of the system to respond in an appropriate way to environmental changes.
3. In the long term, goal-directedness describes the ability of the system to fine-tune aspects of its behavioural repertoire to recurring elements in the environment - the system adapts its behavioural responses. To put it in Tolman's highly suggestive terms: the system adopts certain environmental features as 'means-end ready objects' in a smooth and competent manner. A consequence of adaptation is that the system can not only deal with environmental perturbations by counteracting them, but that a small environmental change can induce a radically different behavioural sequence associated with a different adaptive pathway.

A final feature that flows from Tolman’s work, and one that we shall come back to, is the observation that, while a system may possess a general high-level goal, the constraints operating in the environment have the effect of canalising it into a series of behavioural tasks. It is to these tasks that the coordination process will have to respond.

4.3 Goal-directed systems

Now I turn to the structure of the system responsible for goal-directed behaviour as described above. I shall propose that a test for a goal-directed system is that it has the right functional structure. If I observe a robot that exhibits consistent movement in the direction of Coke cans in the lab, I might conjecture that it is a goal directed system. If, however, I open it up and find only a simple set of servos inside attached to a radio receiver I might deduce that the robot itself is not responsible for this goal directedness, but that its source lies elsewhere – that the system is ‘extended’. What Smortchkova or Tolman give us in their conception of goal-directedness is a clear description of a phenomenon in need of explanation by our (system) theory. Presented with such behaviour the challenge is on to identify the system features responsible for such behaviour.

I have reason to be optimistic in this endeavour. After all, there is a rich vein of thinkers from the cyberneticists and general system theorists of the 1940’s through to the complex systems theorists of the present day, who regard goal-directedness as a feature of the organisation of systems. This is how Dennis Walsh puts it:

> This rich tradition shows us that goal-directedness is an unproblematic causal consequence of the architecture of an adaptive system. It is also an observable feature of the system’s dynamics. It consists in the capacity of a system as a whole to enlist the causal capacities of its parts and direct them toward the attainment of a robustly stable endpoint. That end-point is the system’s goal. (2015a, p. 195).

Wayne Christensen, in a remarkable early paper (1996), draws a distinction with Aristotle (1941) and with more recent accounts of Mayr (1961), Fodor (1981), and Churchland (1989). For Aristotle teleology is imposed from outside the system by a ‘final cause’ existing in the cosmos itself, for Mayr it is something like a linear program or algorithm encoded in genetic material of the organism, while for Fodor or Churchland it is associated with representations in the system. For Christensen on the other hand “(…) teleology is a property of certain types of complex cybernetic systems. The ontology of this type of system crucially includes *relational* properties which emerge through the interaction of the component physical parts of the system” (1996, p. 302 emphasis original). I think this is fundamentally correct and this section is a working out of these intuitions. Where I depart from Christensen is in rejecting the machine metaphor and its associated notions of transducers, taking functions to be normative and systems to be sets of processes rather than things. Nonetheless, I share Christensen’s basic idea that teleology is an emergent property of complex cybernetic systems.

What sort of system structure or process architecture will give rise to coherent and persistent behaviour of the type identified by Tolman as goal-directed? Clearly, we will not be dealing with specific mechanisms. Tolman’s notions of coherence, persistence and docility are broad-brush notions and any mechanistic implementation will be radically under-determined by them. After all, my ambition here is not to produce a theory of cognition, mechanisms and all. I shall leave that to cognitive science and neuroscience. All I want to do is take a description of goal-directedness, albeit at high level of abstraction, to constrain the functional description of the system organisation in broad terms. Remember, the purpose of this chapter is to support the general characterisation of coordination given in the previous one. We want to get from Tolman’s description of goal-directed behaviour to a set of constraints on the system producing it that are the coordination conditions. This is the essence of the CSA and if I can show that the coordination conditions are necessary and sufficient for goal-directed behaviour in the Tolman sense then I have succeeded in my aim. As stated elsewhere, necessary and sufficient is perhaps too ambitious. My hope is that the conditions that I have laid out are at least sufficient and that the reasons given in this section make them likely if not necessary.

The characterisation of goal-directed behaviour as coherent, persistent and exhibits adaptation, sets up a job description for a candidate system for delivering such behaviour. It must exhibit some sort of process continuity - there must be the right sort of time-critical control of system processes to produce coherent arcs of behaviour, and it must be robust in the sense that it must be able to respond to changes in the environment understood as changes in task.

I shall show that the process continuity property of systems, their knack of producing coherent behaviour from a set of disparate system processes, motivates the tracking and triggering condition for coordination (condition (1)) while robustness motivates the task-sensitivity condition (condition (2)).

I have claimed that these connections are made at the functional level of description. Since this is a contentious claim, it will need some justification which will be the job of the next chapter. For now, I shall assume that function talk is meaningful, explanatory, and carries normative force.

4.3.1 Coherent behaviour and Task Specific Devices

Behavioural coherence is achieved by the system possessing a process continuity property: being able to trigger and monitor a sequence of subprocesses in a time-critical manner. A clear example of this is encountered in driving a car. When I learned to drive in the late 1970's (it might be a little different now) the learner was expected to put together sequences of simple actions such as clutch down, gear lever into first, ease clutch up, at the same time press gently on the accelerator and release the handbrake, all the time keeping a watch for traffic. The order and timing of the actions is crucial. The learner driver lacking habituation must perform each of these actions consciously in a domain general manner - that is as a peculiar task for which no special motoric procedures have been put together (and requiring much conscious effort). It is difficult for the first-time driver to coordinate these actions smoothly and continuously. The learner driver must control all the variables of the movement of feet and hands in real time and unsurprisingly, often fails. Over time, each movement becomes simpler. The degree of freedom of each is reduced – fewer variables are required to control the movement. Learning introduces constraints. Movements are chunked together. No longer should we think of the separate movements of left and right foot and (in the UK at least) the left hand but rather of the whole ensemble as ‘changing gear’. The continuity of processes achieved by the experienced driver results in coherent behavioural arcs - that is - goal directed ones.

But how should we think of this continuity of processes? Which processes are involved? The movement psychologist Geoffrey Bingham introduced the idea of a Task Specific Device (TSD) as an attempt to understand several deep problems regarding the dynamics of the motor systems of human beings and animals (1988). These are actions that are “*softly assembled* (…) low-dimensional, deterministic machine[s] that [are] used to achieve the [tasks] specific to a [goal]. Thus, for instance, launching a projectile [as a bodily movement] is achieved via a softly assembled throwing machine” (Bingham, 1988, p. 240). As the name suggests each TSD is set up to deal with a specific task (or as I shall argue later in this chapter, a narrow set of tasks). Task specific devices considerably simplify a complex set of behaviours. The system simply needs to trigger each device when it is needed and then monitor it to gauge the point at which the next device needs to be triggered. Task specific devices for the learner driver might behaviours like changing gear, doing a hill start and so on. The key point is that task specific devices are what the coordination process coordinates.

Bingham describes TSDs as being smart, deterministic, soft-assembled, controlled, scaled and “assembled over the properties of both the organism and the environment” (1988, p. 250). I shall briefly explain these properties.

They are smart in the sense that they exploit the dynamics of the resources from which they are assembled to perform their task. It is not that they are able to make these judgments themselves and I shall argue that it is the creative capacity of coordination that can assemble such devices and align them to the appropriate dynamics (1988, p. 241). Bingham states that TSDs are deterministic (1988, p. 244). If I read him right this is more like an expression of autonomy. Once they have been triggered the TSD works on its task without the need for any higher control[58]. Soft-assembly is the idea that the TSD is put together on the fly for a purpose and then dismantled (1988, p. 245). It is not something that necessarily persists as a permanent system feature. A TSD is ‘flexibly controllable’ meaning that the task that it faces may occur in a number of metric variations (1988, p. 248). For example, the rate of release of the clutch on the car depends on whether the car is on an incline or on flat ground. A single TSD faces a narrow set of tasks indexed by a parameter such as scale. A throwing movement may be large or small, a running action of a mouse in a maze may be long/short, fast/slow.

TSDs are assembled over properties of both the organism and the environment – a key idea in this thesis. The dynamics on which the TSD trades may be inherent in the human action system, or they may lie in the environment or be a combination of the two. Bingham shows his extended system credentials when he says: “absolutely no a priori distinction can be made between the relative contributions of these dynamics to the behaviour eventually exhibited by the [human action system] when organised as a TSD” (1988, p. 250). Finally, Bingham mentions that a TSD has the potential to be modified to serve a new

[58] See my earlier comments in Section 3.5.3 on the distinction between coordination and control.

task. This parallels the process by which coordination may add new behaviours to the repertoire to enhance the adaptiveness of the system.

How should we understand Bingham's claims about TSDs? Is this a theory about the system ontology of the human behaviour or is it more of an epistemological model? Firstly, it is intended to answer questions about the dynamics of the system, as seen by an outside observer. Because TSDs involve greatly reduced numbers of degrees of freedom, the mathematical descriptions are more tractable than those of the entire human action system. Seen in this light then TSDs are an epistemological tool. But what is good for the goose is good for the gander. What helps the investigator describe the human sensorimotor control system also might help the system itself perform its tasks. TSDs reduce the risk of a dimensional explosion in terms of the parameters required for control, so their construction is a good design strategy. TSDs can be thought of as an epistemological model but perhaps one that the system itself adopts.

The TSD argument made here (and also referred to in Clark, 2011a, p. 157) motivates the first co-ordinator condition: *the system processes tracked and triggered by the coordination process are TSDs.* Thinking about coordination this way makes it abundantly clear why it accounts for the production of coherent behaviour in the service of a task – what I have called process continuity. The first coordination condition is therefore a consequence of thinking about goal-directed behaviour being the result of the coordination of TSDs. The second coordination condition, task sensitivity, requires stepping back and looking at features of the whole system.

4.3.2 Robustness

On a short timescale, a system is robust if it can deliver the same functionality (in service of its goal) over perturbations in its environment and variations in the performance of its components. Robustness will require strategies that scale with scaleable tasks facing the system, such as jumping gaps, but also require that the system switch strategies altogether at times, such as switching from jumping to climbing. An example of the first case would be changing a scaling parameter in a TSD such as an experienced golfer compensates for the wind in judging a tee shot. The second case would be the use or creation of a new TSD such as a right-handed golfer without enough space to swing in the rough attempting to play the shot left-handed[59]. Similarly, the neutrophil reacting to a change of direction of the bacterium is an example of the first case, while an amoeba switching between linear locomotion and random tumbling, depending on the detection of a positive sugar gradient is an example of the second.

Robustness is had through the organisation of the system as can be seen through simple engineering applications such as the design of aircraft wings and control systems. An aircraft wing is flexible so responds to unpredictable changes in air pressure, but its aerodynamic configuration is itself variable so that the wing geometry can be changed in flight. To take another example from aviation, the Automatic Flight Control System (AFCS) on modern commercial aircraft is designed to be able to deal with unpredictable situations during flight as well as failures in the system itself. It is effectively a set of three separate computers linked to the aircraft control mechanisms. Each unit performs the same function, but each runs on different hardware to reduce the risk of correlated software or hardware errors (called Swiss Cheese errors in the systems failure literature, see Perrow, 1999). In complex systems, robustness will require the system to have enough self-organisation to be able to redeploy its resources in perhaps novel ways to counteract unforeseen circumstances.

The systems biologist, Hiroaki Kitano, identifies four main sources of robustness: system control, alternative or fail-safe mechanisms (redundancy), modularity,

[59] Rickie Fowler did exactly this at the WGC in 2020.

and decoupling (2004, 2007). All four can be illustrated by the aircraft wing example. AFCS provides system control in order that the wing function is robust. The AFCS uses the aircraft gyroscope to detect whether the horizontal attitude of the aircraft is abnormal, and then triggers a correcting behaviour in the inboard flap which rights the aircraft. This is an example of robustness by control. It is also achieved through changing the wing geometry to perform the different tasks of creating lift on take-off and increasing drag to slow the plane when landing.

Redundancy is built into the AFCS which consists of three units to perform a function that could *in extremis* be performed by one of them, hence maintaining functionality in the face of internal component failure. In addition, the system is modular so that the effect of any damage and perturbations can be limited locally and the consequences for the whole system minimised. Modules are more effective if they are hierarchically organised. Kitano points out that modules might not just be physical but are likely to be functional and temporal (2004, p. 831).

Finally, decoupling can occur when a higher-level function inhibits the functioning of a lower-level component. This happens when one of the units of the AFCS gives control information that is different from that of the other two units. The whole system (comprising the interactions of all three units) closes down the errant unit and prevents its further participation in the control system. In general, decoupling provides a buffer that isolates lower-level variation from high-level functionalities. Kitano uses the term *systems failure* in situations where the system is not sufficiently robust or versatile to cope with environmental vicissitudes.

Robustness motivates the task-sensitivity coordination condition (2) that the coordination function should vary in the appropriate way to a change in task. Clearly this is an example of system control in the sense that coordination is a higher-level function in the system that controls plant functions – those functions concerned with performance of a task rather than system coordination. The requirement that the coordination processes are sensitive to changes of task specifically satisfies the need for alternative and fail-safe strategies. Changes of task are mirrored in changes in coordination which results in alternative plant processes coming into play. Modularity and decoupling are less obviously expressed in the coordination conditions. I believe they are there – not in a mechanistic sense but in a functional one. Modularity can be interpreted as the relative autonomy of the TSDs strung together by the coordination process. Decoupling of the control system from the contingencies of the environment is achieved through the separation of coordination and plant functions. Even if the system is not modular, the coordination functions possess determining power over plant functions – that is the essence of emergence. There is something like a functional hierarchy in operation here even if this does not correspond to a causal one[60].

It is important to note that a change in task might mean pursuing the same goal under different environmental conditions (and remember that in a dynamical situation such as an organism in motion, the environmental conditions are changing constantly). It may also involve facing the same environmental conditions but a different goal. In the next section we shall see that this makes sense because tasks are the triadic relation of system goals, environmental constraints, and the capabilities of the system.

4.4 Goals, tasks, and task spaces

The focus of the last two sections has been trying to make sense of intelligent or goal-directed behaviour as robust and adaptable – meaning that the system produces the right action in the right circumstances. It has been taken for granted

[60] Writers such as Moreno and Mossio (2015) and Christensen (2007) couch decoupling of control in terms of *second-order* constraints.

that we can talk freely about systems having goals and facing tasks, that these notions have some sort of normative force and can be situated in a naturalistic account of normativity (for an overview of this huge area see De Caro and Macarthur, 2010). By normativity I mean that goals and tasks are the sort of thing that have conditions of satisfaction – they can be performed well or badly – and that the system *ought* to perform them well. Naturalism means that this normativity can be accounted for within our usual scientific account of the world. At the risk of opening a Pandora's box, there is, at the very least, a tension between normativity and naturalism; or as Hume put it, a seemingly watertight distinction between *ought* and *is*. Mark Bickhard points out that there are three unattractive options for resolving it: (1) an anti-naturalist dualism between fact and norm (2) a pan-normative idealism – everything has an *ought* component (3) rejection of normativity entirely identifying naturalism and physicalism (2004, p. 121). As I hinted earlier, it is Bickhard's fourth option that is the solution to Hume's *aporia* that I want to pursue here: that norms *emerge* from non-normative phenomena (2004, p. 122).

Norms are needed to understand the second coordination condition: the coordination process is itself sensitive to the *task* faced by the system. Coordination processes *ought* to configure themselves in response to a task and are responsible for a behavioural performance that *ought* to be successful.

What fact about systems makes this true, that coordination processes are normative with respect to tasks? Certainly, some of the systems we are interested in are self-organising in a manner that ensures their survival. In these cases it makes sense to describe this basic survival to be an emergent goal of the system. This goal is general (and radically multiply realisable) and it is system capabilities (and constraints) taken together with environmental constraints that canalise the distal system goal into a more immediate (and less multiply realisable) set of tasks. The system *ought* to perform the tasks in order, ultimately, to survive. To perform the tasks the system needs to be task-sensitive and self-organise in response to them. This means that normativity of coordination function and normativity of performance (behaviour) is inherited from that of tasks, and, ultimately, from the system goal.

There are other systems, such as, for example, social systems which do not obviously possess the same intrinsic goal of survival. I claim that systems of this kind operate without exception within a normative framework; indeed, they could not exist or operate at all without it. One could speculate that social norms operating on a group have a basis in survival norms of the individuals that make it up, or one could adopt an emergentist view of these norms that they are an emergent property of societies that have very little to do with norms governing individual survival. These are again big questions, and I shall remain agnostic about them. All that is required is that there *are* norms that govern goals and tasks.

But what exactly is a goal? Denis Walsh speaks of a goal as being a future state (or process) towards which the system directs itself or towards which it is biased (2013). Take an organism as a paradigm system that is biased towards survival or a contribution towards inclusive fitness - that is securing the survival and reproductive success of self or close kin. Such a goal, it could be argued, is an objective feature of the world that does not depend on an outside observer to attribute function – being biased towards survival is not something that is observer-relative. Claims like this can be supported through arguments such as that sketched in chapter 3 regarding the survival of a precarious system in a thermodynamically disruptive world (Bickhard, 2000, 2004; Christensen, 2012; Christensen and Bickhard, 2002; Maley and Piccinini, 2017, p. 243; Piccinini, 2015, p. 101). For the purposes of this thesis, I shall accept the general thrust of these arguments.

Artefacts inherit their goals from the norms that operate in the social system of which they are part. A washing machine, the stock example of these kinds of account, inherits the goal of washing clothes from the intentions of its manufacturer and the circumstances of its use. It may also possess goals not envisaged by the manufacturer such as wedging the door open or being a convenient horizontal surface on which to pile books. Again, I shall not rehearse a fully-fledged argument to objective goals but instead refer the reader to other sources such as Piccinini (2015).

System goals can be used to define tasks. Whereas a goal is a general high-level feature, a task is local and specific. A task is a required transformation of the world in the service of a goal such as arriving at an end state or process that itself contributes to the fulfilment of a goal. Note that a transformation may also be thought of as the maintenance of some state or process in the presence of disrupting factors. The autopilot of an aircraft has the task of maintaining a steady course despite the presence of crosswinds.

It is useful to think of a task as being the ‘operationalisation’ or ‘canalisation’ of a system goal via environmental and system constraints. The mouse in the maze has a general goal of survival which requires ingesting food. The environment makes this specific by presenting a limited set of options to the mouse because of the constraints operating both in the environment and the mouse. The maze has passages constrained by walls that convert the goal of finding food into rather specific tasks of running in certain directions along the passages of the maze. The mouse is only able to run on more or less flat surfaces, it cannot fly.

Following Tolman, I emphasise the positive role of environmental constraints in this analysis. Without them, a goal is rather too unspecific to lead to the highly specific behaviours required of the system. Constraints are to be thought of as enabling specific behaviours - they are open to being exploited in the service of a goal. The behaviour of the mouse is channelled (literally) by the maze, but as a consequence, it can exploit the structure of the environment to break down the pursuit of the goal into a sequence of specific behaviours. There is a sense then in which environmental constraints reify, or make concrete, abstract and general goals. They also play a role in selecting the task specific devices (TSDs) which may be relevant to the task. This is how I understand Tolman's 'means-end ready' parts of the environment or Gibson's 'affordances'.

The inclusion of internal constraints in defining tasks helps keep the set of tasks manageable and realistic. Actual or possible competences of the system place a limit on the set of world states that are reachable by the system, i.e. flying is not available to the mouse. A task is, conceptually, a triadic relation between the distal system goal, the capabilities of the system and the constraints in the local environment. It is, if you like, the transformation of the world needed to fulfil the goal given the environmental constraints and the capabilities of the system.

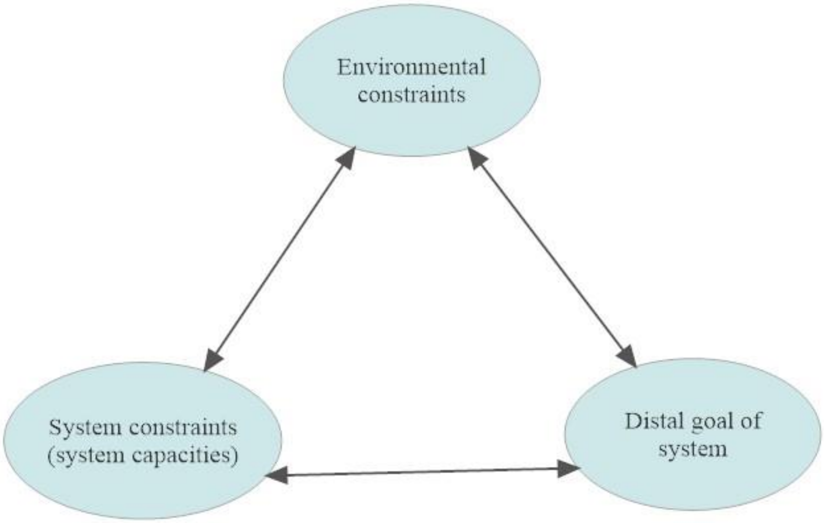
Environmental constraints
System constraints (system capacities)
Distal goal of system

Fig. 4.1 Conceptually, a task is a triadic relation between the distal goal of the system and internal and external constraints.

Environmental constraints therefore help define a set of tasks – or a set of options for the system to exploit in the service of its goal. It is the job of complex systems theory to investigate how the system recognises the tasks reified by environmental constraints and how it puts these tasks together into a coherent sequence conducive to the goal. Different environments generate different constraints and therefore generate different tasks. The key question is how the system responds to these differences – it is this question that is expressed as the second coordination condition – the responsiveness of coordination to changes in task.

There are a couple of worries that the reader might raise regarding this definition. The first is a concern that making tasks relative to system capabilities short-circuits the possibility of task failure. If a golfer fails in a shot, then, surely, she did not have the requisite capabilities. But something feels wrong about this example. A capability is a 'type' while a given performance is a 'token'. The ground for asserting a capability needs to be independent of a given token performance. Some in the enactivist camp such as Julian Kiverstein (see 2020a) appeal to characteristics of whole populations of organisms rather than individuals to establish capabilities. I shall just argue that, practically speaking, the kinds of capability at stake here are, in the first instance, behaviourally established. Although it is a complex matter, we do have ways of inferring capabilities from aggregates of individual performances – this is exactly what (we hope) examinations in schools, universities and driving test centres do. Establishing capabilities can ultimately be established behaviourally.

It seems to me that the question of capabilities introduces an interesting dynamic into the general problem of the extent of systems. If the set of tasks that the system can cope with is partly defined by system capabilities, then by co-opting environmental resources a system can enlarge its repertoire of behavioural responses – that is its capabilities – and therefore increase the size of the task space. In turn, an enlarged task space will require changes to the coordination processes which define the core of the system, which may require co-opting more of the environment, and so on. This is a way in which a system can, developmentally, increase the effectiveness in which it interacts with the world by co-opting parts of that world. Far from being a bug, the mutual dependence of system capabilities, coordination processes, and environment is a feature of the account[61].

A bigger worry is circularity. The aim of this chapter is to create an understanding of what it means for behaviour to be goal-oriented and to derive the coordination conditions as a set of constraints on the organisation of the system so understood. Surely it begs the question to explain goal-directedness by invoking a goal.

A response to this worry trades on the idea that goal of the system that explains the normativity of tasks is an explanatorily distal notion. By this I mean that it plays no direct role in the sort of explanations that are central to these chapters. Saying that a system has a distal goal of survival does not place low level constraints on the organisation of the system – instead it places constraints at a very high level of abstraction – for example, that the system will need to maintain itself and its environment – without saying much about how it should do this. Invoking this distal goal does not tell us much about the functional organisation of the system and therefore does not presuppose an explanation of this organisation. For example, there are many ways that a system can satisfy the requirement of survival and inclusive fitness, witness the variety of lifeforms on this planet. That this Lupin plant has a distal goal of survival does not explain the particular pattern of growth of the plant or its preference for sandy soils. More is needed such as a story about the evolutionary pressures faced by the plant that helped shape the organisation of the processes in the plant. But the system organisation is a ready explanation for goal-directed behaviour in the here-and-now. We can invoke a distal goal, but we still need to explain 'local' goal-directed

[61] Christensen tells an interesting evolutionary story about the development of capabilities and attendant control system. (see Christensen, 2007).

behaviour in relation to the functional organisation of the system. The distal goal supplies the normativity but not the explanation.

4.4.1 Task space

Let us now imagine a complete set of tasks attaching to the rat in the maze. By this I mean that the myriad ways in which the constraints of the maze, coupled with the general capabilities of rats, canalise the distal goal of the rat. The space of all such tasks in every kind of environment I shall call the *task space* of the system. For tasks in the task space the system adjusts itself adequately to perform the task, let us say, most of the time. What interests us here is what form this adjustment takes[62].

The significance of the task space is that it defines the normal range of conditions in which the system typically functions successfully. If we want to understand system characteristics it makes sense to study the systems in the wild within this range. It is here that we stand the best chance of understanding the relation between task and the functional organisation of the system.

Would-be tasks exist that fall outside this range. This happens when a system faces a situation outside its normal operating conditions. A mouse facing a gap of 1000m or a washing machine required to output clean clothes when they have been soaked in paint (Millikan, 2002, p. 120). These are strictly speaking not tasks of the system since they do not lie within a reasonable definition of its capabilities. But there can be *bona fide* tasks at which the system fails. Cricket trousers with particularly bad grass stains might emerge from the machine in a

[62] I do not dwell on an important set of questions surrounding how the system recognises tasks and how it selects between the different task strategies that are available to it. These are clearly important questions that cognitive science must answer, but for the purposes of the current argument it is sufficient to leave the details of these functions blank. I shall come back to these questions in chapter 7 in the specific context of stygmergic coordination.

less than pristine state, a mouse might have trouble jumping 3m in certain circumstances. These tasks lie within the task-space but are nonetheless not performed well.

Equipped with the notion of task space we can revisit the discussion of task specific devices in the previous section. A TSD is a subsystem that is oriented towards a small task space. The task space of such a device will comprise a family of tasks that are similar in kind but perhaps vary in terms of parameters of size and duration. A TSD devoted to ‘overarm throwing’ will encounter different objects to be thrown and different constraints on the throw in terms of distance, arc of throw, whether the throw must meet constraints such as being made quickly or with a straight arm and so on. Within this class then there are an indefinite number of different tasks, but they can be specified by only a few variables. The job of the co-ordinator is to string these TSDs together to allow the system to tackle a much larger task space than that of single TSDs on their own, including tasks such as bowling an inswinging yorker in cricket or skimming a flat stone across a lake.

Not only is the size of the task space of the system a measure of its ability to perform tasks, therefore a measure of its performative power, but also a measure of its success in adapting to its surroundings. The mouse jumping a gap to reach its food will reach a point where the gap is beyond its jumping abilities. In this case the abilities of the mouse and environmental constraints combine in such a way as to impose a task that is outside the task space of the mouse, and system failure is the result.

To remedy system failure, the task space of the system must be enlarged - this is often the focus of a long term developmental or evolutionary approach. It seems plausible that learning, including the processes mentioned above for co-opting environmental constraints, and evolutionary processes generally enlarge the task space of the system. This is where the ‘creative’ aspect of coordination comes to the fore. Novel task performance may require novel behavioural strategies which in turn depend on coordination processes creating novelty. Faced with the large gap between the table and chair the mouse might switch from jumping to running behaviour by running down the table leg, along the floor and up the chair leg, for example.

4.4.2 Of thermostats, and tropistic wasps

As a rule, more sophisticated systems possess larger task spaces because they have at their disposal a larger number of TSDs and hence a larger repertoire of behaviours. The thermostat is a simple system facing a small task space (one-dimensional control space namely the required temperature of the room). The task is to maintain the room temperature at say 21C through the adjustment of the heat output of a heater. In the simpler systems this is done through simply turning the heater on when the temperature is too low as measured by a suitable transponder (thermometer) and turning it off when it is too high - an example of so-called ‘bang-bang’ control where the control variables are either ‘off’ or ‘on’ (Burghes and Downs, 1975, p. 273). As is the case with artefacts, the system could fail because the working assumptions underlying the design are violated (see Stout, 1996). The thermostat, for example, might not function correctly because the room may not be in thermal equilibrium due to a window being open and the patch near the thermometer being warmer/cooler than the rest of the room. The thermostat does not have sufficient self-organising capacities to reconfigure itself when faced with these tasks outside its task-space – the task of maintaining the temperature of a room that is not in thermal equilibrium is unrecognised by the device, and it fails. Working assumptions about operating conditions are a strategy for simplifying design[63]. Relax the assumption about

[63] In fact, the operating assumption enables the temperature of the air to act as, what is called in chapter 7, the ‘salient aspect’ of a trace in a stygmergic system. The thermostat is an example of

thermal equilibrium and suddenly there is a much more complex problem for the thermostat system to solve requiring a more complex set of system processes.

Compare the example of the thermostat to that of the *sphex* wasp which is often used as an example in cognitive science when discussing tropism (Dennett, 1981, p. 65 and p. 245)[64]. The story goes as follows[65]: The wasp stings the prey, usually a grasshopper, to paralyse it. It drags the prey back to the nest and drops it at the opening before going in to check that there are no predators in the nest. If the coast is clear, the wasp brings the prey in to provision the nest for her eventual offspring. If, during the nest inspection, the prey is moved, the wasp will repeat the whole procedure including leaving the prey outside while inspecting the nest. Wooldridge (1963, p. 82) reports that on one occasion the luckless and untiring wasp was made to repeat the same pattern forty times. Another peculiarity of the wasp's behaviour is that it will drag crickets by their antennae, after paralysing them. If an antenna breaks off, the wasp will not drag the cricket by a leg but rather abandon it (see Fabre, 1916). These examples are taken to be evidence for tropistic behaviour, that is, a set routine of specific behaviours that is not sensitive to certain environmental details. In other words, the wasp, already an immensely complex biological system, supports a behavioural system that is effective in normal circumstances in performing the task of catching prey. The task space for the *sphex* is already huge compared to the thermostat. Nonetheless there is a systems failure when the task falls outside the usual range, for example, when the tropistic sequence is systematically interrupted by an interfering investigator. According to the usual story that is told, the wasp is

a stygmergic coordinator. In the problematic case the one-one relation between the trace and the task is broken see 7.5.2.

[64] My thanks to Chris Moore who introduced me to this example in Oxford in 1983.

[65] The original reference to this example is in Wooldridge (1963, p. 82).

not able to adjust its repertoire of behaviours in accordance with what are changes in the task to do with the introduction of new environmental conditions[66].

The point of the story is to illustrate the notion of the task space and link it intuitively to the complexity of problem-solving. The wasp of the story does not supposedly respond to the difference between task 1: dragging prey to opening of nest for the first time and task 2: dragging prey to the opening of the nest for a second time having already checked that the nest is free of predators. This is a clear case of inability to distinguish between different tasks. Moreover, the wasp can perform the dragging task under the condition that the grasshopper antenna affords dragging but the task of dragging an antennae-less grasshopper lies outside its task space. We are in a privileged position as observers and can see that the problem can be solved by an internal re-organisation of the dragging function by hooking it up with a TSD such as a ‘leg-recognition module’. Although the task space of the wasp is large compared to the thermometer it is perhaps smaller compared to the rat.

There are two things going on here. Firstly, the wasp as a system is not sufficiently versatile to respond differently to task 2 than task 1. The robustness of the system regarding the dragging task leads to system failure when the task is repeated disrupted or that conditions are changed to include grasshopper *sans* antennae. Secondly the system is (apparently) not able to adjust its functional organisation in the long-term to produce and enlarge its task space. Not only can the system not cope with antennae-less grasshoppers (this task lies outside the task space of the system), but it cannot enlarge its task space through learning.

[66] Fred Keijzer questions whether the wasp is truly tropistic, or whether, given enough iterations, it will change its behaviour (2013).

Imagine now that a curious cognitive neuroscience student, Zena, inserts some new circuitry in the wasp so that, in an encounter with an antennae-less grasshopper, the wasp adjusts its behavioural repertoire and grabs the leg of the grasshopper instead and drags it back to the nest. The wasp is now able to perform more tasks than before, so the task space is enlarged. However, the behaviour of the wasp continues to be tropistic; the task space is still only enlarged by a single dimension. But suppose now, in a further twist of the story, Zena becomes rather fond of this individual *sphex* wasp, and every time the wasp encounters a new task that lay outside its task space, she swiftly provides a new module for performing the new task – let us say by linking the wasp's sensorimotor systems to a computer that could be reprogrammed/functionally reconfigured as needed. By degrees the system is losing its tropism (or its 'sphexishness' to quote Hofstadter (1985)). But *what* system is losing its tropism? Is it the wasp or is it the wasp-Zena system which is beginning to look like an integrated system? Zena (or her computer) seems to be coordinating the wasp's response to the dragging task. The gist of my argument in these chapters is that this is sufficient to support the conclusion that wasp plus Zena is indeed an integrated system.

To summarise then, constraints in the environment in conjunction with the capabilities of the system and its goal generate a set of tasks faced in pursuing the goal. The coordination process responds differentially to these tasks and ensures the smooth sequencing of their performance which is manifested as behaviour. Under certain circumstances the coordination function expands the list of TSDs that it coordinates thus expanding the task space of the system.

4.5 The task-relativity of coordination

I mentioned earlier in this section that the question of what component(s) play(s) a coordination function depends on the prior question of what set of tasks we are investigating. Since we have set up the conceptual machinery of coordination to make sense of goal-directed systems, this means that whether a system is goal-directed or not depends on what tasks are at stake. A neutrophil is goal-directed with respect to the activities of bacteria but not with respect to red blood cells, the mouse in the maze is goal directed with respect to getting the cheese but not, I take it, with respect to stock market investment (Douglas Adams notwithstanding). Systems are goal-directed with respect to sets of tasks within their task spaces. Which processes play coordination roles depends on the tasks under consideration. We have seen that this chimes with the idea that we can expect specialist subsystems within a system dedicated to specific sets of tasks.

However, if we are using the coordination function in an argument about the demarcation of cognitive systems this result raises some awkward questions. Which set of tasks should be taken to be representative of the relevant cognitive capacity? What can be said about systems facing radically differently sized task spaces?

I shall start with the second question although it leads us to the first one. Recall that the task space of a system is the set of tasks that, typically systems of this kind can successfully perform. By taking the size of the task space as a proxy for the extent of cognitive ability, the CSA offers the possibility of comparison of the cognitive abilities of different systems. The thermostat has a miniscule task space in comparison with the *sphex* wasp, and the *sphex* in comparison with the mouse, hence a corresponding comparison regarding cognitive abilities. The thermostat is minimally cognitive while the human being possesses considerably greater cognitive power. Scaling cognition in this way is a recipe for cognitive gradualism – cognition itself being a scalable attribute rather than an absolute one[67].

[67] This idea is developed further in chapter 6. It is not seen as a bug but rather as a feature of the account.

I shall plant a seed here of an idea that I will develop in part III of the thesis. The comparison of differently sized task spaces may also explain a misunderstanding within the extended cognition literature. Some systems may be put together to serve rather specific purposes, so their task spaces are relatively small compared to a much more general set of tasks. Scientific research might be like this, or navigating a naval vessel, playing Tetris, or walking to MOMA. Adherents of extended or distributed cognition such as Andy Clark or Ed Hutchins would be right in claiming that *in these cases* cognition is extended. Rob Rupert takes basic everyday human tasks involving perceiving, remembering, and moving around the environment as typical. By taking a completely general set of human tasks as characteristic of cognition the relevant coordination processes may well be intra-cranial and Rob Rupert would be right. Put in these terms, the disagreement between Clark and Rupert may be down to a question of what task space to take as a benchmark for cognition. Rupert insists on a large typically human task space while Clark (and Hutchins) consider systems to be cognitive that face smaller and more specialised task spaces. The CSA places these on a continuum and therefore, from this perspective, there is a sense in which they are both right. If you look at the general picture you might appeal to a Rupertian analysis while if you look at a specific kind of cognitive task you might appeal to an extended analysis.

This takes us back the question of associating tasks with cognitive capacities. Right from the start the CSA has associated cognition with the production of intelligent, that is goal-directed, behaviour. It was never the intention that it would be able to individuate cognitive capacities let alone explain them. The production of goal-directed behaviour with respect to a set of tasks is sufficient to assert that there is *some* cognition going on. Therefore, the CSA avoids the circularity difficulties that Kaplan encounters when he seeks to find tasks that are representative of cognitive capacities.

It is of course true that experimental psychology presents tasks to its subjects that are deemed to involve one or more cognitive capacities. The same assumptions are made in cognitive neuroscience when compiling databases of fMRI analyses indexed by cognitive categories (see Eisenberg *et al.*, 2018, 2019; Janssen *et al.*, 2017; Poldrack and Yarkoni, 2016; Varoquaux *et al.*, 2018; Viola, 2016; Viola and Zanin, 2017). But a strength of the CSA is its generality. It demarcates the cognitive system along general functional lines, and this is prior to questions of cognitive ontology. When it comes to the question of the basic functional organisation required to produce performances of tasks, *any tasks*, which I am taking to be goal-directed, the CSA sidesteps these immediate problems. Of course, this might saddle the CSA with another the worry that this makes any task performing system cognitive – but I shall deal with this in Part III.

4.6 Concluding remarks

This chapter has made a start on the job of supporting the claims in the previous chapter. The first task was to establish a clear understanding of what we mean when we can talk of a system's being goal-directed. The strategy was to find a way of describing goal-directed behaviour that led naturally to a general functional characterisation of coordination. The coordination conditions ensure goal-directedness and the identification of processes that realise functions that satisfy them constitutes an explanation of goal-directedness. In the next chapter I shall show that this is enough to establish HEC.

We started by characterising goal-directed behaviour in terms of objective observable features such as coherence and persistence. Tolman tells us that goal-directed behaviour is a coherent process rather than a set of micro-behaviours. The system responds to its environment in a way that makes its trajectory to some extent independent of its starting point. To do this it uses environmental constraints (and internal capabilities) to select between a set of relevant behavioural strategies. Tolman points out general behavioural mechanisms by which behaviours become smoothly adapted to these environmental constraints.

We then moved on to the system side of the story to ask what kind of functional organisation is sufficient to produce such persistent and adaptive behaviour – what produces robustness in a system? To produce coherent behaviour the system must be able to link together smaller behaviour-producing units, TSDs, in a seamless fashion. To do this the system must be able to track the progress of each TSD and initiate new TSD processes in a time-sensitive manner. Hence the coherent aspect of goal-directed behaviour gives rise to the tracking and triggering condition for coordination processes. Selecting between elements of behavioural repertoire requires that the system, in a broad sense, is sensitive to the relation between the environmental constraints, its overall goal, and its capabilities – in short, the task faced by the system. This justifies the second coordination condition – task-sensitivity.

Having shown that the coordination conditions deliver goal-directedness the remainder of the chapter developed an understanding of overall system goals and tasks. In basic systems a distal goal might be something like system survival and inclusive fitness – but it is acknowledged that other kinds of goal may well be emergent features of more complex systems such as social systems. Tasks are more concrete and immediate than goals and are taken to be the way the system goal is canalised through environmental constraints relative to the capabilities of the system (system constraints). Goal-directedness is therefore best thought of as a structural feature of the system in relation to tasks rather than the system goal.

One of the lessons of this chapter is that coordination depends on the set of tasks under investigation. This is not to say that it is an entirely observer-dependent notion; given a specific set of tasks, the relevant coordination is an objective empirical fact of the matter. But care is needed when discussing goal-directedness, or cognition, *per se*. There is only goal-directedness relative to a task space.

Systems are taken to support adequate explanations of intelligent behaviour – they are explanatorily encapsulated – one does not need to refer to processes outside the system to account for the behaviour under investigation. Indeed coordination processes are claimed to be responsible for the goal-directedness of the system with respect to a task space. But coordination of system processes is described in these chapters at a certain level of abstraction – at the level of function. What notion of function is appropriate to coordination of processes that may be deeply entangled with the environment, are not machine-like, and that do not possess clear-cut input and output channels? Is this notion of function sufficient to explain intelligent behaviour? Answering these questions is the task of the next chapter.

Chapter 5
Ecological functionalism

5.1 Introduction

The CSA developed so far looks promising in terms of solving some of the problems identified in Chapter 1. In understanding goal-directed behaviour as being coherent, persistent, and adaptive, the previous chapter showed that the system responsible for such behaviour must possess a certain basic form of functional (self-) organisation as manifest in the coordination conditions. By stating coordination conditions in functional terms, the CSA builds a defence against the coupling-constitution problem discussed in section 1.4.1 as well as the problem of sterile effects discussed in 2.5.3. Functional relations are something more than just causal relations as I will show in this chapter – causal correlation is not enough to guarantee similarity of function. A functional approach may not only reveal a difference between causation and constitution but also separate causally correlated effects from those that are functionally involved in cognition.

If this is correct, then function does significant work in this argument. In this chapter I examine what is meant by function and how it can be used to ground an explanation of goal-directedness and therefore identify the core of the cognitive system. There is certainly work to do here. For a start, traditional functions are defined in terms of transformations of inputs to outputs. Given that the starting point of the CSA is a rejection of the container metaphor and therefore of well-defined and persisting input and output channels, the traditional notion will need to be modified to be able to encompass systems that are entangled with the world and have no clear boundaries. Moreover, systems in the CSA are taken to be self-sufficient explanatory units. They are a minimal set of resources required to explain the phenomenon – in this case goal-directed behaviour. If systems, or their core processes, are picked out by function then there is an assumption that function-talk is explanatory. One way the CSA can be challenged is by denying the adequacy of functional explanation and therefore denying the possibility of picking out systems, as explanatory units, by functional analysis. This might be a position taken by some hard-core new mechanists.

What kind of function fits the bill? We can construct a job description with three main requirements. Because the systems in the account are taken to be facing normative tasks, the notion of function required needs to be normative. Malfunction must be a conceptual possibility. This rules out descriptive function – something has a function by virtue of what it happens to do. Secondly, the kind of function must be able to cope with systems that are entangled with the world such as ant colonies or house builders. Finally, function must be explanatory.

I tackle the first and last job description conditions together in Section 5.2. Section 5.3 deals with the second job description condition regarding a system as a set of processes unfolding over time that may be entangled with their environment without clear cut boundaries and clear input and output channels (see chapter 3). The kind of function meeting all three conditions I call *ecological function.* Ecological function helps solve another puzzle which is how to reconcile talk of system behaviour, traditionally conceived as the output of a system, with a system that is entangled with the environment. If system and behaviour are both sets of processes, how can they be distinguished?

Section 5.4 argues that the coordination conditions are explanatory if couched in terms of ecological function and suggests how they may be used in practice to demarcate cognitive systems.

Finally, I return to the argument set out in chapter 3 to check that all the jigsaw pieces are now accounted for.

5.2 Functional explanation

Why the interest in explanation at this point in the thesis? Let us step back a little. In Chapter 3 we discussed the proliferation of systems – what I called promiscuous systemhood. There are many of them and therefore many system boundaries in the world. I proposed that the appropriate boundary is picked out by explanatory needs. The braking system of the car is picked out by the set of resources needed to explain how the system can deal with the task of slowing the car. The system is picked out by looking at the functional arrangement of processes responsible for performance of the system tasks. Plausibly this would include (at least on older vehicles) the brake pedal, the hydraulics, and the brake pad assembly. We do not need to appeal to anything else in understanding how given a press on the pedal the car slows. Similarly, for cognitive systems, identifying the system responsible for producing goal-directed behaviour *qua* its goal-directedness means exhibiting a candidate set of processes such that we can account for this feature without having to appeal specifically to anything outside them. These processes themselves must account for the observed goal-directedness without, for example, taking highly coordinated external inputs. It is the emphasis on systems being explanatory units that brings explanation to the forefront of the argument.

To add a bit of weight to this point imagine a hard-core new mechanist assessing the CSA. She might pull apart the justification for the coordination conditions from the argument to HEC. “Sure, I accept that coordination function is a high-level description of the core of the system, but this is too abstract to identify any part of the system *on the ground*. What we need is a mechanism. Trying to identify the system by saying that it is necessary for a functional explanation won’t be enough to identify it, because a functional explanation is an incomplete mechanistic one” (see Adams, F. and Aizawa, 2001 for such an argument). There is of course a sense that she is right. We are hoping that the functional description will identify, not a mechanism, but implementing causal processes. But these are picked out by their function. Function tells us a bit more than cause and there is a complex relation between them. Two causally correlated processes may well have different functions – think about the rotations of cogs in a clock. Additionally, the same causal process may have different functions – think about choke and carburettor in an engine (McClamrock, 1991, p. 3).

This chapter makes a case for function as the appropriate level of description of coordination processes and at the same time pinpoints the kind of function relevant to the CSA. I shall start the argument with two remarks.

The first point to note is that once committed to a process ontology, the jump to function is not a large one. In chapter 3 we saw that processes possess some kind of directedness or coherence. Processes are essentially functional entities since we identify them by what they *do*. Activities constitute a process when they "come together in a coordinated fashion to bring about a particular end" (Dupre and Nicholson, 2018, p. 13), The task of identifying processes that implement given functions within a system is part of the everyday business of working with processes. For example, we are used to speaking of the washing and spinning cycles of a washing machine and distinguishing them without any real technical understanding of the details of the machine. This is not to say that identifying the processes that implement coordination functions is going to be easy in every case, but rather that this job is on a par with the kinds of explanation that suffice in many everyday applications. To jump the gun a little, once the coordination conditions are in place, it seems possible to check whether Otto's interactions with his notebook satisfy them or not without any extra technical detail.

Of course, an argument will need to be made that this kind of activity constitutes an explanation of the kind we are looking for. How well do we understand a system by identifying processes with the functions that we have deemed necessary for producing the right kind of behaviour (goal-directed behaviour in our case)? Providing such an argument is the main job of this section.

The second remark concerns the need to supplement or go beyond mechanistic-style explanation for the reasons discussed in chapter 2 (see Chirimuuta, 2018a, p. 894; Walsh, 2012, 2013; Weiskopf, 2011). As Dan Weiskopf puts it: "(…) cognitive models are often non-mechanistic in form. Despite this, they still have explanatory force. Their status as explanations derives from the fact that they are able to capture facts about the causal structure of a system" (2017, p. 61). As I shall argue later, functions allow more generality and capture the essence of a variety of implementations.

But functions do more than this, they support statements such as "the system *ought* to implement processes X, Y, Z in the context of the tasks that it faces in relation to its distal goal and its constraints and those of the environment". I propose that the kind of function in play here is (prescriptively) normative because it describes the organisation of the system needed to perform tasks. Since tasks are normative then so are the functions implemented by the systems that perform them. Functions are prescriptively normative because they contribute to the performance of tasks and the reaching of goals.

Some authors call this kind of function *teleological* because of the relation to goals, and explanations involving such functions are teleological explanations. Denis Walsh writes that this kind of explanation is able to answer *why* questions out of reach of purely causal theories which are more suited to *how* questions (2015a, Chapter 9). Teleological function can be found in a broad class of explanations including computational, psychological, and neuroscientific explanations. We shall explore some of them in the next section.

5.2.1 Functions and functional explanations

Both descriptive and prescriptive functions have the same form. They capture a regular relationship between a prior and a posterior state of the world described in some abstract terms.

$$F: S_1 \longrightarrow S_2$$

Classical function theory envisages S_1 to be an input state and S_2 to be an output state. A machine classically performs function F if it reliably transforms S_1 to S_2 whenever S_1 is input. S_1 and S_2 are specific states or more likely classes of equivalent states with respect to an equivalence relation[68]. In the more general functionalism envisaged in this chapter S_1 and S_2 are just taken to be states of the world and without any assumptions about inputs and outputs.

Used descriptively F is a summary for a connection between states observed in the world. In this sense a river functions to transport water and silt downstream. Used prescriptively F is a general statement of what a system *ought* to do: the waffle maker in my kitchen should transform a mixture of butter, milk, egg, flour and baking soda into waffles. If it does not do this, it has failed to perform its function. Complex arrangements of functions purporting to relate to the world are called functional models.

A functional model of the first type could be a block and arrow diagram showing how the erosion processes of rainfall leads to silt being created and transported via streams into larger rivers. When the speed of the water reduces when the river water comes up against that of the ocean it drops its load to produce the build-up of silt in a river delta.

[68] I note in passing Bickhard's useful distinction between *performing* a function and *having* a function (Bickhard, 2000, 2004; Christensen and Bickhard, 2002) see also the footnote below.

A computer program is a functional model of the second type. Often a programmer will start by sketching a flowchart that represents the different functional units that the program *should* implement to compute the required function. She then goes on to write the program, as set of instructions for the computer to follow, to implement the function. The functional model is a high-level prescription for the causal workings of the computer hardware.

Functional models of both kinds use abstract language to break down the overall behaviour of a system into a particular organisation of smaller (abstract) pieces corresponding to causal processes. In both cases, the functional model tells us something about causal organisation of the system. The difference is that in the second case there is the possibility of error, computer programmers can (and do) get their coding wrong. In the first, the only error is that the functional model does not adequately represent what actually happens in reality. The first type of model is descriptively normative but the second type, and what interests us here, is prescriptively normative.

It has already been remarked that the relation between a function and its realisation is complex. By a realiser P of a function F I mean a process such that, when it is combined with the effects of the other processes in the system it performs F that is it transforms S_1 to S_2. It does this by virtue of its causal organisation in the context of the causal organisation of the other processes in the system. The pumping process of the heart performs the function of producing blood circulation because of the configuration of the system and other functions such as the channelling of blood flow in the veins and arteries and so on. In this sense both function and its realisation are system-wide ideas while at the same time being localisable. The pumping process of the heart is a local process, but it performs the circulatory function because of the functional organisation of the whole system.

Summarising then, functions map prior world states to posterior world states in a regular fashion according to suitable categories. Descriptive functions must satisfy accuracy conditions, what Searle might call a function-to-world direction of fit, while for prescriptive functions it is the system that must satisfy adequacy conditions a world-to-function direction of fit (Searle, 1998, pp. 101–103)[69]. Functions only make sense in relation to other functions in the system, but they are realised by processes that can be local.

Functions can be invoked in functional explanation. Three common forms are:

(1) The system exhibits behaviour B in situation C because it has functional organisation F.

(2) In order to realise goal G in a set of situations X the system *must* have functional organisation F. So, the existence of the realisers of F is explained by G.

(3) In order to realise goal G in a set of situations X the system *can* have functional organisation F. So, the existence of the realisers of F is explained by G.

[69] In this thesis I focus on the notion of *performing* a function. There is a parallel notion of *having* a function often associated with the idea of *proper function*. This is the common argument in philosophy of biology, that the existence of a trait or a structure resulting from an evolutionary process can be explained by having a function for which it was selected (see Hardcastle, 2002; Millikan, 1984, 1989, 2002; Neander, 1991, 1995; Wright, 1973; for an overview see Walsh and Ariew, 1996). Consequently, the theory of proper function builds in a dependence on the evolutionary history of the organism – it is an etiological theory. Mark Bickhard argues that this move makes having a function epiphenomenal (Bickhard, 2004). In a swampman situation where a molecule by molecule copy of an existing human comes into existence, the heart of the human has a function because of its evolutionary history while that of swampman does not. Yet they are identical hence have the same causal powers, therefore the concept of function is irrelevant in the causal universe, hence is epiphenomenal. I am not entirely convinced by this argument because a process can have the same causal powers yet different functions (think of choke and accelerator pedal). But this does not render function epiphenomenal – just that the relation between function and cause is not straightforward.

In each of these explanations F is a model of the functional organisation of the system which I shall explore in more detail below. Explanation type (1) uses the functional model F to explain a dispositional behavioural property of the system – that is the behaviour that the system exhibits given a certain set of input conditions (the situation C). This type of explanation is commonly associated with the work of Robert Cummins (1975, 1983; Roth and Cummins, 2014, 2017).

There are two main criticisms of this approach: that the functions identified in F are descriptive functions, so there is no possibility of malfunction, and that any part of any system whatsoever has a function, namely, to do what it in fact does. As Moreno and Mossio put it, lacking a task or goal structure it "lacks a principled criterion for identifying the relevant set of contributions for which functional analysis makes sense" (2015, p. 66; see also Millikan, 1989, 2002; Neander, 1991, 1995, 2017; Walsh, 1996, 2002, 2013; Walsh and Ariew, 1996). Although, I draw upon other aspects of Cummins' work in what follows, lacking prescriptive normativity, Cummins functions cannot help us understand the relation between function and task and cannot account for malfunction or non-function.

Functional explanation of type (2) and type (3) are *goal contribution* accounts preferred in this thesis. One accounts for the functional organisation of the system, and hence eventually the functions of specific system processes by their contributions to a goal of the system. Type (2) is strongly prescriptive in that there is only one type of functional organisation that will deliver the goal. There will be occasions where the environment and the nature of the system impose strict constraints thus canalising the goal into a single task choice. Situations such as this may admit a strong functional explanation of this form. As Denis Walsh so aptly puts it: "(…) [T]here is no fact of the matter, I take it, whether one ought to prefer to watch FC Barcelona over Real Madrid, but given an intention to watch Barcelona rather than Real, one *ought* to go to Camp Nou rather than Bernabeu" (2008, p. 120). Given a goal of watching Barcelona the system needs to be configured to respond to the task of getting to Camp Nou. The question: 'why is the system configured to navigate to Camp Nou?' is adequately answered by 'because watching Barcelona play is a goal of the system and the only way to watch Barcelona (let us say) is to go to Camp Nou'. Therefore, the functional organisation of the system is explained by reference to this task.

However, it is more likely that there are many different tasks and consequently many different functional strategies that lead to the same goal, prompting the weaker teleological explanation exhibited in (3). The functional configuration of the system is explained by its being *one way* of meeting the goal. The question: 'why is X taking the metro to Collblanc?' is answered by 'Taking the metro to Collblanc is one way of getting to Camp Nou and X wants to see Barcelona play'. It is not the only way to get to Camp Nou, but that does not weaken the explanation.

The functional explanations type (2) and type (3) are ordinarily encountered in examples of artifact design or the operation of machines. Consider the programming example discussed above. Suppose that it occurs in the context of a computer science class, where the instructor gives the students a task to write code in C++ that should execute the computation of a given mathematical function f. Often there is more than one way to complete the task so the example might fit better into explanation (3). What does the flowchart explain? It explains one way of computing mathematical function f. Usually there are adequacy conditions that derive from the social practice of computing - such as a certain simplicity or elegance in the design of the flowchart/program that act as further constraints. Conceivably, these might be strict enough to push the explanation into (2) – where there is a single 'right answer' to the programming exercise.

Adopting a goal contribution account of function avoids the problems associated with Cummins functions. The contributions of system processes are picked out because of their relevance to a system goal. Moreover, not every system possesses these functions because not every system is goal-directed.

Mark Bedau (1991) worries that natural equilibrium-tending systems like a piece of space rock attracted to the earth's gravity will also end up possessing functional processes by virtue of the 'goal' of the rock being to crash into the earth (see also Moreno and Mossio, 2015, Chapter 3). Therefore, the function of its solid parts is to provide mass to contribute to the goal. This objection has some force as a criticism of a general theory of normative function. However, it is not the aim here to commit to such a theory but rather to have a notion of function to appeal to in understanding coordination functions. The systems we are interested in face a distal goal of self-maintenance and so on which is not the case with the space rock. But perhaps more telling is that the systems of interest in this thesis are characterised by what happens when there is a *change* in task. The space rock 'system' is not sensitive to most changes of task, such the goal being changed to being repelled by the earth. However, it is sensitive to changes of task within a narrow range, for example if the earth were moved, the motion of the rock might respond appropriately. This much can be conceded without thereby losing sight of the fact that the task space of such a system is vanishingly small in comparison with something as apparently simple as the *Sphex* wasp. So, I might bite this particular bullet (see 6.7).

5.2.2 An example of a functional model

Design situations like the computer programming class, generate functional explanations that have normative force because they contribute towards a goal that is imposed from outside - in this case in the social context of a computer science class. The program has a certain functional organisation *because* it *must* compute a certain output for a given input - as specified by the teacher. The functional model shows, in an abstract way like a flowchart, how the different processes must interact to produce the required behaviour. This model is then realised in some concrete manner – in this example by the writing and running of computer code.

There are other examples that can help us see how prescriptive function contributes to explanations in cognitive science or this one due to Mazviita Chirimuuta in computational neuroscience (2018b, p. 865). Before I introduce the task addressed by the example let us look at a parallel example from engineering that can help set it up. Consider a problem confronting the engineer who wants to extract frequency and timing information from a signal (the value of a variable over time). The classic example is processing of sound. There are two typical tasks: identifying sound events (such as the onset of a musical tone) and extracting frequency information (such as identifying the pitch of a musical tone). There is a well-known trade-off between these tasks, known as the *Heisenberg-Weyl Uncertainty Principle*[70]. Because there is no such thing as frequency at a point, extracting frequency information requires a duration; the longer the duration of the 'window' over which frequency information is gathered the more precise that information. Making the window larger, however, makes the timing information about the event less precise. This is not to say that there are not better or worse methods for extracting this data. In the 1940's Dennis Gabor showed that there was an optimum solution to this trade-off got by computing a Laplacian (wave operator) to extract frequency information on top of a Gaussian envelope to extract timing information (Gabor, 1946)[71].

Computational neuroscience meets a two-dimensional analogue to this problem in primary visual processing (area V1 in the mammalian brain). Instead of frequency of sound waves over time, we are presented with an image extended over space and asking what are the spatial frequency components of the image?

[70] This arises from exactly the same mathematics as the Heisenberg Uncertainty Principle in particle physics.

[71] The same principle is used by modern frequency analysers, for example in sound processing applications, except they tend to use triangular or trapezoidal windows to reduce the computational load.

For example, is the image stripy and what is the orientation of the stripes (see figures 5.1 and 5.2)? This may be important in the identification of predators, but it could also aid other interpretive tasks such information about depth and the existence of edges.

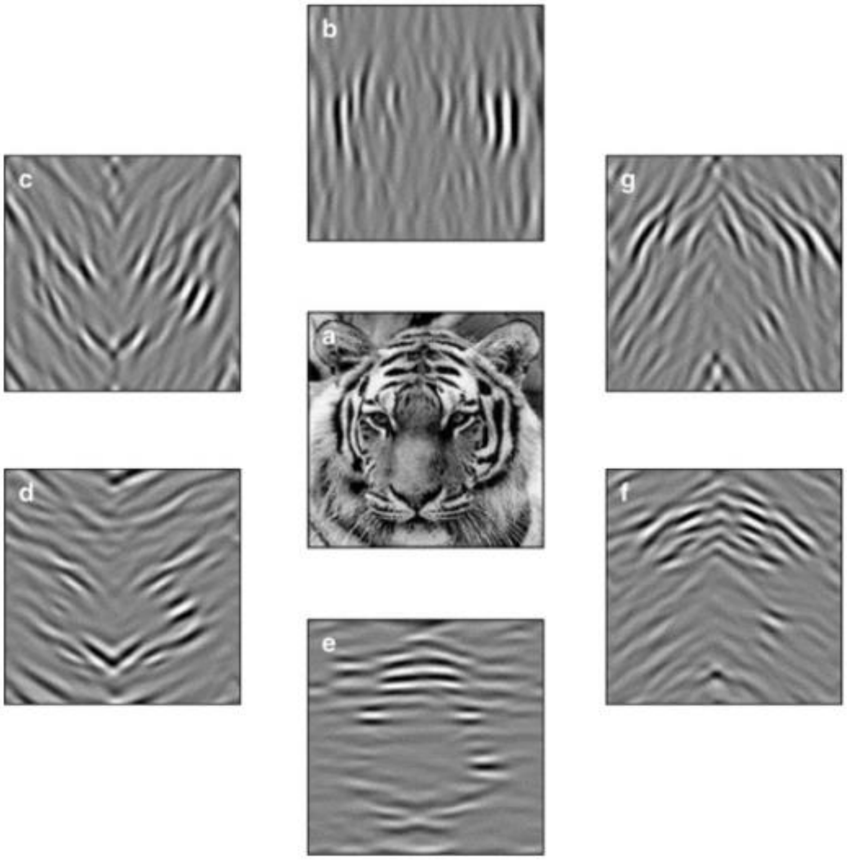

Fig. 5.1 The spatial frequency components of an image depending on orientation obtained by the application of an oriented Gabor filter. Image a decomposed according to vertical components b, horizontal components e and other orientations c,d,f,g (image courtesy of M. Chirimuuta).

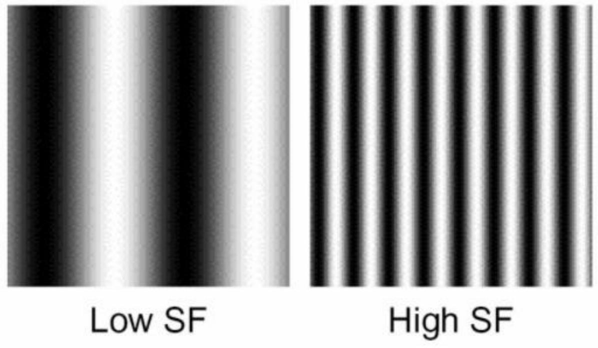

Fig. 5.2 Low and High spatial frequency components (image courtesy of M.Chirimuuta).

An optimum solution to the problem of extracting position and frequency information from the two-dimensional array of signals from the retina computes a two-dimensional Gabor function. Is there any evidence that this function is being computed in primary visual processing in the brains of mammals? It turns out that computational models based on the Gabor function fit well with empirical data about classical receptive fields - that is the set of visual stimuli to which the relevant neurons respond - in mammalian primary visual processing (Chirimuuta, 2018a, p. 866; Hyvarinen and Hoyer, 2001). There is evidence that this function is what is being 'computed' by V1.

What is going on here? Plausibly, what is at stake in visual processing is the distal goal of survival and inclusive fitness as well as efficiency; there are metabolic costs involved: what is needed is a system that is accurate, reliable, and cheap. Under these conditions the optimum mathematical solution to the problem can be seen as a useful benchmark against which to compare the solution that evolution has come up with – mathematics puts forward a prescriptive model - how the system *ought* to go about performing a task functionally in an ideal world. This is a bit like the flowchart in the computer programming example – providing a high-level functional description of the actual physical operations that the computer should perform. The model is ideal in that it abstracts away from the actual implementation details of the physical system.

This kind of functional model fits into type (2) or perhaps more likely type (3) above. Thinking of primary visual processing as a system in its own right with its own goal - say extracting accurate timing and frequency information from the visual array, insofar as this is possible given Heisenberg-Weyl uncertainty, Gabor gives a mathematical reason why a certain set of functions should be computed for reasons of efficiency. Neuroscience tells us that the basic processes in the brain are neuronal. The functional analysis prompts a conjecture that these neural processes should be computing the Gabor function. Experimenters then examine whether there is an interpretation of these neuronal processes, consistent with other general knowledge about neural processes, in which they can be said to be computing such a function. According to Chirimuuta there is such an interpretation. Therefore, the presence of these processes in V1 is *explained* by the functional model.

This is truly an explanation in a Woodwardian sense in that it can answer w-questions: *what* if things had been different? (Woodward, 2003, p. 221). It can explain *why* certain processes with the required properties appear at this stage in visual processing (if they did not, the required function could not be computed). Chirimuuta claims that *efficient coding* arguments like this are good explanations that are not mechanistic explanations. They are not mechanistic explanations because they are not ultimately concerned with describing the causal details of the realisers. Instead they look at the functional organisation and thereby make predictions about the kinds of realisation processes they will find. It is a top-down approach that imposes an explanatory framework on the causal nexus.

5.2.3 Functional analysis

In general, a functional model considers a function that is to be performed by the whole system and explores ways in which it can be broken down into a co-ordinated array of simpler functions; think about a complex task being broken down into a set of performances by Task Specific Devices. This decomposition is called *functional analysis* by Cummins in his work on the nature of psychological explanation (1975, 1983). Functional analysis explains how the system-level function (this will be the task faced by the system) can be performed by decomposing it into (simpler) sub-functions combined according a ‘program’ – that is organised in a way that can be specified by a flowchart. The explanation is ultimately grounded in the causality of the system – functions in the analysis are implemented by physical processes related causally. For example, the cells in V1 whose receptive fields are rectangular together perform the Gaussian window function in the Gabor model because of their causal properties. These are neurons that “respond to short bar-like stimuli of a specific width and orientation” (Chirimuuta, 2018b, p. 865). Processes involving these cells depend on visual events within the vertical rectangle (say) and not others. A given process performs this function because of its causal structure and its relation to the other functions in the system.

There are two points connecting functional analysis to the CSA developed here. The first is that the ‘program’ that combines the operation of the different functions in a time critical way can be thought of as a function itself – the coordination function. What Cummins writes about the program applies *mutatis mutandis* to the coordination function. The second point is that to demarcate a system we must be able to identify the processes that realise the model functions in the system. This is done by finding processes that have the right causal profile for the required function like those involved in the Gabor function in the example above. The ability of a functional analysis to pick out the implementing processes is what Martin Roth and Robert Cummins call a *causal relevance filter* (Roth and Cummins, 2014, 2017).

According to Cummins, at least in his earlier works, one of the strengths of this type of explanation is that the functional model does not depend on the exact details of the implementing processes. In the jargon of the field, functions can be *multiply realised*. Cummins refers to the causal processes as ‘componential analysis' and writes: "form-function correlation is certainly absent in many cases … and it is therefore important to keep functional analysis and componential analysis conceptually distinct (…) Functional analysis puts very indirect constraints on componential analysis" (1983, p. 29). This means that a given functional analysis can be realised in different processes, and functional analysis can serve as a tool for unifying disparate phenomena. I shall say more about this in a moment.

5.2.4 Explanatory power of functional models

In this section I examine the explanatory power of such a functional model. Much of this argument already exists in the literature so what I am doing here is adapting it to my own ends in terms of arguing for the adequacy of a functional explanation of systems capable of producing goal-directed behaviour. The explanatory power of such a functional model derives from the degree that lower-order functions are simpler than higher-order functions. In some cases, a complex system can be broken down into simple pieces, much in the way TSDs can be put together by coordination into a more complex result. Functional analysis allows us to understand how a complex behaviour can be the result of the coordinated ensemble of much simpler processes which is at the heart of the CSA. Explanatory complexity is transferred by functional analysis from individual processes to their coordination.

A complex system function can be expressed in terms of simple functions if the coordinating program is itself sophisticated (Cummins, 1983, p. 30). Interestingly Cummins also asserts that:

> As the program absorbs more and more of the explanatory burden, the physical facts underlying the analysing capacities become less and less special to the analysed system. This is why it is plausible to suppose that the capacity of a person and of a machine to solve a certain problem might have substantially the same explanation, although it is not plausible to

> suppose that the capacities of a synthesiser and a bell to make similar sounds have substantially similar explanations. (Cummins, 1983, p. 30)[72].

Of course, there are systems in which such a clear-cut breakdown is not possible. These are systems in which higher level functions emerge from the coordination of large numbers of simpler functions. The traditional linear functional analytic methods break down in these cases. Nonetheless, even if coordination itself is an emergent function of a system, it is plausible that its realising processes can be identified. Indeed, such a lack of functional modularity benefits the demarcation project because in these cases the coordination function is implemented *by the whole system*.

Either way there is a tension with mechanistic explanation in which a 'more details are better' approach is often taken (see Craver and Kaplan, 2020; Kaplan, 2011)[73]. The CSA only requires enough detail to secure the identification of coordination function. Since these functions are relatively basic to the system, there is no requirement that the explanation details the complete implementation of all system functions.

There is another source of tension between the functionalism of the CSA and the new mechanisms literature which has its roots in the differences in the metaphysics of sub-mechanisms and subprocesses discussed in Chapter 3. Sub-mechanisms are mereological parts of mechanisms. This relation is transitive. If

[72] This surely depends on the synthesiser. In the case of a physical modelling synthesizers like Chromophone (*Chromophone 2 physical object modelling synthesizer VST plugin*, 2014) there is a functional isomorphism between the model of the bell in the synth and an actual physical bell. The explanation for the similarity of the sounds then derives from the accuracy of the physical model. But Cummins probably had in mind something like the Yamaha DX7 which was very popular at the time which uses FM synthesis for which his statement is true.

[73] This tension has not gone unremarked in the literature, despite the historical links between functional analysis and mechanistic explanation. Craver spends the first part of chapter 4 of his (2007a) discussing how functional analysis is actually a mechanism sketch that needs to be 'completed' by supplying the requisite mechanism.

mechanism C is part of mechanism B, and B is part of mechanism A then C is part of A. However, as we saw in Chapter 3, the General Process Theory of Johanna Seibt denies transitivity of process belonging: if process C is a part of process B and process B is a part of process A, it is not necessarily the case that C is part of A. Process part-whole relations are not transitive. Given the strong connection of processes with functions we might suspect that the same intransitivity could be possible in terms of functions. If true, this would imply the possibility of functions existing at a low level in Cummins' tree diagram of a system function that are not actually part of the system. Although, I do not develop this idea here, something like it is explored in the discussion of symbiotic cognition in Chapter 7.

An advantage of functional explanation rather than, say, a mechanistic one, is that it can achieve a level of generality that unifies a disparate set of causal mechanisms under a general functional description. For example, there are indeed many and varied systems that are functionally described as washing machines from the washboard and mangle of my grandmother, my mother's old twin tub, through the Hotpoint Keymatic we had in the 1970's to the 'Elektro Helios Flexi Dose' currently in my bathroom. Their diverse causal arrangements are unified by implementing the same broad high-level function: to wash clothes.

Functional explanation handles the functional promiscuity of physical processes well. By this I mean that the same physical process may play different roles in quite different functional contexts as mentioned earlier. It is this fact more than anything else that prevents our reading off the higher-level function from the causal organisation of physical processes. For example, in human cognition, Chirimuuta writes about canonical neural computations as being standard neural modules "that apply the same fundamental operations in a variety of contexts" (Carandini and Heeger, 2012; quoted in Chirimuuta, 2014, p. 127). CNC's perform different functions depending on the context.

Functional explanation is adopted in this thesis as a way of describing the features of cognitive systems at a high enough level of abstraction to be able to capture the surface properties required to produce intelligent behaviour. At this level it detects patterns not visible at the causal process level. As Bechtel puts it: “information about how the parts are organised goes beyond the account of the parts and their operations” (2007b, pp. 182–183). While not perhaps reaching the high bar on explanation set by new mechanists, functional explanation does the job required of it in this argument.

5.3 Ecological functionalism

One issue we will need to sort out is how to reconcile functionalism espoused here with the process ontology described in chapter 3.

5.3.1 Functions without containers

Traditional functionalism grew out of a machine-like view of the system as being a nested set of ‘virtual’ machines each with its own input and output channels. Indeed, this view seems to underly the mechanistic accounts we discussed in chapter 2 as well. A machine is an entity that is enclosed by a spatial envelope and its transactions with the environment take place through clearly defined input and output channels. I referred to this as the container metaphor.

Section 3.3 proposed that applying the container metaphor to cognitive systems is a mistake. It is a mistake because it assumes that there is a clear-cut system-environment boundary, and clear-cut input and output channels, which in many of the systems that we investigate is not the case. Adopting a process view allows that a system could be entangled with the environment in a way that made drawing boundaries and identifying input/output channels difficult.

So how are we to understand functionalism from the perspective of a process ontology? There are two remarks that are pertinent here. The first is made by

Mike Wheeler - the functionalism that we adopt will need to be able to transgress the boundary of the organism if it is to play a non-question-begging role in the extended mind debate. Wheeler calls this *extended functionalism* (2010, pp. 248–249, 2017). By taking this line Wheeler is already accepting that inputs and outputs to the relevant functions do not have to be the same as the organism's inputs and outputs. The second is to remind the reader of the point made in 3.3 that we need to think of the system not just as a passive receiver of an input from the world, manipulating that input, and delivering an output, but as something that actively transforms world states[74] and processes through being entangled with it. It is for this reason that I coin the term *ecological functionalism* for this brand of extended functionalism[75].

I propose the following notion of function. Function is a prescription of a transformation of the current state (or set of processes) of the world (the 'before' picture) to the final state (or set of processes) of the world (the 'after' picture) in some suitable abstract language, including constraints internal and external to the system. If the function in question is that of the whole system, then it is none other than the *system task* as discussed in the previous section. But this makes a lot of sense - the system is prescribed to perform the system task, and coordination is the key process in making sure that this happens.

5.3.2 Distinguishing system and behaviour

This way of viewing function also helps to clarify what we mean by behaviour of a system, viewed from a process ontological perspective, that is entangled with its environment. Ecological functionalism sees goal-directed behaviour as a set of coherent, persistent and adaptive processes in the world. It need not be

[74] In a process ontology I think of a state as being simply a relatively stable process.

[75] No link with ecological psychology is intended here.

thought of as behaviour of a ‘thing’. This is what Johanna Seibt refers to as ‘subjectless activities’. “Subjectless activities are temporally extended and, like things, they are good illustrations of the category feature of being an *enduring* entity, that is, an entity that persists through time (…)” (Seibt, 2018, p. 116). Building of a house is a subjectless activity in the sense that there might be no enduring entity whose behaviour it is but rather a feature of a set of processes themselves with temporal profiles and without clear-cut boundaries. I can imagine that the building system consists in a variable collection of individuals whose collective interactions with their environment constitute building behaviour; such behaviour takes place without being the behaviour *of* a fixed-object with clear-cut boundaries. There is a distinction to be made here between the behaviour of individual elements of a system and behaviour of the system itself. We can of course speak of individual builders being engaged in building behaviour. But the building of the house is a behaviour despite not being specifically behaviours *of* someone or something[76].

Some readers may accept this but then worry that we lack criteria for distinguishing behaviour from system. I would agree that the CSA brings these two sets of processes into close proximity. There are a number of possible responses. One can point to the causal relations between system and behaviour. System tends to cause behaviour rather than *vice versa*. Moreover, system will be subject to functional organisation constraints such as the coordination conditions while behaviour perhaps less so. Of course, it is an advantage of the CSA that, in the right context, behaviour is incorporated into system. The behaviour of an individual ant is part of the system of the colony. Perhaps the dividing line is drawn again by explanatory interests. If we are investigating how the colony can intelligently decide between two food sources then the behaviour

[76] The propensity to want to reify the subject of an activity – to speak of a behaving entity – may be a remnant of mechanistic or container-like thinking.

of the individual ant is a system process that is part of the whole colony, while if we are interested in how individual ants respond to pheromones then talk of ant behaviour makes sense.

5.3.3 Circularity worries

If I am right that ecological functionalism is the correct form of functionalism to understand the functional organisation of a system, in particular the coordination function, then we are free to jettison the idea that a system must possess a long-term permanence or that it can be expressed in terms of a permanent architecture possessing inputs and outputs, *pace* Rob Rupert[77].

But now we encounter a potential circularity problem. In the CSA, the system faces tasks that derive from distal system goals. Even if we accept, following teleological emergentists such as Walsh and Bickhard, that tasks emerge from the working of the system, it does seem to be the case that the boundaries of the system are set prior to talk of it having goals or facing tasks. Given that having a goal is a property of the whole system, shouldn't the demarcation question be settled before we start to talk about system goals?

First, let me be clear that teleological emergentism is on the right track. I *do* think that having goals and facing tasks are system-level features. And at the same time, I also think that the boundaries of the system shift and change and are task-dependent in a critical manner. I do not think that this is a vicious circle, however.

[77] Some readers might detect a tension between my fondness for cybernetics and general system theory and the worries I have about the container metaphor. Ashby is explicit about his view of systems (including biological systems) as machine-like (see Ashby, 1960; Pickering, 2010). Modern conceptions of systems in terms of networks and complex systems theory does not, in my view, commit to the sort of machines as containers that Nicholson rightly rejects (see, for example, Bickhard, 2011; Ladyman and Wiesner, 2020; Thurner *et al.*, 2018). I develop this view in part III in the thesis.

There is a distinction here that may help. The boundaries that are relevant to a system having goals may be one kind of boundary. The boundary that is relevant when discussing cognition may be a different kind of boundary. Autopoietic enactivism of a classical kind (see Maturana and Varela, 1980; Varela *et al.*, 1993) asserts that these boundaries are the same because it identifies cognition essentially with self-maintenance, but I don't think this is right. According to the CSA the sphex wasp has goals by virtue of its being an open system in a far from thermodynamic equilibrium state. But the cognitive system may also encompass the grasshopper insofar as this is a physical trace of its previous actions and its position is coordinating of further actions. There are two kinds of emergence at work here: that of normativity and cognition. The proposal is the emergence base for these features is not necessarily the same – i.e. that the systems from which they emerge might not have the same boundaries. There may be cases in which they coincide, but this is not a conceptual necessity. That they are the same may be another presupposition that potentially skews the debate about HEC.

5.3.4 Using the coordination conditions to demarcate cognitive systems

There are two ways in which the coordination conditions can do work in an investigation. We can investigate a given system, and on discovering that the coordination conditions are satisfied with respect to a set of tasks, conclude that it is goal-directed in relation to these tasks. But we can also work the other way around. Given that there is a patch of the world that, on independent grounds, we have reason to believe is goal-directed with respect to a set of tasks, we can use the coordination conditions to investigate what processes play a coordination role with respect to these tasks and thereby set the limits of the system responsible. It is this second strategy that is exploited in this thesis to cast light on the demarcation problem.

Either way, the existence of coordination is proposed as an explanation of goal-directedness in a system. In the first case coordination is evidence that the system is goal-directed. In the second case, goal-directedness is established by other means, so coordination is taken as evidence of the extent of the system - so explains systemhood. The first case is illustrated by the neutrophil. A set of processes implementing the coordination function found in the neutrophil would be evidence for its goal-directedness. As I understand it, empirical investigation into the mechanisms of neutrophil motility is in its infancy (see Liu *et al.*, 2012). But I would expect that empirical investigation will reveal the appropriate processes for coordination eventually. The second case is illustrated by a slew of examples from the 4E literature. The question with the Otto-notebook system is not whether it is goal-directed, but rather how far does it extend?

In adopting the second strategy, the investigator does not start from zero but from a provisional idea of the extent of the system on independent grounds. For example, given the previous discussion, a starting point may well be the system as a bearer of goals. Otto is himself a bearer of goals that we have independent reasons to know. The investigator therefore considers a region of the world that includes this system as a candidate, for example processes involving Otto and the notebook. Then application of the coordination conditions will single out the processes involved in coordination. The only caveat to this process is that the investigator should initially err on the side of making the system too large. Coordination, being an emergent feature, can fail to appear if the circle is too tightly drawn.

For another example, here let us return to mammalian vision and the efficient coding argument of Chirimuuta. The goal of survival and inclusive fitness, which boils down to detecting predators and the like, belongs to the whole organism. But from previous research we can narrow down a first guess for the location of the system responsible for performing the function of detecting the timing and frequency of visual events to the primary visual area V1. The functional model prescribes a function that would be expected to play a part in the process, the Gabor function, which is broken down, Cummins-wise, into the Gaussian window and the Laplacian operator. The search, then, is on for the appropriate neural processes at the implementational level which turn out to involve neurons with the appropriate classical receptive fields. Finding neurons with the appropriate properties to implement the functions helps refine ideas about which processes are involved in the system responsible for coordination of these tasks[78].

5.4 The explanatory power of the task-coordinator relation

But what work is done by the coordination conditions in these arguments? The tracking and triggering condition ensures that the behaviour produced by the system is coherent – that it strings together TSDs to produce continuous rather than erratic behaviour. It does not on its own account for goal-directedness. It could be argued that dissipative systems such as candle flames or Bernard cells satisfy the first condition, because they possess a kind of self-maintenance that ensures smooth and continuous behaviour that can resist small environmental perturbations, but their ability of adapt to big changes of task is minimal. Like the popular view of solar system, they stably do what they do[79]. It is the second condition that really does the lion's share of the work in establishing goal-directedness by ensuring that the system is suitably sensitive to the tasks that it faces. This seems to be more than just a functional condition. What reasons do

[78] Larry Shapiro raises the issue that functionalism alone cannot provide criteria for demarcating cognitive systems (2008). Without going into these arguments here it seems that they apply more to mechanisms than processes where it is understood that boundaries are in any case less well defined. Chirimuuta's example seems to contradict Shapiro's argument. In any case, internalist functionalist positions are just as vulnerable to this objection as externalists.

[79] Laplace showed that the solar system was not actually a stable dynamic configuration in the long run (see Roy, 1973).

we have for supposing that a relation such as this is explanatory (and therefore the processes identified by it should belong to a minimal set of explanatory resources)?

To answer this question, I shall use an analogy with the parallel problem of causation. I want to emphasise up-front that it is only an analogy - I am not committing to a particular theory of causation. The tactic is just to show that if the causal relation is explanatory by virtue of its formal properties, then the coordination relation is also explanatory *because it possesses the same formal properties*. One view of a causal explanation is that it posits the right sort of counterfactual relation between cause and effect. Let us say for the sake of the argument, that an event (or value or process) C causes event (or value or process) E if it is the case that when C is reliably present then E is too, and if C is absent then E is also absent. Denis Walsh calls this a counterfactual invariance relation (Walsh, 2013, p. 51). A causal explanation appeals to this counterfactual relation: E is present *because* C is present. As Denis Walsh puts it: “The cause of a phenomenon is the set of conditions that makes the difference between its occurrence and its non-occurrence” (2013, p. 51). In the argot of this literature, the cause is a ‘difference maker’ for the effect. Suppose ᴛ is a task space for goal-directed system S with coordinator C. If the task changes from T_1 to T_2 then the state of the coordinator (in functional terms) changes from C_1 to C_2 in a reliable manner. The task is a difference-maker for the coordinator state. Thus, the change in task reliably produces a change in coordinator state which explains task-sensitivity required for goal-directedness. It is the counterfactual relation between task space and coordination process that does the explanatory work in an analogous manner to the causal case. Why did the neutrophil change direction? Because the bacterium did, and the direction of travel of the bacterium is a task setter for the neutrophil. In other words, the system changed behaviour because the task changed, and the system is goal-directed. Alternatively, in relation to the first kind of explanatory scenario: is the system goal-directed with respect to direction of travel of the bacterium? Yes, because we can point to the coordination process[80].

Satisfaction of the coordination conditions as a requisite for goal-directedness with respect to a set of tasks does not mean that the system is necessarily successful in carrying them out. The neutrophil is goal-directed even if it never catches the bacterium. The washing machine is goal-directed with respect to washing tasks even if it is never used, providing it has the right control architecture. A malfunctioning system whose control system is intact but nonetheless signally fails in its tasks is still goal-directed.

Of course, there will be pathological cases where this argument fails. Suppose the damage to the system is so severe that it destroys the appropriate relation between the task and the coordinator state. Although these cases fall in a grey area there will be a point at which the pathology is so great as to undermine goal-directedness itself. However, it may be hard to draw the line between a broken goal-directed system and a non-goal directed system. In these situations, one could be guided by other considerations such as the organisation of functioning systems of the same kind.

[80] Some theories of causation would describe the relation between the task the coordination process as being causal (Woodward, 2003). I remain agnostic. Woodwardian accounts are not fussy about the relata in causal relations referring to them as ‘variables’. I wonder whether the variables in the task-coordinator relation are of the right kind to be causal. The first is an abstract function (a transformation of world states itself a relation between goal and constraints) and the second is a set of processes implementing it. I prefer to say that it is a determining relation of some kind.

5.5 The coordination argument to HEC and explanatory encapsulation

Let us take stock where we are in establishing the coordination argument to the HEC. Recall the argument from 3.7.

(1) Cognitive systems produce goal-directed behaviour.

(2) Coordination is responsible for goal-directness.

(3) In order to understand the goal-directedness of the system the explanation must include how the system processes implement coordination functions.

(4) Therefore, if the system is thought of as being a core explanatory unit, then the process realising the coordination function must be part of the system.

(5) Therefore, if an ‘external’ process plays a coordinating function then it is part of the system.

(6) Therefore, if a situation can be found in which this is the case, then HEC is true.

Premise (1) has been taken as read for the sake of the argument for the whole thesis. If I am on the right track so far, chapters 3 – 5 have established (2) and (3). It remains to establish premise (4).

I can imagine a reader, perhaps someone like Rob Rupert, who goes along with the gist of the argument but does not buy premise (4): the move from coordination to (cognitive) systemhood. Maybe Rupert would list coordination as a function that could be performed by resources external to the system. What is the argument for including these in the system?

Recall that in earlier discussion of systems, I proposed that boundaries of systems depend on our explanatory interests. The braking system of a car consisting of processes involving the pedal, hydraulics, and brake pad assembly is picked out by our explanatory interest in explaining how braking occurs. This is the line taken by the CSA that systems are task relative, and the set of tasks chosen for investigation is in the hands of the investigator. Attempts to shake off the observer relativity of systems seems to end in circularity since one needs criteria for selecting the relevant task set which requires a prior notion of cognition, for example through taking some tasks as representative of cognitive capacities (see Section 4.5)[81].

What happens if we omit a component, say the processes involving the hydraulic fluid? The system is now two separate processes the pedal movements and the brake pad movements. To explain the braking action we would need to invoke an output from the pedal action that mysteriously correlates with an input of the brake pad action. There are two problems here. The first is that this explanation is intuitively more complicated than an explanation involving the hydraulic processes. The second problem is that the correlation seems like magic. The first problem is similar to that raised by Clark, discussed in Chapter 1 (Section 1.4.3). The second I shall dub the *magic coordination* problem and is discussed below.

In Chapter 1 I introduced Sprevak’s argument that it does not seem to matter from the point of view of simplicity whether ‘external’ components are considered part of the system or not. I argue that it does matter. If the system is a set of minimal resources required to support the explanation, an assumption that we

[81] It might be useful to distinguish between weak and strong observer dependence. Weak observer dependence would be a requirement that there are many token instances of a kind K, and it is the observer that provides criteria for picking one of them out – as in the case here. Strong observer dependence would be akin to something like attributivism where properties of a token depend on the attribution of an observer (see Dennett, 1987).

make in this thesis, then the hydraulic process, because it is needed in the explanation is part of the system.

But there is a second argument to be made here concerning problem 2. The CSA argues that coordination is responsible for goal-directed behaviour *qua* goal-directedness. Were coordination processes omitted from the system S it would omit the very items of the explanation responsible for the phenomenon to be explained. Inputs to S would be correlated through some prior process outside the system and these inputs would seem *magical*, in the same way as inputs to the brake pad assembly when the hydraulics are not included.

I can imagine some readers objecting that I have *stipulated* that systems, as sets of processes, are explanatorily encapsulated – minimal sets of resources involved in an explanation of a phenomenon. But my point throughout this part of the thesis has been that because of the multiplicity of boundaries that can be drawn around systems, there must be some investigation-relative fact that picks out one boundary rather than others. A second point that I have highlighted in the thesis is that the investigation relativity of systems does not in any way undermine their reality. The braking system of the car really does exist. The third point concerns the fact that in some systems functional properties such as coordination are emergent through the interaction of system elements. One needs the whole to explain how the dynamics of coordination comes about through complex non-linear interactions (see 3.3.2). Removing items in this interaction from the system, again, makes the emergence of these functions magical. Systems then are natural explanatory units that really do exist in the world.

These arguments establish premise (4). A system should be sufficient to support a functional explanation of goal-directed behaviour, without having to appeal (magically) to an external source of goal-directedness. But coordination is the source of the system's goal-directedness. Therefore, coordination processes lie within the system.

5.6 Concluding remarks

This chapter has dealt with the issue of drawing boundaries around systems using functional criteria. There are many boundaries and therefore (nested) systems in the world. Explanatory requirements pick out the boundary of interest. This chapter investigated how a functional explanation picks out the boundaries of the system.

First, functional explanation was explained and then defended against some arguments from the new mechanisms literature. Traditional functionalism was replaced by ecological functionalism that can operate in the CSA context, that is, collections of processes entangled with the world.

We discussed how behaviour processes could be distinguished from system processes and how we might avoid circularity worries concerning system goals. Then we showed how the coordination conditions could be used ‘in the field’ to demarcate cognitive systems.

The second coordination condition asserts a relation between task and function. As such, then, it goes somewhat further than a functional explanation in which processes are explained through the functions that they perform. I showed that this condition is explanatory in the same way that causal explanation is explanatory since they both have the same formal structure.

Therefore, the coordination conditions are explanatory and therefore pick out the system considered to be an explanatory unit.

Finally, we can return to the five areas of contention discussed at the end of chapter 1. How does the CSA answer them? Here is a brief summary.

(1) What is the appropriate style of explanation in cognitive science?

The CSA takes the view that (ecological) functional explanation identifies the appropriate kinds which are realised by token processes (see 5.2)

(2) Should cognitive systems be conceived of as agent-centred or distributed?

The CSA is able to accommodate both but does not need the assumption of a locus of central control (see 3.2, 3.5.3, and 8.2.2 to come)

(3) What theory fixes the reference of cognitive concepts?

The CSA takes the view that cognitive concepts are fixed empirically by the best cognitive science. The CSA is largely agnostic regarding cognitive categories, the only requirement being that cognition is responsible for goal-directed behaviour as described in this part of the thesis. (see the main discussions in chapters 3, 4, and 5)

(4) What role does functionalism play in the theory?

Ecological functionalism is central to the CSA. It identifies coordination processes as high-level functional kinds. (see 5.3).

(5) What is the role of representation in the theory?

The theory does not rely on cognition being essentially representation-involving although it turns out that a minimal kind of representation may be the outcome of the theory (see 8.4 to come).

Introduction to part III

> When we turn to the coordination events and see all the media that are simultaneously in coordination (some inside the actor, some outside), we get a different sense of the units in the system. (Hutchins, 1995a, p. 158).

There were four broad themes running through part II: the nature of systems, the characterisation of goal-directedness, a functional and causal characterisation of coordination, and the nature of functions and functional explanation. I shall summarise each briefly before turning to the business of the third section of the thesis.

Systems are epistemic units that provide the minimal resources for adequate explanation. A major piece of the jigsaw is the argument that when the explanandum is behaviour, oriented towards a set of tasks, coordination processes with respect to these tasks must be part of the system that supports the explanation. Would-be systems omitting the coordination processes are explanatorily inadequate. A ‘no magic’ argument is given for bridging the gap between explanatory adequacy and systemhood.

Such an epistemological account raises questions regarding the ontological status of systems. Part II dealt with the task-relativity of coordination and therefore of systems. Coordinator talk, and therefore system talk, is always relative to the target of an investigation. But this does not make systems and their coordination processes less real. This position is compatible with a pluralistic realism about coordinators and their systems. There are many real systems out there individuated by their task spaces; this is, as we have observed, promiscuous systemhood.

Goal-directed behaviour is observably coherent, persistent, and adaptive. It is a coherent molar whole that is stable under changes of environmental conditions.

When environmental conditions change in the right sort of way behaviour can change somewhat more radically – the system selects a different element from its behavioural repertoire – hence it is adaptive.

The character of goal-directedness places demands on the system responsible for this behaviour. The coordination conditions express these demands in functional terms. The first condition ensures that the individual components of behaviour, produced by task-specific devices, are strung together in a time critical manner. The second condition, the sensitivity of the coordination process to changes in task, is required for adaptiveness. Taken together the system is robust with respect to the relevant set of tasks.

The final theme dealt with in part II concerns normativity and function. Function is normative and is conceived as being the contribution of a process to a 'distal' goal of the system. Because the CSA is a processual theory functions are defined in an ecological manner – a system performs a function through the transformation of the world from one state or process to another, rather than thinking of functions in terms of transformations of inputs to outputs of an isolated, machine-like, thing.

How can the coordination conditions help in the task of demarcating cognitive systems? First, coordination processes can be identified through their causal role. This means looking for causal connections that are characteristic of the functions that they are to perform. For example, in the car engine, coordination for combustion timing can be traced backwards along the causal chain from spark plugs to processes in the distributor which are responsible for both tracking and triggering. In this case, task variation occurs along a single dimension that is the required frequency of the combustion stroke, which is linked to the distributor through the causal link with the camshaft. Identifying a part of the world as a co-ordinator for the performance of a set of tasks also identifies it as belonging to the system as a minimal set of resources required by the explanation. Hence the co-ordinator approach has the potential to provide a sufficient (but not a necessary) condition for a component to be part of an extended cognitive system.

The third part of this thesis uses the CSA to explore a selection of examples and to move the debate on to slightly different ground. Some of these are familiar from the first two parts, while others are new. The aim of this is twofold: firstly, to ascertain to what extent the method allows progress to be made on the demarcation issue, and secondly to provide an overview of some of the problems that are encountered in applying the method in practice.

Applying the CSA to the examples that underly a parity driven approach to extending cognitive systems through the use of material resources in the environment, is the main aim of chapter 6. Here we apply the machinery of part II to issues that were raised in chapter 1 such as the problem of cognitive bloat and the coupling-constitution question. Does the CSA provide support for the HEC? Certainly, the examples examined in this chapter seem to point to this possibility.

Perhaps the greatest support of a version of the HEC is to be found in exploiting the CSA in relation to socially extended systems which is the main concern of Chapter 7. The second half of chapter 7 introduces an important class of systems, those involving so-called *stygmergic* coordination processes, which play an important part in the rest of the thesis.

Finally, chapter 8 takes stock theoretically. How does the CSA fit with existing approaches in the field? It is perhaps not surprising that the CSA sits well with second and third wave approaches and distributed cognition because of its non-agent-centred diachronic stance. The role of representations in the CSA is discussed and it turns out that a minimal kind of representation might be involved in coordination despite not being put there at the beginning. The striking link with enactivist approaches noticed in chapter 4 is held under the microscope. While there are certainly strong connections here, the suggestion I want to pursue is that the CSA, by virtue of its normative assumptions, is distinct from enactivist approaches of a radical bent. The chapter ends by sketching answers to some difficult questions raised in chapter 2.

Chapter 6

The CSA and materially extended systems

6.1 Introduction

This chapter and the next investigate whether the CSA approach developed in section II of the thesis can be used to support the HEC. They will also assess whether a coordinated systems approach can respond to the questions and problems raised in chapter 1.

This chapter focusses on aspects of the original parity-driven defence of extended cognition. The hope is that the CSA will overcome some of the difficulties of this approach. It was developed along different lines to the parity-principle with the express intention of avoiding the grain size problem and any vestige of human or neural chauvinism, without thereby producing cognitive bloat. The hope is that any system that fits the functional bill can be considered cognitive, whatever its origins, location, or material composition. The aim therefore is to check whether the CSA fulfils these ambitions.

This chapter focusses on examples in the literature of material extension as a way of showing how the CSA might work in practice. The original Otto example is discussed along with the Tetris case from CC. The former turns out not to be quite so straightforward in a CSA context and is quite nuanced and contains many layers of subtlety.

Finally, this chapter examines whether the HEC casts any light on the hard problem of sterile effects emerging from chapter 2.

6.2 Otto's notebook

Let us revisit Clark and Chalmers' canonical example (see chapter 1) to see whether the CSA agrees that it is an example of an extended cognitive system. Recall that Otto and Inga have decided to visit the Museum of Modern Art in New York's 53rd street. Inga uses her biological memory to locate the museum, but Otto, who has mild Alzheimer's, relies on a notebook to find the address. CC assert that the notebook plays the same functional role for Otto as Inga's biological memory plays in her trip to MOMA, therefore by the Parity Principle it is part of Otto's cognitive system (1998).

The difficulties with parity, especially those to do with what characteristics are relevant and how to decide what degree of similarity is acceptable, motivated us to develop the CSA in the first place. This means that we were less interested in whether the notebook extended Otto's cognitive system by playing the same functional role as some folk-psychological capacity like memory or belief, and more interested in how processes involving the notebook contributed functionally to the coordination of *the whole system*. In this case the whole system might be conceived as a coalition of three different groups of processes: Otto's internal neural processes, the interactions with the notebook, plus possibly some other kinds of interaction with the environment.

There is a sense, as I have indicated, that the CSA approach taken here combines three insights. The first is AA's requirement for a mark of the cognitive – they cash this out in terms of a non-derived content condition – while the CSA proposes the weaker coordination conditions. The second is Rupert's insistence that we take the concept of system seriously enough for it to do work in the argument, the CSA suggests that a system is a self-sufficient explanatory unit and in some cases the whole from which coordination functions emerge. Finally, Wheeler's insight that questions about cognition should be settled by reference to a theory of cognition rather than on a functional comparison with the human case, the CSA proposes that this theory is a set of minimal (ecological) functional conditions for robustness. Therefore, the argument as stated in CC is not one that is endorsed by the CSA.

However, this is not to say that the CSA has nothing to say about this example, and it may be that it arrives at the same conclusion as CC but via a different route. But we must keep in mind that the basic question asked by the two approaches is different. For CC, the question is whether the notebook serves as an extension of Otto's cognition in the sense that it spits out a representation in an analogous manner to the spitting out of a representation by Inga's biological memory. For the CSA it is whether the notebook contributes to the coordination required by the goal-directed behaviour of the Otto-notebook system, namely the performance of a navigation task in midtown Manhattan. To address this question, we need to ask: is it task-sensitive regarding the tasks marking the relevant navigational capacity, and does it track and trigger the relevant navigational processes?

How we answer this question depends on exactly how Otto uses the notebook. We need to decide, for example, whether manipulation of the notebook corresponds in the right way to changes in the task in order that it satisfies the task-sensitivity condition. It is conceivable that, if the task is changed, then the manipulation changes too. Suppose that instead of turning to page 31 for MOMA, Otto turns to page 37 when faced with the task of navigating to the Birdland Jazz club. This would be *prima facie* evidence that manipulation of the notebook is task sensitive. But how do we make sense of the tracking and triggering condition? Does the notebook track and trigger the processes involved in performing the navigation task? Let us suppose that there is a map in the notebook and that Otto interacts with the map by placing his finger at his current position, then it seems plausible that the tracking and triggering condition is satisfied. The finger records the position on the map and triggers navigation actions such as changing direction at street intersections.

The tracking condition is more demanding than the triggering condition. A trigger can be as little as a conditional causal link - a switch turning on a time-critical action. Tracking either requires something like a trace of recently performed actions, or an ongoing causal process in which a physical state is a stand-in for the stage of the process being tracked. For example, in a washing machine the physical state of the pressure switch is a stand-in for the pressure of the air in the drum, which is itself a proxy for the amount of water in the drum compartment.

This tracking condition places constraints on the way Otto's notebook should be used if it is to participate in coordination. For example, there may be a set of robust practices that include writing to the notebook as well as reading from it, or as in the example above using some tracking indexical like a pointing finger on a map. In terms of general notebook usage, there may be constraints on the kind of thing that should be written, and how it should be read[82]. The use of the notebook will be subject to norms that allow it to perform a coordinative function.

In the case, then, where the notebook contains a map or some other feature that allows it to perform the tracking function, the CSA seems to allow notebook-based processes to be part of an extended system as CC assert. This is not because it plays a similar function to human memory, but rather because it is involved in coordinating goal-directed behaviour.

But in their original example, CC make no mention of a map. In its absence, the notebook does not seem to satisfy the coordination conditions. While it might be conceded that it performs a triggering function in that the address of MOMA is retrieved from it, it does not seem to track the processes involved with navigation to the museum. The trigger provided by the notebook is not conditional on tracking – it is the integration of the two that is required by the first coordination

[82] See Weiskopf (2008) on informational integration and Clark (2005) for a reply.

condition. Moreover, the notebook does not seem to be at the centre of what makes Otto's behaviour goal directed in the sense that the coordination dynamics responsible for navigation to MOMA do not involve the notebook in an ongoing fashion. This may violate even CC's own requirement that external components are *actively* involved in cognitive processes. I shall say more about this in a moment.

Summarising then, under these conditions the notebook is not sufficiently closely integrated into the coordination process to be part of it. The CSA cannot be used to argue for the inclusion of these processes in the system. Because the coordination conditions are sufficient but not necessary it does not show that the notebook is *not* part of the system – it just does not provide evidence that it does. In the situation where a process *does* receive tracking updates from Otto, for example on a GPS system, it is far more plausible that the GPS algorithm is part of Otto's coordination processes with respect to the navigation task.

The tight integration implied by tracking-conditional triggering is crucial to the CSA. The friend of extended cognition gets into trouble precisely in those cases where the coordination functions seem to come apart. Richard Menary warns us against the scenarios where we attribute cognitive status to the whole of the BBC website because it plays a role in informing our intelligent conversational behaviour, for example (Menary, 2013, p. 32). I assume that the BBC does not receive updates on the state of completion of Menary's conversations. The BBC website fails as a tracker of the processes involved in this task. It therefore fails the coordination conditions. More importantly, the example goes against the intuition on which the coordination condition is based, and for that matter CC's *active externalism*, that the core processes in the cognitive system are involved in *real-time* coordination dynamics. This is the reason that the tracking condition is so important. Again, keep in mind that failing the coordination conditions says nothing about cognitive status. It just means that the CSA does not supply positive reasons for including the BBC in the system responsible for Menary's goal-directed conversational behaviour.

How is it then that a central example in the debate such as Otto's notebook turns out to be not so clear-cut as CC would have us believe? The answer turns on the fact that CC are arguing with a very specific notion of cognition in mind. For Otto's notebook to work as an intuition pump for the HEC, requires the acceptance of three claims: that cognition is bound up with representations – what is written in the notebook is a representation, that the categories of folk psychology such as beliefs and desires are appropriate for analysing cognition – Otto and Inga harbour beliefs and desires, and that the parity principle holds – the representation in the notebook is constitutive of a belief because it plays the same functional role regarding the navigation task as that of Inga. The more general approach taken in this thesis does not tie us to acceptance of these claims. We do not assume that fully-fledged representation is involved in cognition. Secondly, the CSA does not require that cognition involves folk psychological categories. Systems such as the amoeba may be deemed cognitive even if they do not have beliefs and desires, or indeed first-person propositional attitudes at all[83]. Finally, we deem the parity principle to be too deeply problematic to be of use. The Martian intuition arguments of Sprevak undermining the parity argument are compelling – either the parity argument is faulty, or it leads to severe cognitive bloat (see Section 1.3.4). Moreover, there is a related question of which functional properties should be selected for judging parity which seems to require a prior

[83] Some may argue that we attribute beliefs and desires to the amoeba as a fiction allowing us to predict its actions (see Dennett, 1987; Hutto, 2008, 2013, 2021; Toon, 2016, 2021). The CSA is not couched in terms of beliefs and desires but rather in cybernetic terms of tracking, triggering and tasks and makes no claims about the links between the two sets of categories, so seems compatible with fictionalist views of folk-psychology, without affirming them. Indeed, as I speculate later, coordination offers an alternative to folk-psychological explanations of, for example, joint action.

theory of cognition to be in place and therefore puts the cart before the horse (see Section 1.3.5).

The CSA does not find that operations involving the notebook are necessarily part of the cognitive system. The arguments made in its favour are based on three premises, two of which the CSA rejects outright and a third that it does not accept.

6.3 Tetris

The game of Tetris discussed in chapter 1 provides another test case. I described how Kirsh and Maglio studied the behaviour of players of Tetris and discovered that they tended to rotate the pieces before fitting them even though there was a (time) cost to this in the game (Kirsh and Maglio, 1994). Their paper makes a distinction between two functions played by the interaction with the Tetris pieces. The first function is rotation to orient the piece. The second involves moving the piece and fitting it to the wall. Although rotation is normally required to fit the piece in the wall anyway players engaged in more rotation actions than was needed for the game. Kirsh and Maglio suggested that the extra rotations played a cognitive role that they labelled as ‘epistemic’ as opposed to ‘pragmatic’ for those actions that involved game moves. The CSA sees these actions again as distinguishing coordinative and plant processes. The Tetris rotations are described by Clark as actions “designed to extract or uncover information” (2011a, p. 71, see also 1997, p. 66). What information could it be that these rotations uncover? After all, in a technical sense, all the information required to determine where to fit the Zoid into the wall is available on the screen, rotating the object does not add anything that is not already there. One possibility is that it re-presents information already available in a more tractable form for other processes in the system. Better still is to sidestep an informational analysis altogether and talk in terms of the coordination dynamics of the system - something that is not obviously best expressed in informational terms. Instead of simulating the rotation internally, finding a fit, and then performing the successful simulated rotation on the screen, the idea is that rotating the piece can serve a dual role - not only as part of a search routine but also as part of the performance routine. This is because whatever epistemic actions have taken place beforehand, in the end, the zoid needs to be rotated to fit into the wall[84]. It is conceivable that, by rotating on the screen, the player simplifies the internal processing needed. The internal sequence would be something like ‘rotate’, ‘match piece to wall’, ‘translate piece on screen without rotating so that it fits the hole in the wall’ (remember that translation and rotation are controlled by different buttons in Tetris). Such a sequence would not require any in-the-head rotation which, it can be conjectured, is a relative difficult and perhaps slow cognitive process.

Instead of seeing this as ‘offloading’ cognitive tasks on the environment, the CSA prefers to frame it in terms of an integrated system in which the ‘extra’ rotations are part of the coordination process. To show this we need to check whether the processes involving the rotation of the zoids satisfies the coordination conditions. If this is the case then, the CSA story goes, we will be justified in treating the whole as part of the cognitive system. We would thus arrive at the same result as CC but for different reasons.

Let us begin with task-sensitivity, the second coordination condition, since it is more tractable in this example. Recall, that in chapter 4 I defined a task as being the transformation required to bring about a goal state or process from the current state or process. In the Tetris game a task corresponds to the relation between the current state of the zoid and an appropriate hole in the wall. The shape of the wall, incidentally, constitutes a set of environmental constraints that define the action possibilities of the player. Given a candidate gap in the wall, this task relation amounts to a rotation of the zoid around its centre (of multiples of 90 degrees) to give it the correct orientation, followed by a translation (a movement

[84] I describe this dual functional role below.

without rotation) to fit it into the gap. Of course, some gaps in the wall make better targets than others; they are more efficient in that they leave fewer 'holes' when the zoid is fitted – in the game you want the zoid to fit as snugly as possible. The skilled player need only think of two dimensions to the task: rotation and the translation. I argue that the position of the zoid in relation to the wall is trivially task-sensitive since it corresponds, in a 1-1 fashion, to the task. A change in the task is constituted by a change in the wall-zoid relation for example if the configuration of the wall should change - or the shape of the zoid.

Let us now check the first co-ordinator condition. We need to show that the zoid-wall relation tracks and triggers other system processes involved in the playing of Tetris. This is an empirical question. What other processes are involved? How are they related to the rotation of the zoid? I do not know of further studies in this case, but I propose that something like the following functional organisation of lower-level processes is plausible:

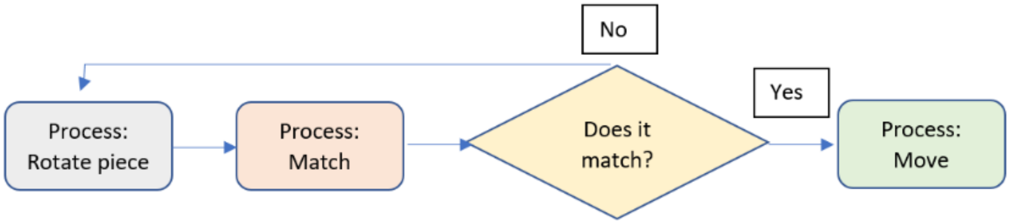

Fig. 6.1 The processes of a Tetris player co-ordinated by the state of the piece relative to the wall – simple algorithm.

If this is anything like the correct picture, then it the rotation of the zoid that triggers an internal matching routine. After each 90-degree turn the procedure kicks in and performs a task that delivers a simple yes/no answer to the question 'does it match?' If there is a match, the system exits the matching process and initiates the moving process. I claim, perhaps surprisingly, that the state of the Zoid tracks the matching procedure. This is because once a match is found the system automatically exits the rotation routine so the fact of still being in the rotation phase implies that a match has not yet been found. Under these conditions then the rotation of the zoid keeps track of state of the matching process and satisfies the first coordination condition.

But this analysis assumes that matching is an all-or-nothing activity and that a match is inevitably found. In some game situations this is not possible for example when the wall does not admit a zoid shaped hole. In this case a more complex set of processes may be in train involving a minimal use of memory. Suppose that degree of fit is measurable. Then a simple algorithm along the lines of "a match is found if it is a better fit than that of the previous rotation" will prove to be relatively efficient. This would involve keeping in memory the degree of fit of the last attempted fit operation. If the current fit is better than this, then the zoid is moved[85]. The algorithm will only perform badly when the sequence of fit scores for each rotation is monotonically increasing. Otherwise, while not always picking out the best fit, it scores reasonably well on average.

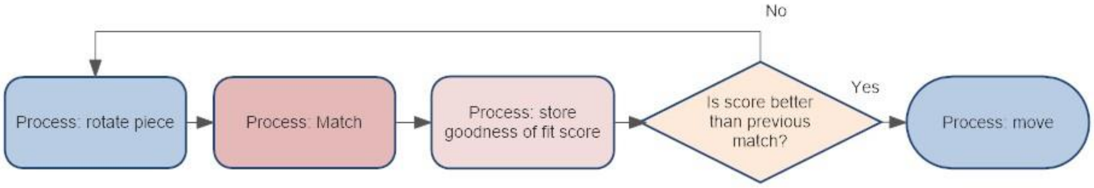
No
Process: rotate piece
Process: Match
Process: store goodness of fit score
Is score better than previous match?
Yes
Process: move

Fig. 6.2 The processes of a Tetris player co-ordinated by the state of the piece relative to the wall – more sophisticated algorithm.

Of course, these are *how possibly* arguments, and a full analysis would need to investigate the actual processes involved empirically, but the CSA provides a

[85] This problem is similar to the optimal stopping time problem of choosing a restaurant from a sequence of n restaurants on a long straight one-way road. The 'take it if it is better than the previous one' algorithm is both computationally simple and effective.

framework to these empirical questions. These arguments are plausible because they place emphasis on the main work being done by the rotation of the piece and the prior development of habits that sustain it, rather than on brain-based spatial rotation and memory.

Such findings would support the claims of Kirsh and Maglio that these actions are epistemic. If they are coordination processes, then they are exactly what gives Tetris play its goal-directedness and by our definition are cognitive hence furninshing evidence for HEC[86]

Kirsh and Maglio do not discount the possibility that the same actions can be both pragmatic and epistemic. In CSA terms this means that the same behaviour is both a performance of a task and coordinative of other performances. How can a single physical process play both roles? How can something that is part of the game also be part of the system that plays the game? AA put this in terms of representation: the shapes on the screen do not represent pieces in the Tetris game, they *are* pieces in the Tetris game (2001, p. 54). Surely the game does not solve itself?

The reader may be reminded of the situations studied in the previous chapter on ecological functionalism, where the system operated upon the world at the same time as parts of the world could be co-opted into the system; there is the same dual function. In chapter 7 we shall see that a central feature of stygmergic systems is that the results of the action of the system on the world help coordinate future actions. These systems are surprisingly common.

[86] It is tempting to speculate that Kirsh and Maglio's epistemic actions are always coordination processes. But it is possible that the class of epistemic actions may be too broad and fuzzy to make this claim stick.

In this light AA's puzzle seems pressing. The answer requires us to acknowledge that the shapes in the Tetris game do indeed play two different functions. Firstly, AA are right that the shapes on the screen are constitutive of the actual game; manipulation of them amounts to playing the game. As pieces in the game, the Tetris shapes play no representational role, they do not represent the shapes – they *are* the shapes. The second function of the zoid is as part of the coordination process for a goal-directed system as described above. It is in this second function that we may permit talk of representation. It just happens that the same physical object plays these two functions. The CSA allows that this can happen. The same physical structure can function both as plant (a TSD) and as coordinator, just as the same component in an aircraft can function both as a wing and as a fuel tank, or the same cylinder is both a printer roller and a battery. Indeed, I suggest below that this is a common state of affairs where building separate coordination structures and processes is expensive to the system.

This becomes clearer in the Tetris example if we devise a thought experiment in which these functions are pulled apart. Imagine an extremely slow (!) game of Tetris where the zoids are almost motionless. Suppose that there is a plastic duplicate of the game, available to the player, consisting of models of the zoids and the wall into which they must be plugged. The player can rotate the plastic zoids and try out moves with them before she makes them in the real (slow) computer game. The plastic duplicate of the game is like a scratch pad. Let the plastic zoids in the model correspond to the real ones on the screen – they are in the correct positions – so that the plastic model is a 'map' of the screen. Now the game function is performed solely by the game and the coordination function is performed solely by the model. In this case, the 'pragmatic' and the 'epistemic' functions are separated and there is no problem asserting that the shapes of the game are just that, while the plastic model is a co-ordinator and part of the cognitive system. Combining the two functions in the actual Tetris human-game system makes it much cheaper for the system because it does not have to generate the ‘extra’ duplicate of the game and keep it up to date. Hence, this analysis suggests that the CSA agrees with CC’s claim that manipulation of the Tetris zoid is partially constitutive of the system responsible for playing the game[87].

6.4 Cognitive bloat and the coupling-constitution question.

> “Look at the size of yer”
>
> “I’m not that big!”
>
> “*He* ate the dormouse else it was me!”
>
> *Beatrix Potter Nursery Rhyme Book*, enclosed CD, Beatrix Potter and Peter Cobbins, (2008)

In the second part of this chapter, we turn to some problems arising from these examples and the parity-style arguments of chapter 1. The first issue is cognitive bloat – that the system extends unreasonably far into the environment in a way that is not explanatorily useful.

Cognitive bloat is linked to the problem of separating system constitution and causal coupling. The CSA can be read as an attempt to show under what conditions (namely the coordination condition) causal coupling counts as system constitution. By saying in effect that not all causal couplings are coordinative it is directly answering the coupling-constitution question. Thus, the hope is to stem

[87] Richard Menary expresses the separation between epistemic actions and pragmatic actions in the following similar terms: “Tetris avoids the coupling-constitution fallacy because epistemic actions are in the problem solving space, not just as a clever strategy for offloading complexity onto the environment but as “part of our cognitive economy”” (Menary, 2010b, p. 568)

the unwanted spread of cognition into the environment. Again, we shall find that things might not be quite so simple.

CSA provides a functional criterion for distinguishing processes connected causally to the system from processes that coordinate the system and therefore belong to it. Simple causal connections are not enough to establish that a component plays a role in coordinating a system with respect to a set of tasks. It must exhibit task sensitivity and bear the right tracking and triggering relations to other processes in performing the requisite tasks. I claim that this removes the immediate risk of bloat.

Recall from chapter 2 the example of Gareth Jones the long-suffering coalminer in South Wales, whose travails help to power my computer. On a purely causal account of system, Gareth would count as part of my computer system. Recall that, for the sake of the argument, because of some strange quirk of Glamorgan Coal Board, Gareth is personally responsible for providing power for my computer. Gareth's absence one day when he went down sick resulted in my computer not working at all. Gareth, it turns out, is causally necessary for the tasks performed by my computer. On causal accounts, he would be co-opted into my computer system. But as we saw from chapter 1, a weakness of these accounts is that they cannot easily distinguish between active components and causally necessary background conditions. On the CSA, however, the question is not whether he is causally connected, but rather whether he plays a functional role that contributes to the coordination function with respect to the normal task space of the computer. I claim that he does not. His work does not bear any sensitivity to changes of task on the computer - it makes no difference to Gareth whether I am writing my thesis or, more likely, playing solitaire. Moreover, Gareth's activities neither track nor trigger the processes responsible for task performance by the computer. Gareth is not part of the coordination process. The most we can say in the absence of necessary conditions is that a CSA account provides no evidence that Gareth is part of the system.

But have we thereby prevented bloat? Might there be a case in which bloat returns but not spatially but in the time domain? Let us return to Otto and his notebook. Let us say that he used a website at some point in the past to enter the address of MOMA in the notebook, and to download a map, print it out and stick it in the notebook. The notebook, and the map it contains, can perform the coordination role only by virtue of its particular causal history including the interaction with the website. Should this causal history be included as part of the coordination process? The worry here is that bloat raises its ugly head again because of the temporal extension of the coordination process.

Examples like this explain why CC found it necessary to add the previous conscious endorsement condition to their glue and trust conditions (Clark and Chalmers, 1998, p. 17), although, as mentioned, it disappears in a later rehearsal of this argument (Clark, 2010a, p. 46). In a footnote, they address the issue of causal history as being constitutive of belief:

> The constancy and past-endorsement criteria may suggest that history is partly constitutive of belief. One might react to this by removing any historical component (giving a purely dispositional reading of the constancy criterion and eliminating the past-endorsement criterion, for example), or one might allow such a component as long as the main burden is carried by features of the present (1998, p. 17 fn).

In writing this CC are keen to point out that the coupling between agents and artefacts is active, as opposed to Putnam's passive content externalism where long causal chains connect representation to content (see Putnam, 1973)[88].

While the CSA does not commit to couching cognition in terms of belief-desire psychology, the same criticism could be raised in terms of the constitution of the coordination process. Coordination processes may take place over long

[88] The same could be said of teleosemantic theories such as that of Millikan or Papineau (see MacDonald and Papineau, 2006; Millikan, 1984).

timescales, and such processes according to the CSA guarantee cognitive systemhood. Does this not imply the unwanted temporal spread of cognitive systems? How can we avoid the conclusion that the long cultural process of establishing norms regarding the use of a public symbol system is not part of the cognitive system and that the system responsible for Otto's navigation to MOMA includes the whole history of the written word?

To head off these objections we might take the same route as CC and draw a contrast between the coordination processes in the here-and-now that relate to the navigation task on the one hand, and the long-term processes that establish the norms for activities such as notebook usage, or the intermediate timescale of the formation of Otto's *habit* of notebook usage on the other. The coordination conditions get this job done. While these long timescale processes provide a background in which notebook usage satisfies the coordination conditions, they themselves violate them. These long-term processes are not themselves sensitive to Otto's task of going to MOMA or the Birdland Jazz Club, neither do they track the processes involved in performing these tasks.

On this basis then I would argue that the notebook potentially functions as part of the system if it satisfies the tracking, triggering and task-sensitivity conditions, even if the causal antecedents are in some way problematic. Consider the example of Twin Otto who wrote that the Museum of Modern Art was on 51st Street in his notebook (Clark and Chalmers, 1998, p. 14). Let us suppose that Twin Otto also pasted the wrong map into the notebook at this time (CC speak of the analogy with memory tampering (1998, p. 17)). The notebook now coordinates the task of going to MOMA, but Twin Otto ends up on 51st street. In this case I argue that the coordination conditions are still satisfied: tracking and triggering work well, and the notebook still bears a counterfactual relation to the task – were the task navigating to the Birdland club, then Twin Otto would open the notebook at a different page. The problem is that the coordination process went wrong because of problems further up the causal chain. I would like to view this as a broken coordinator rather than no coordinator at all. In situations like this there is still an extended cognitive system in place even if it malfunctions.

I said in section 6.2 that the notebook example is complicated because of the confluence of a number of factors including the question of the origins of representational content as this discussion shows. Nonetheless, we learn that cognition may well consist in the coordination of different processes meshing together on widely different timescales. The CSA seems well-equipped to deal with bloat of both a spatial and temporal variety.

6.5 From parity to gradualism?

All well and good, but does CSA saddle itself with other problems instead?

One worry is that the coordination conditions are rather too easily satisfied with respect to systems with small task spaces. This would mean that instead of the extent of cognitive systems being weakly controlled in bloat, there is an unjustified proliferation of cognitive systems. If we accept that a system can possess a zero-dimensional task space, that is perform a single task, then every system is cognitive, and the CSA implies pancognitivism. We could argue that the second coordination condition is vacuously satisfied.

Alternatively, the CSA theorist can argue that such a system has not shown that it can adapt to changes of task since there are none available – therefore does not satisfy the spirit of the second condition and is therefore ruled out. I am inclined to take this view since I see no explanatory advantage in allowing single task systems. Even the TSD’s of Chapter 4 admitted metric variation.

But wouldn’t this strategy still make the lowly thermostat minimally cognitive because it does possess a one-dimensional task space indexed by the difference between the current temperature and the required temperature? One could argue that the state of the bimetallic strip was a tracker and a trigger of the heating processes in the system, and that the same structure was sensitive to the ongoing continuously variable task of transforming the air in the room to the pre-assigned temperature via the built-in temperature control. In some sense, then, the thermostat system is indeed a simple goal-directed system. But is the task space sufficiently large for us to call this capacity a cognitive capacity?

One route out of this problem is simply to say that the thermostat task space is not representative of any cognitive capacity. These are not *cognitive* tasks *tout court*. Richard Menary takes something like this route when he defines a cognitive process in the following terms: “A process is cognitive when it aims at completing a *cognitive task*; and it is constituted by manipulating a vehicle” (2007, p. 57 emphasis added). Even ignoring the appeal to representational vehicles, we adopt this tactic on pain of circularity. How do we avoid appealing to a mark of the cognitive in deciding whether a task is cognitive when the whole *raison d’etre* of the CSA is to supply such a mark?

There is a second objection to the idea that tasks can be sorted into the cognitive and the non-cognitive. In section II the starting point for CSA analysis was that cognition was essentially a property of the processes responsible for a certain kind of persistent and adaptive behaviour. While this is intimately bound up with a sequence of tasks unfolding, there are different ways in which these tasks can be performed. The persistence and adaptability of the behaviour produced by processes possessing a certain type of self-organisation which I have described as robust and versatile is the key to understanding cognition. A *Blockhead* type device (see Block, 1981) which consisted in a giant look-up table may be able to perform a relatively limited set of tasks. But what makes this system cognitive or not is, according to my account, not the specific tasks that it can perform, but rather how sensitive it is to task change (what I have called system versatility). Blockhead would need to be able to adjust its behaviour in line with changing environmental constraints, its own changing abilities, and changes in its distal goal. This goes far beyond its original specifications.

A different tack is taken by P.D. Magnus, who applies a sort of parity principle for tasks: “[for a task to be cognitive] the task must be such that *it would be cognitive* if the process were contained entirely within the epidermis of one individual” (Magnus, 2007, p. 300 emphasis original). The worry here is that we shall end up with the same problems with a parity principle for tasks, as we face with a parity principle for processes. I am not convinced that Magnus’ shift from functions to tasks gets us any further in our problem.

The bold theorist could bite the bullet of gradualism and say that cognition comes in degrees and there is spectrum of cognitive systems with human beings at one end and the humble thermostat at the other. On this view the thermostat is cognitive but minimally so because its task space is just so minuscule. While counter-intuitive, this view finds support in, amongst others, Adam Toon (2016, 2021), Daniel Dennett (1987, Chapter 2), John Dewey (1938), and Wayne Christensen (1996) although for different reasons. Dewey’s naturalism embraced what he called the *principle of continuity* according to which “there is no breach of continuity between operations of inquiry and biological operations and physical operations. ‘Continuity’ (…) means that rational operations *grow out of* organic activities, without being identical with that from which they emerge” (Dewey, 1938, p. 26). Dewey wants to explain mind and its features in terms of bodily operations of creatures who “are themselves the result of an evolutionary history and who have typically passed through a crucial sequence of developmental stages that have shaped their cognitive capacities and their identity” (Johnson, 2010, p. 125). Dewey is not subscribing to pan-cognitivism, but neither is he suggesting that there is a boundary beyond which a creature is cognitive. Christensen’s early work on this question parallels the CSA in terms of the central role for normativity and the sensitivity of the system to the environment: “[The point of this section] has been to stress that intentionality is a continuously graded, multidimensional concept, varying with the foregroundedness of the goal, the degree of explicitness and cultural sensitivity of the information processing in relation to the goal, and the resultant behavioural flexibility" (1996, p. 317). 'Amen' to all that except that the CSA distances itself from an explicit involvement of information processing.

The continuity approach suggests that there is not a sudden gap between basic cognition and higher cognition. Explaining how rational operations 'grow out of organic activities' is the task that faces enactivist theories, themselves inspired by Dewey, that explain basic cognition in terms of embodiment and close coupling between agent and environmental structures, such as the *skilled intentionality framework* of Erik Rietveld and Julian Kiverstein (Kiverstein and Rietveld, 2020; Rietveld *et al.*, 2018; Rietveld and Kiverstein, 2014 see also section 8.4). The problem identified by Richard Menary, among others, is whether the move from basic cognition to higher level cognition on this account involves a discontinuity that is inconsistent with the continuity of evolutionary processes (Hutto and Satne, 2017; Menary, 2015, p. 3). Examining this issue is beyond the scope of this thesis but my intuition is that the CSA's gradualist position is compatible with enactivist stories of basic cognition and promises no sudden discontinuity between basic cognition and higher cognition. But the devil is in the detail.

Alternatively, the slightly more hesitant theorist could venture in and say that there is a point in the continuum where the task space becomes sufficiently large for a capacity to be cognitive. This would mean that there would be a sudden transition from non-cognitive systems to cognitive ones. Again, the danger of circularity seeps back in. The placing of the limit will need careful principled justification to avoid it being arbitrary with the consequence that such a justification would itself end up being a mark of the cognitive.

In view of this, a gradualist view of cognition, although radical, may well be the only realistic option. Any system with a task space and a coordinator is cognitive but not equally so. I shall suspend judgment in the current work – this is not the main aim of my writing - but it is a question that needs to be settled eventually and could be used as some as a *reductio* against the CSA.

6.6 The problem of sterile effects

To complete the chapter, I return to the difficult problem encountered at the end of chapter 2. I shall sketch the outline of a solution made possible by the CSA. It is the whole theoretical structure of the CSA, its process roots, its insistence on functional explanation, and the acid test of the coordination conditions that play a part.

Formulated in process terms solving the problem of sterile effects requires that a theory of cognition can distinguish between ‘active’ processes in the system that are responsible for cognition and those that are causally correlated with system behaviour. Earlier I described this as the problem of separating ‘plant’ processes from ‘cognitive’ processes. It is a more acute version of the coupling-constitution question. It is not just the problem of sorting between processes that are causally connected to the system and provide the background conditions for its operations from the sharp-end processes that constitute the system, like the coalminer problem, difficult though that problem is. It is like a two-way coalminer problem where the computer is coupled to Gareth and, implausibly, Gareth is causally coupled to the computer. Purely causal criteria, such as MUMA cannot serve to pull apart such processes. Chapter 2 introduced the example of blood flow to the hippocampus of a mouse navigating a maze. The activity in the hemodynamic system of the hippocampus is causally correlated with performance of the navigation task. A straightforward explanation is that hippocampal activity needs an energy supply proportionate to the amount of cognitive work it is doing. But no one in the field would argue that the hemodynamic system is the seat of cognitive processing regarding the navigational capacities of the mouse. But why not?

CSA can help solve the puzzle because it is a functionalist theory not a causal one. I stated in 5.2 that two causally-correlated processes may not have the same function. The meshed cogs in a clock are causally correlated but they perform different functions. The coiled spring may be attached to a large cog whose function is to drive the whole system. This is meshed (causally correlated with) a smaller cog with the function of changing the drive ratio. They play different functional roles even though they are causally correlated.

The CSA asks questions about the functional roles of processes in order to determine whether they contribute to the coordination function. Does the hemodynamic system perform coordination duties regarding the navigation tasks of the maze? A Martian without a background in neuroscience might answer that the causal connections lead to that conclusion. But earthbound experts with a wealth of background knowledge to draw upon in addition to the immediate causal data will give a different answer. No. Regarding the function of blood flow as coordinative of navigation tasks means ripping up everything we know about mammal biology. As any high school biology student will tell you, blood flow performs the function of supplying nutrients to and waste products from cells, period. Other experts will chip in with empirical evidence for navigation function attributions for place cells in the hippocampus of mammals and so on (see for example Maguire *et al.*, 2000, 2006). Questions of function are ultimately empirical questions.

A similar question might occur in the investigation of ant behaviour. We might be tempted to point to the mechanisms of the individual ant neurophysiology that make it sensitive to the pheromone trail, as constituting the active component of the system responsible for producing intelligent foraging behaviour. But this would be to miss the point of the emergent organisation of the coordination process which in this case involves the whole system of pheromone trails plus the actions of the ants that produced them.

In each case the subsystem in question is located at the wrong level of functional description to belong to the coordination process and therefore to be part of the explanation of intelligent behaviour. The hemodynamic system is a support system for the processes involved in coordinating the navigation tasks. While it might show some task sensitivity it does not trigger other processes. It does not satisfy the coordination conditions itself. The pheromone sensitivity system within an individual ant is at the wrong level of description to be explanatory of the observed goal-directed behaviour. Even though it is a subprocess in a mereological sense it does not belong to the coordination process. It does not bear the right relations to the task, and while it might contribute to the triggering of the ant's activities it does not track them. That is done by the pheromone trail. On the other hand, it does make sense to think of the ants themselves as constituting part of the coordinator.

There are of course practical difficulties in producing the right kind of functional explanation here. Empirical work does rely on causal correlations which introduces problems for the experimenter. But these questions lie more at the general level of function attributions and rely also upon the historical corpus of knowledge assembled over time in many related fields and not only on causal correlations. Functional explanation must be allied with more substantial theoretical work. There is no silver bullet. The coordination condition is the first part of the explanation and suggests where to look for the appropriate evidence of causal processes.

But the point is that there is no longer any *conceptual* problem. The problem of sterile effects dissolves once we move away from purely mechanistic explanation.

Hence the hard work in chapter 2 showing that it was not only possible but correct to do so[89].

6.7 Concluding remarks

This chapter applied the theory of section II to some of the problems discussed in chapter 1 concerning the material extension of a cognitive system. By analysing examples such as Otto's notebook and Tetris the CSA comes to the tentative conclusion that in these cases there is at least a provisional support for the HEC. In contrast to classical extended cognition arguments, the CSA treats these cases not as cognitive extension of a central agent - i.e. not as agent-centred cognition - but rather as systems in their own right where coordination roles may be distributed over aspects of the agent and material artefacts. In these cases, the arguments of the previous chapters support the idea that the coordinator, as the explanatory core of the system, extends beyond the boundaries of the individual agent and therefore offers support for extended cognition.

This approach suggests an answer to the immediate problems of cognitive bloat and the coupling-constitution question. It has something important to say about the Tetris example and avoids the problems of deciding the functional grain parameter by avoiding a parity driven account altogether replacing it by a broad organisational mark of the cognitive.

Otto's notebook is, in many ways, more puzzling from a CSA viewpoint than the Tetris example. On one reading, it can be taken to play a role in coordinating

[89] The enthusiastic reader might ask why we cannot use the coordination conditions to distinguish between sterile effects and processes in a cognitive system. The problem lies with the possibility that coordination has been established as a sufficient condition for a process to be part of the system. But it is not a necessary condition. There may be processes involved in cognition that are not themselves coordinative.

Otto’s navigation task to MOMA. For it to play this role it must satisfy the tracking condition - in other words its use must be causally loopy. Read in another way, the notebook is just a causal input into the system and therefore does not satisfy the conditions for playing a coordination role and therefore not a core system component.

The CSA does introduce a new problem that does not afflict parity driven theories which is the question of cognitive gradualism. Systems may possess coordination processes with respect to tiny task spaces. If the CSA is correct, and coordination implies cognition, then these systems are cognitive. The CSA theorist trying to avoid this counterintuitive conclusion by stipulating a minimum size of task space for cognition is caught between either an arbitrary limit or a question-begging one. Compared to these options cognitive gradualism might well be the best bet, meaning that cognition is a spectrum from minimally cognitive systems facing tiny task spaces to the unfathomably large space of tasks characterising human cognition.

Finally, I sketched an answer to the question that emerged in chapter 2. The problem of sterile effects which is quite awkward for causal accounts, is more tractable for functionalist ones like the CSA. Each strand in the paraphernalia of the approach has something to say in this example: process, ecological function, and the coordination conditions.

The CSA is general but broad. In addition to material extension, it has a lot to say about socially constituted systems. It is to these that we now turn.

Chapter 7
The CSA and socially extended systems

7.1 Introduction

There is an influential position straggling the collective intentionality, group agency and social psychology literature that asserts that the joint production of goal-directed action by a group of individuals requires shared representations of some kind (Bratman, 1992, 2014; Gilbert, 2014, Chapter 5). These views typically rely on some version of belief-desire psychology. For example, Christian List and Philip Pettit require that members of groups together promote a given goal, each intend to do their allotted part in a salient plan to achieve this goal, believe that others form such intentions too, and each believe that the others believe this and so on (List and Pettit, 2011, p. 33; see also Bratman, 2014, p. 10; Butterfill, 2018, p. 73). Something similar is found in the writings of Adam Morton (2003). The cognitive psychologist and philosopher of science Giovanni Pezzulo suggests that action coordination in a cooperative game such as two people building a tower together with coloured bricks requires the players to align their personal representations of the goal and state of the tower (2011, 2015). Margaret Gilbert writes that “two or more people are acting together if they collectively espouse a certain goal” which involves her notion of joint commitment which is expressed in terms of the same goal being shared by individuals (2014, p. 34). Typically, ‘shared representation’ means that each participant maintains an identical representation to the others, whether it be a goal state or current state of the action. On the other hand there is a minority view that holds that ‘shared’ representations are distributed across individuals and that they do not need to possess identical or even similar representations in order to engage in joint actions (Hutchins, 1991, 1995a, 2014). The CSA can be used as an argument for this latter approach to social cognition. Access to joint coordination processes operating in the world can take the place of individual belief-desire pairs locked away in the brains of participants in the joint action. No one need have an ‘overview’ of the whole action and it is therefore more parsimonious and cognitively cheap than that proposed by the standard view.

The argument is substantially the same as in the case of cognitive systems extending over material artefacts. Afterall, the CSA, being blind to the exact nature of the processes involved in the coordinated system, does not make a distinction between systems involving material processes and those involving social processes. All that is required is that the functional conditions are satisfied for coordination. I introduce three new examples of socially extended systems to explore the CSA in these contexts. The first is based on representations, the second and third are not representational.

The second part of the chapter takes us into neo-cybernetic territory. I argue that a common feature seen across many of the examples in this chapter is *stygmergy*. A system is stygmergic if the material result of action of the system in the world coordinates further actions by the system. The example given in the chapter concerns a group of people clearing a messy classroom. In this case, under the operation of suitable social norms, the state of the classroom, which is the result of previous clearing actions, coordinates further actions. One can find this in other examples such as Tetris, formation of paths in the snow, and cases involving systems of organisms such as eusocial insects. Indeed, if there is a problem with the original example of Otto’s notebook, it is that it is not stygmergic enough! Combining stygmergy with the CSA produces a powerful model for understanding many examples of cognition in the wild. At the end of the chapter, I use this to explore recent questions about the cognitive status of social institutions such as the legal system.

The reader might ask how clearing a messy classroom, providing babysitting services, or settling legal disputes is cognition. But these social systems, just like the Tetris or Otto systems, produce goal-directed behaviour, which is the notion of cognition that guides this thesis. What is interesting about these examples is that they involve a system comprising a group of individuals subject to social norms and conventions. They provide a helpful contrast to the examples considered in the previous chapter, since it is not possible to conceive these systems as an extension of the system of a single individual. Therefore, these systems move the argument away from the conception of extension in CC, where a system such as Otto-plus-notebook possesses a central locus of control, namely Otto. They align nicely with the coordinated systems model where control is potentially distributed throughout the system. I shall say more about distributed cognition in the next chapter.

7.2 1960's babysitting collective

I shall introduce a new example to show how the CSA can cope with the production of coordinated action in groups while still possessing some parallels with the original Otto-like cases in CC.

Growing up in the 1960's I experienced many instances of social co-operation where people organised themselves into self-help or mutualist groups devoted to making life easier through collective effort. The general idea was that individuals agreed to perform duties as part of the collective in order to reap individual benefits in turn. For example, my family belonged to a babysitting cooperative comprising a dozen families. The economy of the group consisted of an internal currency of 'babysitting credits'. The system was organised by a 'Babysitting Book'. Each family was responsible for operating the book for a month. If a family needed a babysitter on a given occasion, they would telephone the bookkeeper giving details, and the bookkeeper would systematically write in the book the name of the family requiring the service, and the date and time it was required. The bookkeeper would then go through the list of names of families in the collective, starting with those who had the lowest net babysitting credits (since it is a zero-sum game the search always started with those who were in negative territory; the total ‘money supply’ of credits is always zero). When a babysitter was found who agreed to the task, an entry was written in the book and the original caller informed as to who would be coming. After the duty had been discharged, the entry would be marked closed in the book, the contributing family would be awarded with one credit, and the receiving family debited by one credit.

This is a loosely connected system of semi-autonomous agents directed towards the goal of providing babysitting services. According to the stance taken in this thesis the system is cognitive in the sense that it produces goal-directed behaviour. Admittedly the task space is relatively small, given the formal architecture of the system, so it lies more towards the thermostat end than the human end of a spectrum of cognitive systems (see section 6.5)[90].

The elements of the system are the different parents in the cooperative, the babysitting book, and the various communications media. The general goal translates, in the babysitting environment, into specific tasks of the form: babysitting for family X at time t. The collective is disposed to perform the task through its organisational structure, and through the skills of its members.

I claim that the Babysitting Book, and actions of the bookkeeper in writing and reading from it, play a coordination function with respect to the set of babysitting tasks. The claim is justified by checking off the coordinator conditions from the functional organisation of the system. The entries in the book, together with their

[90] This is interesting given that it possesses human members. The formal constitution of the system might serve to constrain the kinds of action available to its human membership. Nonetheless there is considerable potential for the system to overthrow its formal structure and reconfigure itself on the fly to respond to unforeseen eventualities (although those with experience of any sort of committee work might not be quite so optimistic regarding the flexibility of such systems in the face of unusual problems!). This potential for self-organisation pushes it along the spectrum away from the thermostat end. These considerations will be ignored, and this toy example will be studied as formally constituted.

manipulations, trigger the behaviour of subprocesses in the system, namely the actions of a particular parent going out to babysit for X at t. There are also representations tracking the progress of this activity, namely the inscriptions that close the entry when the task is complete. Thus, the Babysitting Book contains representations that trigger systems activity directed towards the goal and track that activity. In the words of computer science, it is a *task scheduler*. But these inscriptions also track the task facing the system taken to be the current babysitting request in the form of an entry in the 'requests' column. There is regularity in the relation between the task, that is, the babysitting appointment, and the representation in the book. This relation supports counterfactuals: if the task were changed, then the book would be changed to reflect this. For example, suppose a parent of the family X called subsequently and said their babysitting requirement had changed and that they required a sitter at a different time, then bookkeeper would amend the entry. Therefore, the babysitting book satisfies the task-sensitivity condition.

The book does not perform this function on its own. It plays this role by virtue of the manipulations of the bookkeeper; it is the whole process that coordinates. The coordination condition allows us to proceed from function to constitution. Because of its broad functional profile, the book is, perhaps surprisingly, part of the system responsible for the performance of babysitting tasks of this kind.

There are lessons to learn from this example. Babysitting action takes place within a complex set of social norms, practices, and conventions that help define, or constitute, the babysitting task. These norms place constraints on the type of activity that counts as babysitting; it would not do, I take it, when looking after young children, to bring along one's Stratocaster and a 300W amplifier and practice thrash metal riffs downstairs when the little ones were trying to sleep. This normative framework governs the task of babysitting and imposes constraints and success conditions upon it. Similarly, the Babysitting Book can only function as part of a coordinator within the system because of the existence of symbols in a public language, norms and standards for public timekeeping, norms that govern the writing and reading from it, norms around the order in which potential babysitters are polled in arranging a match of demand and supply, and so on. Without these norms or something like them, the Babysitting Book could not play its part in tracking the babysitting task and its implementation.[91]

In this example the process of manipulation of the Babysitting Book by the bookkeeper constitutes the coordination function. But there are other processes in the system that are not coordinative such as the actual performance of the babysitting tasks. Recall that in section 3.4.1 these were called *plant* functions or processes and they were associated with task specific devices. In this case coordination functions can be pulled apart from plant functions. This means that the coordination conditions do not identify the whole system in this case. Nevertheless, coordination processes *do* belong to the system wherever they are. Worth mentioning here too is the fact that the Babysitting Book coordinates other local coordinators such as the diaries and agendas of individual family members, that coordinate other tasks that are incidental to the babysitting performance. These coordinators together form an 'ecological' system which I shall say more about later in the chapter.

A detailed examination of the example helps give clarity to the distinction between coordination and plant functions. Consider the telephone lines (such things existed in the 1960's) passing messages between the different members of the babysitting group. I claim that they do not perform coordination duties despite playing a crucial causal role. One reason for this is that they do not satisfy the

[91] Considerations similar to these suggest that it might be fruitful to apply the CSA to economic markets (for the opposite view see Huebner, 2014). It would be interesting to see to what extent the concept of coordination meshes with Hayek's notion of a *catallaxy* a spontaneous emergence of social organisational structures (see Hayek, 2003). Interestingly, there are a number of stygmergist texts analysing Hayek (see section 7.7).

task-sensitivity condition. If the babysitting task changes, there is no corresponding change in the state of the telephone wires. They are best considered as part of plant rather than coordination processes – they are not responsible for the goal-directedness of the system. On the internalist side, it seems plausible that there are structures in the brain that are merely plant, perhaps some long-term memory structures play this role[92]. Again, this supplies further evidence that the satisfaction of the coordination conditions is sufficient but not necessary for cognition.

This discussion illustrates the difference between the CSA and standard informational accounts. Dan Weiskopf (2010) identifies cognition as occurring in an informational medium between transducers and effectors. Susan Hurley describes such a view as a sandwich model - where the cognition takes place in the middle layer between perception and behaviour – and she sees it as a mistake (see 1998, pp. 247–249). Weiskopf's criterion would bring the telephone lines into the cognitive system because they link transducers - perceiving information about the babysitting environment, and effectors - the action of babysitters themselves. In cases like this, the CSA is more deflationary than information-based approaches, since it identifies only the core coordinating processes as the key to generating goal-directed behaviour rather than supporting informational infrastructure.

In this sense the CSA sees coordination as one function among many, and there are plausibly cognitive functions that are not coordinative. The CSA licenses a tripartite ontology of subsystems: cognitive processes that are coordinative, cognitive processes that are not coordinative that I call plant, and non-cognitive,

[92] The cognitive scientist Tarja Susi includes 'placeholding' as a characteristic of coordination. Were she to take a CSA view she would be more likely to include passive memory processes as coordinative (see Susi, 2016).

non-coordinative supporting processes (units outside the system but causally connected to it).

It really does depend on one’s explanatory project how wide the system net is cast. Explanation of a system that exhibits goal-directed behaviour with respect to a task set τ will be based on a system whose core plays a coordination role with respect to τ. There may also be plant processes involved in supporting these coordination processes; the explanation will take these for granted even though they contribute to the performance in an essential manner, such as functioning as constraints that canalise active coordination processes. For example, cognitive neuroscience is interested in the firing of neurons but not primarily in processes in the cytoskeleton of the neuron, even though it defines the infrastructure in which makes neuron firing not only possible but intelligible. These supporting and constraining processes come into the spotlight when the investigation moves to questions about how the system processes are constituted but they are not relevant in the first instance regarding questions about the source of intelligent behaviour.

Having had our fill of babysitting let us now tackle the thorny problem of clearing up after a children’s party.

7.3 Clearing a messy classroom.

Katya Abramova and Marc Slors (2015, 2019) invite us to consider the following real life example showing how context provides norms for a set of tasks and their performance. The scene is a messy classroom after a children’s party at a pre-school in Nijmegen, Holland. A group of dismayed parents view the scene of devastation before them. There is orange juice spilled on the desk, cake on the floor, dirty plates and cutlery scattered around, balloons and streamers everywhere. The goal of the parents is to return the classroom to its former pristine state ready for a fresh start on Monday morning. Social norms define the goal and the constraints acting on performance of the goal. Only certain states of the classroom count as clean and tidy, and only certain methods of clearing it up are socially legitimate. For example, blowing the classroom up, although it does remove the mess, violates the constraints of the clearing up operation which include leaving the basic infrastructure of the classroom intact and close to its original state. There are tools to hand that can be used in the cleaning up operation such as vacuum cleaner, mop, cloths, dustpan, and brush and so on. One parent grabs the vacuum cleaner to get the crumbs off the floor, another a cloth to clean the tables and a third takes the dirty plates to the sink. There is no central controller telling them what to do. Seeing what the others do helps to define the tasks available for a given parent. When the vacuuming of the floor is complete, the socially sensitive parent now looks around to see what other tasks are available and selects the mop to wash the floor without being told what to do. In time all the tasks are done, and the classroom is clean and tidy again.

Slors uses this example to explore the cognitive role of cultural convention in allowing action coordination in servicing a group goal (Abramova and Slors, 2015; Slors, 2019, forthcoming b). I shall use it as raw material on which to try the CSA. I claim, perhaps surprisingly, that the state of the room plays a part in the coordination process and therefore should be included in the system responsible for intelligent clearing-up behaviour. Let us examine each of the co-ordinator conditions one-by-one.

Starting with the triggering condition, the state of the classroom, jam smears on the table and so on, trigger the actions of a parent who, seeing that no one is using it, grabs the cloth and scrubs away. Gradually the jam smears and crumbs are removed - the state of the table measures the progress of the task so there is a sense in which the state of the room tracks the cleaning processes. This is only possible because of the social norms that govern turn-taking in co-operative ventures, cleanliness of classrooms, and the mode and order of cleaning tasks. The first coordination condition is satisfied.

The second condition is that the state of the classroom bears the right relation to the task. Again, just like the Tetris case, we find that this relation is trivially satisfied by the situation. Against the horizon of norms for cleanliness, the existence of crumbs and jam on the table correspond to the task that is their removal. The relation of the state of the world in relation to an ideal goal state corresponds to a task. Therefore, the state of the classroom automatically satisfies the second coordination condition and is part of the coordination process[93].

There is some hidden subtlety here in that the goal toward which the clearing up process is directed is not unique. The social norms operating on clean classrooms do not define a unique end state but rather an equivalence class of them[94]. Perhaps one clean classroom has the chairs under the tables while another has the chairs on the tables. This means that the tasks facing the system are not uniquely determined and there is an important stage in coordination dynamics in which the system settles on one or other of these goal states[95]. How the parents decide collectively on whether chairs should be on tables or underneath them depends on many factors. There might be an explicit agreement made through a conversation, but there might also be a tacit understanding arrived at through interlocking behaviour governed by the relevant norms of social interaction. One person puts a chair under a table, and another just puts it on the

[93] We shall see that task-sensitivity is automatically satisfied in cases of stygmergic coordination of which this is one see section 7.5.

[94] This fits with our notion of function as being a transformation between two equivalence classes of states of the world.

[95] Perhaps there is a sort of meta-stable dynamic process at work where the system is pulled towards one rather than another attractor state (see Bruineberg and Rietveld, 2014).

table. The first feels that it is churlish to undo this action so the default becomes chairs-on-tables. The way in which a goal is selected from an equivalence class is not our immediate concern. All that matters is that the system is sensitive to the general task and that, by one means or another, more specific tasks become available.

Some readers might feel uneasy about the idea something could be both part of the output of the process (the gradual cleaning of the classroom) and part of the system that is producing this output (part of the coordination process for the system). I agree that this looks dubious if our idea of a system is a fixed machine-like item into which inputs are deposited, flow through the system, and outputs emerge. But remember that we rejected this metaphor in chapter 5. There it was argued that the system and its environment can become entangled and that the operations of the system on the environment can also be construed as playing a role in coordinating the system, which is precisely the situation we have here.

A further intuition pump might help to take the sting out of this objection. Let us suppose that the goal is the clearing of a messy after-party classroom as before, but this time let us suppose that one of the parents, Frida, personally coordinates the clearing up task. She does this by drawing a map of the classroom on a piece of paper indicating all the subtasks that need doing such as clearing dishes from a table, vacuuming a section of floor and so on. She then assigns a parent to each subtask. The parents go off and perform their tasks as Task Specific Devices. When they are done, they report back to Frida who checks their work, and if she is satisfied with it, crosses the task off on her map, thereby updating the available tasks. Clearly, Frida’s map is related to the tasks in the right way, and dynamically tracks and triggers the various cleaning subprocesses. Frida’s map seems to play the same functional role as the Babysitting Book in 6.3. By the same reasoning then Fridas manipulation of the map are part of the coordination process for the clearing tasks. Functionally, Frida and her map are no different from the conjunction of the room and the parents in the original example. The classroom itself simply takes the place of the map – we would expect them to possess a degree of similarity in the relevant way because that is the nature of the 'mapping' relation. Frida must be able to construct and read her map (and manipulate it to show the completion of tasks). But parents must be able to 'read' the state of the classroom and its deviation from that of a 'clean' classroom as well as being able to read what each other are doing. If Frida and her map play coordination role, then so do the parents and the classroom. In other words, the parent-classroom constellation is a distributed co-ordinator[96].

This example is interesting because it captures the sort of generic co-operative set-up characteristic of many collaborative social situations. If the analysis is correct, then in this kind of scenario, the material features of the task itself play a double function as part of the co-ordinator. The system engulfs parts of the environment on which it is operating. The relevance to extended mind cases should be clear: rotating Tetris zoids and the classroom are on a par.

7.4 Herding virtual sheep

The examples discussed so far in this thesis are typical philosopher's examples, taken from everyday situations and not amenable to rigorous experimentation. To offer a contrast, I introduce a new example that is at the forefront of multiagent systems and social coordination research to investigate how the CSA fairs with a more experimentally rigorous situation. Patrick Nalepka and his colleagues (2017) ran a series of experiments involving two participants in a game of herding

[96] The sensitive reader might worry that this thought-experiment, just like that for the Tetris game, seems to rely on a new Parity Principle; namely that if a process is functionally equivalent to a process that we would 'unhesitatingly' regard as coordinative, then the original process plays a coordinator role. If this is an argumentative strategy, then the reader would be right that it is vulnerable to the same problems as the original PP (see chapter 1). But as stated at the top of the paragraph the comparison is purely an intuition pump not a principle. The CSA does not rely on functional comparison because it employs *absolute* functional conditions for coordination.

virtual sheep. The experiment takes place on a semi-opaque table on which a 'playing surface' is projected from below, consisting of symbols representing a set of sheep, a desired circular area in which they are to be herded, and a controllable sheepdog per player. The players adopt positions on opposite sides of the table, and can clearly see each other's movements, the position of the sheepdogs, and the sheep on the table. The sheep are programmed to move according to Brownian motion but are repelled by the presence of a sheep dog within a certain distance in a manner such that the speed of repulsion is inversely proportional to the distance from the sheepdog. Each player is to move her sheepdog so that the sheep move into the circular target area.

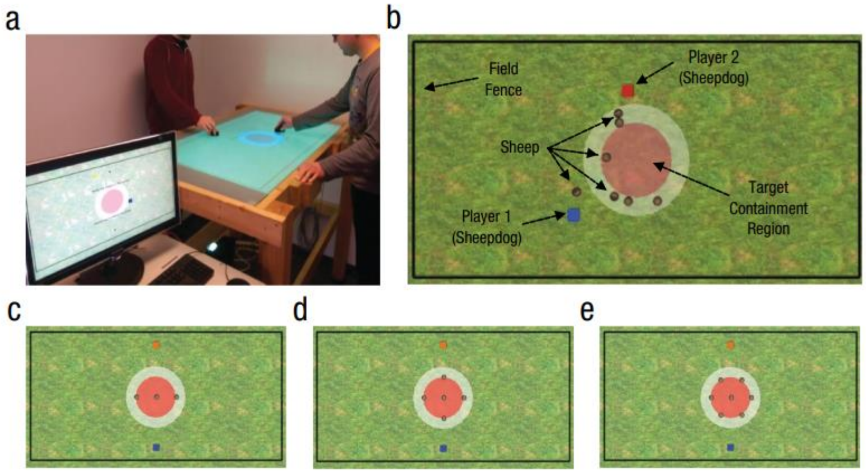

Figure 7.1 The set-up of the sheep herding game Nalepka et al (2017). Photo a shows the game setup, b the projection on the table, c, d, and e the state of the experiment in the 3, 5, and 7 sheep condition (reproduced courtesy of Sage journals).

The main finding of this experiment was that there were two main modes of behaviour of the players. The first mode usually adopted in the game is described as 'search and recover' (S&R). When a sheep ventured too far from the target area the player would direct the sheepdog to herd the miscreant back into the flock. Typically, this was not a successful strategy, since this single sheep herding procedure tended to cause the deviation of other sheep. The S&R strategy is essentially a single player strategy. Once the players tacitly agreed areas of the playing area to patrol, they become responsible only for the deviant sheep on their patch and were not interested in what the other player is doing.

More successful was a strategy named ‘coupled oscillatory containment’ (COC) where players would sweep their sheepdogs in an arc, in phase with each other around the central flock of sheep (much like the operation of trained sheepdogs in fact). Players stumbled upon the second strategy later in the series of trials. Once discovered, this strategy was never abandoned. The moment at which the behaviour ceased being individual S&R and became COC, was a distinct transition in the control dynamics - a type of phase transition. The COC strategy is true two-player cooperation. The coupled sweeping of the sheepdogs in arcs around the central group of sheep is a coordinated action. I contend that once it emerges in the game, the COC action is part of the coordination process.

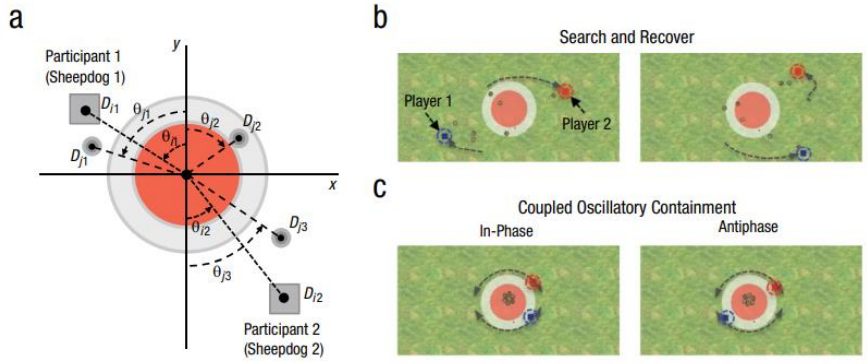

Fig. 7.2 The two types of behaviour dynamic in the sheep herding game. Picture a showing the angles measured in the experiment, b showing the S&R strategy and c the COC strategy (Nalepka *et al.*, 2017 reproduced courtesy of Sage journals).

Let us examine this claim in more detail. There can be no denying that the movements of the players in the game contribute towards to the performance of the game tasks, they are part of plant processes. But my claim is that they are more than this, that, like so many of the examples we have examined, these movements play a dual role. At the same time that they enact actions in the game, they also perform coordination duties. The idea is that each player can see the movements of her partner and that there can be a process of entrainment or phase-locking. This is the phenomenon where the oscillatory motions of two agents end up locked together, with the same frequency and a constant phase; much like two pendulum clocks fixed to the same wall. Processes such as this have been studied extensively not least in the action coordination literature in connection with musical performance and dynamical systems (Bosga and Meulenbroek, 2007; Goupil *et al.*, 2021; Høffding and Satne, 2019; Nalepka *et al.*, 2018; Richardson *et al.*, 2016; Walton *et al.*, 2015, 2018; Washburn *et al.*, 2019; Wolf *et al.*, 2019, 2020; Zamm *et al.*, 2015). Classical entrainment studies show that phase locking dynamics have a strong attractor at phase=0 (in phase locking) and a weaker attractor at phase=180 (antiphase locking). Nalepka et al find evidence for both types of coupling in their experiment (2017). The details need not concern us here, the key idea being that the behaviour of the players in the game is best described as a single coupled system rather than two separate autonomous systems. Sufficient for my purposes is that there is a process by which the coordination function is realised.

The claim that the actions of the players also function as a coordinator needs to be checked against the coordinator conditions. Do the movements of the sheepdogs trigger and track processes in the system? There are at least two ways in which tracking can take place. If a player is attending to the bodily motions of her colleague, then these roughly track the motion of the sheepdog. If a player attends only to the position of the sheepdog on the screen, then, of course, it tracks the positioning process, and because of the periodic nature of the COC motions can be used as a predictor for future movements. If the motion of the other player and the position of the sheepdog is hidden, then it is possible that the position of the sheep track the behaviour of the other player in an indirect manner[97]. The tracking condition is clearly satisfied in the first two cases – and in the actual experiment I suspect that it is satisfied through a combination of observing bodily movements and the motion of the sheepdog on the table. There are many examples to be found where behaviour of a system is also a tracker of the processes of the system - think of the phenomenon of the Mexican wave - the wave itself tracks the behaviour of its components, namely the up and down movements of the many people constituting the wave. The behaviour of each player acts as a trigger for the other. The oscillations phase lock and become stable. This is what the dynamics of rhythmic coupling tell us.

The second condition is that the coordinator is task sensitive. In a complex dynamical system like this one, evidence in favour of task-sensitivity would be that the co-ordinator process changes in a significant manner with the task. Nalepka and his team report that as they change the task by changing the number of sheep to be herded, the behaviour of the players changes. Players are far more likely to use S&R successfully in the 3 sheep condition than the other conditions and (therefore) less likely to stumble upon the COC strategy. In the seven sheep condition, players are far more likely to employ the COC strategy. These different strategy behaviours are indicative of the different coordination requirements of the two tasks. Therefore, the coupled behaviour of the players is part of the coordination process of the system.

This result has some interesting consequences. The first is that something interesting happens at the point where S&R behaviour gives way to COC behaviour. Nalepka et al point out that this can be represented dynamically as a classical Hopf bifurcation (see Thom, 1975, p. 96). This is significant from the view of systems ontology. It seems plausible that S&R behaviour is best viewed

[97] I suggested this possibility to Patrick Nalepka who tells me that he has plans in the pipeline for such an experiment (Nalepka private communication).

in terms of two separate and autonomous systems, perhaps thought of as similar to two games of Tetris played on the same screen with each player attending to just one part of the playing area[98][99] However, after the bifurcation to COC, we may view the whole as a single system coordinated by the coupled movements of both players. Nalepka and colleagues are clear about the quite different characteristics of this system.

An important lesson to learn from this example is that it is not the physical set-up that marks out systemhood, but rather the dynamical and functional organisation. The same physical set-up can be two autonomous systems in one dynamic regime and a single coupled system in another. The transition between the two is a point of interest both in this example and in general. Moreover, the example seems to vindicate the starting assumption in the analysis that we can think of a system as a coalition of elements that may change dynamically.

This emphasis of the dynamical structure of the system rather than its mechanistic structure is highly reminiscent of enactivist approaches to cognition which are discussed in the next chapter. The move away from synchronic ideas of system constitution towards dynamic diachronicity is characteristic of Second and third Wave extended cognition theories. Moreover, the CSA offers an alternative to the 'mindreading' approaches to joint action coordination mentioned in the introduction. Instead of coordinating joint action by attributing beliefs and

[98] In S & R mode, once the tacit agreement is made to divide up the playing space, each player performs more or less independently of the other. They are not entirely independent since the actions of one player do have a causal bearing on the actions of the other via the movement of the sheep. If there is coupling, then it is very weak compared to the COC mode.

[99] Basil Wahn has performed some interesting experiments on joint search tasks where player coordination emerges. These also make good examples of emergent coordination processes where the search of each player is coordinated by the movements of the other, eye-tracking of the other player is made available via a cursor on the screen. Unfortunately there is not space here to provide further details of these excellent examples of distributed cognitive systems (see Wahn *et al.*, 2017; Wahn, Karlinsky, *et al.*, 2018; Wahn, Kingstone, *et al.*, 2018).

goals of the other person, the action can be coordinated through the dynamics of the interaction as we see in this example[100].

7.5 Stygmergic coordination

The Tetris, classroom, and herding examples, and others we have dealt with are examples of *stygmergic systems*. This is such a large and important class of systems that it will occupy us for the rest of the chapter. I should say at the start that although there is a large literature on stygmergic systems in the engineering, multi-agent systems, and cybernetics literature, it has not really made an impact in the 4E sphere, so the analysis in these last three sections is, to the best of my knowledge, new.

What made things relatively straightforward in these examples, was that the task sensitivity coordination condition was automatically satisfied and the tracking and triggering condition easy to verify. This is because *part of the world on which the system was acting functioned as part of the coordination process*. This can be summarised in four basic conditions.

(1) The system acts on some locally defined part of the world X

(2) Some aspect of X corresponds to the task facing the system

(3) The current state of this aspect of X serves to track the system’s actions

(4) The current state of this aspect of X triggers future actions.

[100] More generally the CSA suggests a way to understand human interaction without positing the use of folk-psychological mental state attributions. The CSA provides resources for explaining joint action, especially in the context of stygmergic coordination, without needing mindreading or simulation theories.

Because of these conditions X is part of the coordination process. In the Tetris case X was the state of the zoid in relation to the wall. In the classroom case, X was the state of the classroom in relation to an ideal clean state. In the herding task, X was the state of the sheep, and the sheepdogs relative to the goal circle. In each of these systems, *the result of previous actions of the system coordinates further actions.* This helps us to answer the standard question raised in each of these examples: how can something be both a result of the action of the system and a coordinator for it? This was effectively the question that AA asked about Tetris, and they could have asked it about the classroom and herding examples too. I want to show that this is not a bug of the analysis but a powerful feature of stygmergic coordination. The characteristic of stygmergic systems is that plant function and coordinator functions are realised in the same physical process[101]. To understand how this works we need to do some modern cybernetics.

The Belgian cyberneticist Francis Heylighen wrote: “[s]tygmergy is an indirect mediated mechanism of coordination between actions in which the trace of an action in a medium stimulates the performance of a subsequent action” (Heylighen 2016:4). I shall explain what this means in terms of the systems that we have been studying. The word ‘stygmergy’ comes from two Greek stems: ‘stygma’ meaning mark or sign and ‘ergon’ meaning action or work. The idea is that elements of the system leave marks or signs in the environment which stimulate other elements to act. These elements may themselves be agents - but in my analysis I shall just take them to be processes in the system. Heylighen refers to the part of the environment on which the system acts to be the *medium*. By making a distinction between the whole environment and the medium he wants to draw our attention to the fact that the system does not have access to the whole environment. The medium then is the accessible environment that

[101] In sematectonic stygmergy see below.

would be described by von Uexkull as the *Umwelt* (von Uexkull and O'Neil, 2010)*.* It is the immediate part of the environment that matters to the system. The distinction need not concern us too much at this point though when I refer to a system acting upon the world, properly I mean the system is acting on that portion of the world that it has access to – the medium.

The key idea then, is that the action of the agent leaves a *trace* in the medium that then acts to stimulate further actions. The trace is X in the definition above. The trace may be something like the pheromone marker laid down by ants when foraging, specifically for the purpose of coordinating ant foraging activities - this is called *marker stygmergy*. Or the trace could just be the result of the action of the system in performing its tasks, such as the state of the classroom in which case it is called *sematectonic stygmergy* (Marsh and Onof, 2008, p. 137). The 'agent' may be a process in the system such as the behaviour of individual parents in the classroom situation, or it may be the whole system itself. Indeed, the trace might be the result of the action of the whole system, but it might nonetheless stimulate[102] individual elements within the system. In nearly all the interesting examples we have looked at so far, the trace is the result of action of the system on part of the world, playing a coordination role for the system. A stygmergic system, then, is one where the agent is sensitive to the trace in the right way so that the trace plays a coordinative role with respect to the task set of which the trace is the performance. *Stygmergic traces automatically satisfy the coordinator conditions with respect to the systems that produce them*.

[102] Heylighen uses the term 'stimulate' rather than 'trigger' because he wants to allow a probabilistic relation between the trace and system processes rather than a deterministic one.

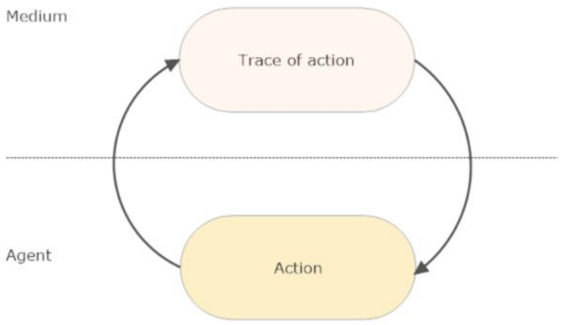

Fig 7.5 The basic principle of stygmergy.

Heylighen offers the familiar example of building a house. The foundations need to be laid before the construction of walls and floors and roof. All these need to be complete before the electrical system is installed. The well-trained electrician knows the state that the building must be in before she can ply her trade – she knows what exactly she needs to look for. This state then stimulates actions involved in constructing the next stage. The electrician is, in effect, a task specific device discussed in chapter 4. She engages in specific processes triggered by the state of the house and tracked by it. In general, a TSD is a soft assembled subsystem that performs just a single quite narrowly define task or set of tasks. Task specific devices are coordinated by condition-action pairs - when the trace satisfies a given condition it stimulates a certain action of the TSD. In a perfect world, the electrician would get to work when the structural aspects of the house are complete. In the real world, of course, she may have other jobs on that delay the installation (the voice of bitter experience!). The state of the building both tracks and triggers the next phase and defines the current task, explaining why a trace of a stygmergic system automatically satisfies the coordination conditions.

Heylighen's stygmergic account leads us to understand how part of the world – the trace – can be both the result of action by the system and can coordinate further actions. By showing how certain physical processes can be realisers of both plant (non-coordinative) functions and of coordinator functions at the same time – it solves the puzzle of the ‘double function’ that occurred in the Tetris and classroom examples. This means that it tracks current (and past) actions - because it is a trace of those actions, and it triggers new actions by the system which produce alteration to the trace which alter the tracking property and so on until the end (goal) state is reached (see figure 7.6). This description not only answers AA’s worries about the state of the Tetris game being both a game and a coordinator but hints that this a common situation in the world especially regarding social systems. In the words of systems theorist H. Van Dyke Parunak: “[it] would be more difficult to show a functioning human institution that is not stygmergic, than it is to find examples of human stygmergy” (Parunak, 2006). Readers may worry that this signals the re-emergence of bloat, or more accurately an unwarranted proliferation of cognitive systems, but I would rather put this more positively, that many human social institutions perform cognitive functions. Indeed, one may argue that this is not a coincidence and speculate upon the role of such structures in the evolution of human cognition.

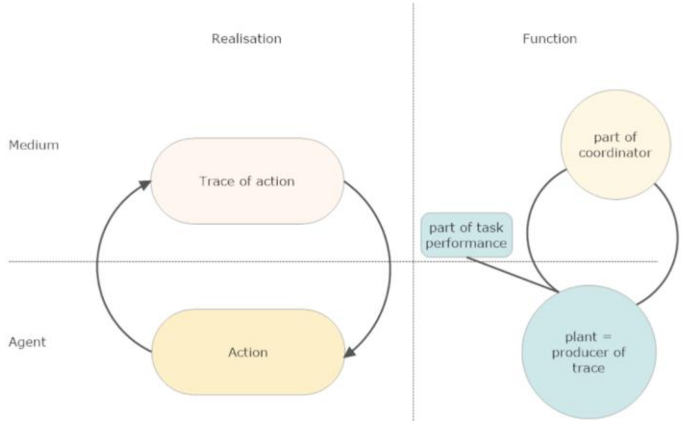

Fig. 7.4 How stygmergy solves the plant/coordination duality problem. The left column shows the processes at work and the right column the functional roles that they play. The horizonal panels correspond to whether the process belongs to an agent or the medium.

7.5.1 Stygmergic coordination and the classroom example

The stygmergic approach solves the problem of how the state of the classroom can be both the result of actions of the group of parents clearing it, but also part of the coordination process for further clearing actions. This is because the system is stygmergic and the state of the classroom is a trace. The stygmergic analysis shows that the classroom example belongs to a common kind of coordinated system.

How the system becomes sensitive to the trace is a separate matter and is explained by a story (that needs to be told) of the establishment of social and cultural norms, practices and conventions (see for example Boyd and Richerson, 1985; Henrich, 2016; Menary, 2013, 2018; Tomasello, 1999; Tomasello *et al.*, 2005). In the case of the messy classroom, it is the capabilities of the parents and their ability to apply social norms regarding what counts as tidy and what kind of state specifies a given task - what I shall call their cleaning-up abilities - that gives the trace of the cleaning process - the state of the classroom - a coordination role. The socialisation that gives them these capacities also includes what we might call ‘common sense’ about the order in which tasks should be performed and about keeping out of each other’s way. The parents in this example are semi-autonomous agents who, in their interactions with the classroom, also participate in the coordination function. This suggests that the coordination function is emergent from the actions of the parents.

Some might consider this to be a problem with stygmergic systems - that we cannot neatly separate system performance functions from coordination functions. But it is precisely this that can be seen as a feature that makes the system amenable to the CSA in the demarcation problem and is a consequence of a lack of modularity[103]. The coordination function being distributed throughout the system means that the coordination conditions also identify the system. The lack of separation of task performance and coordination is not a bug, but a feature of the approach.

Stygmergic systems are generally of this type, especially if they are, like the classroom example, sematectonic - that is the trace is the result of system plant actions - as opposed to marker stygmergy that involves a special marker whose function is purely coordinative. This means that the coordination function is linked to the action of the whole system, making it relatively easy to identify.

I have said that loopy or distributed systems that fail to be modular or nearly decomposable, turn out to be good systems to demarcate because the coordination demarcation criterion serves equally well as a system demarcation criterion. In the light of this discussion, not only does this this loopiness appear in a concrete form in stygmergic systems, but that such systems are remarkably common.

7.5.2 Tropism and trace-aspect sensitivity

This section focuses in more detail on how to understand the ‘salient aspect’ of the trace that does a lot of work in explaining stygmergic coordination. It also casts light on the nature of tropism (see chapter 4) which turns out to be a feature of stygmergic coordination.

Stygmergic coordination depends on the existence of some aspect of the trace that reliably correlates with the completion of the main task. For example, there is some aspect of the incomplete building, such as the existence of basic walls

[103] Note that lack of modularity in a system is not the same as a lack of division of labour as this example shows.

and floors, but not finished or painted surfaces (I am guessing here), which indicate, in the eyes of a suitably qualified electrician, that the time has come for installing the ring main and other electrical features. It is this aspect of the building rather than, say, its height, that indicates the state of completion and the readiness for electrical installation. This aspect is correlated with the relevant notion of completeness and therefore the salient aspect of the trace for performing coordination duties. Let us look at this relation in more detail.

Let us start by considering what we mean by an aspect of a trace and by reliable correlation with task completion. By an aspect of the trace, I mean a value of some environmental variable that is a property of the trace. In the example of the house, it is the existence of basic fittings like floors and ceilings. In any house building project then, let us say, this is a reliable tracker and trigger, and is correlated both with what has been done already but also with what tasks are still needed.

The salient aspect of the trace functions as a kind of proxy indicator for the degree of completion of the main task. What constraints are there on an environmental variable for it to function in this manner? What is the nature of the relation between this variable and the tasks facing the system? To answer these questions let us consider a parable.

Imagine a somewhat absent-minded philosophical lobster fisherman, Ludwig, who goes out in his boat every morning with empty lobster pots to set them in a promising place in the sea. Then he returns home. In the evening he goes out again to the pots and hauls them into his boat before returning to land. Let us suppose that our lobster fisherman does not want to think too hard about what he is doing because he is lost in thought regarding lobster cognition so is a bit absent-minded. What aspect of his work could he use to keep tabs on where he is in the cycle?

One candidate is the position of his boat. He can look up from his musings and spy the distant shore to try and work out what he was doing. But this will not work because he cannot tell from his distance from the shore whether he is travelling out to the pots or returning to land. The position of the boat does not distinguish between them.

But what if he took the position of the boat and the direction in which the boat was travelling as the relevant aspects of his trace. Again, this would not work, because on looking up from his musings and noticing the position of the boat and the direction of travel, he still cannot distinguish between the morning session when he places the pots and the evening session when he hauls them in again.

But what if he took a further aspect into account: the contents of his boat? Now, we have something! The position, direction and contents together bear a regular relation to the task. If he is out at sea and moving away from the shore and the boat contains pots then he is going out to place them, if the boat is empty then he is going out to collect them. If he is out at sea and moving towards the shore, and the boat is empty then he has just placed the pots, if the boat contains pots, then he has just collected them. Were nature to make philosophical lobster fishermen stygmergic, she would do well to make them sensitive to position and direction and make them aware of the presence of pots in the boat. Together the three aspects of position, direction, and presence or absence of pots are salient to the lobster fishing goal.

What is special about the relation is (1) it is a regular relation between the task performances and the trace and (2) that it is a one-one relation. (1) is necessary to tether triggering of the task to the correct stage in the process and (2) is to ensure that a performance is not triggered by the wrong stage of the trace. In the first two suggestions in the parable, the mapping between the performance process and the suggested trace was regular but violated the one-one condition and therefore would not do as salient trace-aspect variables[104].

What happens when a usually reliable one-one correspondence is disturbed? Imagine an evil demon who wants to disrupt the routine of poor Ludwig. Let us assume that as Ludwig was on his way out one afternoon to bring in the pots, the demon filled the boat with empty pots. Ludwig comes out of his philosophical reverie, sees the pots, and triggers the setting pot routine instead of the collecting pot routine. The demon's interference has broken the regular relation between trace-aspect and triggering function and broken the stygmergic coordinator.

But this is precisely the same problem that we encountered in the discussion of the *sphex* wasp of chapter 4 and its peculiar antics. Recall, that in dragging a grasshopper back to its nest, the wasp deposited it at the entrance of the nest and then checked the nest for predators. If none were found it would enter the nest with the grasshopper. If, while checking the nest, the grasshopper were moved, the wasp would go through the whole process again. In that chapter I claimed that the task space of the grasshopper was rather small because it did not encompass (to us) obvious adaptive solutions to the 'displaced grasshopper' problem. In the light of the Ludwig case, we can see that we were perhaps a bit premature in writing off the wasp. By analogy with the Ludwig case, it is better to think of the wasp-grasshopper system as stygmergic, and the position of the grasshopper as a salient aspect of the coordinating trace of the wasp's actions. The position of the grasshopper generally correlates well with progress of the dragging/checking tasks because it is in a reliable one-one correspondence with the completion of this task. Intervening and moving the grasshopper, just as with the demon and Ludwig, breaks the regular relation between the position of the grasshopper and the stage of completion of the dragging/checking tasks. The

[104] This is the same condition that we required of trackers in chapter 3.

position of the grasshopper ceases to be a tracker and therefore ceases to be coordinative. This explains why the wasp's behaviour stops being optimally goal-directed. Similarly, if we mess around with coordinating neural structures in human primary visual processing, we would not expect the same optimal performance of visual tasks either. Kudos to the wasp!

An even better example is *Abispa ephippium* which is a favourite in the stygmergy literature. It is a solitary Eumenid wasp that builds a remarkable funnel-shaped nest out of clay.

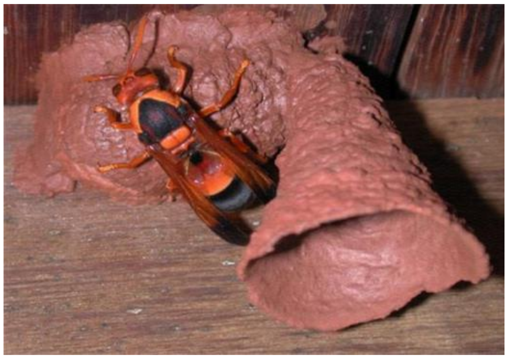

Fig. 7.5 The funnel-shaped nest of the Eumenid wasp Abispa ephippium plays a stygmergic coordination role (photo from Matthews and Matthews, 2009 creative commons attribution licence - Hindawi Journals).

Theraulaz and Bonabeau (1999, p. 102) discuss how each stage of the building process of the nest is triggered by the previous stage - the nest itself acting as a coordinating trace. In a sense this is the perfect insect analogue of the human house-building example discussed above. The wasp is tropistic with respect to the nest-building tasks. If, in the final stage of the project, a hole is made in the side of the nest structure, the wasp will start again from the beginning building a completely new nest on top of the old one. Clearly the nest topology is the aspect of the trace which is salient in coordinating the construction process. A nest with a hole in it is not part of the sequence and is, from the wasp's point of view, not recognisable as any stage of the nest building process. Hence the whole project needs to be started from scratch. The normally reliable link between the aspect of the trace used as a tracker, its topology, and the completed task sequence has been broken. While the wasp undoubtedly has the physical capability of repairing the hole, it is not a task that is triggered by the trace.

This example shows that a system that is stygmergically coordinated relies upon a particular aspect of the trace for its tracking and triggering functions. Nature, society, or development picks out an aspect that is reliably correlated with the completion of the task and ensures that the agent is sensitive to the aspect so that it can operate as a trigger and a tracker. If this correlation is broken, then the trace cannot function properly to coordinate the system performance and tropism is the result.

There are other interesting studies of stygmergy, for example in internet software projects (Bolici *et al.*, 2016), path formation (Goldstone and Roberts, 2006), or a fascinating study of the adversarial stygmergic coordination of drug cartels and law enforcement across the Mexico-US border (Nieto-Gomez, 2016). In the latter it turns out that the busting activities of Homeland Security act like a stygmergic signal pointing out to the cartels the weaknesses in their systems which they can then correct. While I do not have space to develop an analysis of these systems using the CSA in the present work it is worth noting that stygmergy need not always be cooperative.

To summarise this section then, systems possess stygmergic coordinators when the trace of their action in the world plays the coordination role. It can do so because agents in the system have developed sensitivity to certain aspects of the trace that correlate reliably in a one-one fashion with the completion of the task. This can be built-in by evolution in the case of wasps or encultured by society in the case of the parents cleaning the classroom.

7.5.3 Advantages of stygmergic coordination: the world as its own model

From the point of view of the system itself, adopting a stygmergic coordination strategy offers a number of advantages. The use of a salient aspect of the trace of system performance is a way of getting coordination on the cheap. The system does not have to spend resources on building an elaborate internal structure to represent task planning and completion. All that is needed is the ability to recognise the appropriate aspect of the system's own changes to its local environment as suitable for tracking and triggering tasks. The art to building stygmergic systemhood is creating this trace-aspect sensitivity in the first place. First, an appropriate aspect of the trace needs to be selected that bears a one-one correspondence with the completion of the main task. Then the agent or agents need to develop a sensitivity to this aspect in order that they may be triggered by it. In the example of the classroom this is created through socialisation, while in the examples of stygmergic insects this is a result of evolutionary processes[105]. The main advantages to the system are reduction of costs through doing away with the need for generating expensive internal representations. As the roboticist Rodney Brooks wrote: "[w]hen we examine any simple level intelligence, we find that explicit representations and models of the world simply get in the way. It turns out to be better to use the world as its own model" (1991, p. 140). Stygmergic coordination is one way of understanding this claim.

There are other advantages for stygmergic coordination especially in multi-agent systems. The trace of a system composed of many autonomous agents allows for asynchronicity, or buffering, of task performance that would be difficult to

[105] Another example is the topology of highways and the provision of roadsigns that transform position in a journey into a sygmergic trace.

achieve using a centralised coordinator. Since stygmergic triggering happens through condition-action pairs, agents latch on to the appropriate condition when it is manifested in the trace. If there are enough agents, then this becomes a highly optimal ‘just-in-time’ system, because as soon as the system is ready for the agent the relevant trace condition triggers the right sort of agent. Eusocial insects, such as ants or termites, are a good example of this in action. The behaviour of ants can be modelled easily on a computer using an object-oriented platform such as *Starlogo* (Resnick and Klopfer, 2018). Each ant can be conceived as an autonomous agent programmed very simply to be interested in food, to disperse pheromones when it is carrying food, to be attracted to pheromone trails and to engage in basic proximate avoidance behaviour to avoid traffic jams (Resnick, 1997, p. 60). In addition, ants are programmed to wander around the screen in a random Brownian pattern. Once an ant has discovered the food source through its random wandering it is programmed to return to the nest. Doing so leaves a pheromone trail which will attract other ants. The more ants are attracted the stronger the trail becomes. When the food source runs out ants will no longer return to the nest and the trail will lose its strength, because of the diffusion properties of pheromone trails. Other work by Resnick (1997, p. 67) shows that automatons programmed in this simple manner collectively solve simple optimisation problems, for example, being able to make economically rational decisions between richness of food source and distance from the nest.

Another reason for a system to adopt stygmergic coordination is that it solves the tricky problem of task sequencing. In our discussion of Tolman (chapter 4) we arrived at the idea that goal-directed behaviour consisted in a coherent sequence of tasks performed in the right order. This places big demands on coordination since tasks not only have to be selected corresponding to the repertoire of performances available to the system, but they must be selected in the right order to deliver on the main goal of the performance[106]. Coordination of task performances by a trace takes care of task sequencing for a system. Imagine a succession of states of the trace X_1, X_2, X_3, triggering task specific devices T_1, T_2, T_3. Suppose that the states of the trace are such that X_2 can only appear after X_1 and so on, then the tasks T_1, T_2, T_3 will always appear in the correct order. Since a house with walls but no roof occurs before a house with a roof and basic fittings, the task of building the roof will always be before the task of putting in the electricals. The key question is whether there is a trace that satisfies the one-one trace-aspect to task condition. As we saw in our parable, different environmental variables might need to be combined to from a trace-aspect that satisfies the one-one condition. Moreover, there may always be situations in which the reliable one-one trace-aspect to task relation breaks down. When this happens, there will be system failure, and however sophisticated the system, it will appear tropistic. Tropism is a necessary consequence of stygmergic coordination.

These are advantages of a stygmergic coordinator from the point of view of the system with its finite resources.

7.6 Symbiotic cognition

Having introduced the notion of stygmergic coordination it is time to return to the more familiar territory of extended or distributed cognitive systems and use the CSA to adjudicate on a new debate in the literature which has stygmergic underpinnings, concerning a new type of cognitive ecology proposed by Marc Slors (2019).

[106] An alternate formulation of this problem is that at any level in the system, tasks typically underdetermine behaviour constituting their performance. The order of tasks in classroom clearing is a good example here.

Slors' idea is a response to an interesting and provocative paper by Shaun Gallagher and Anthony Crisafi (2009), who produce an argument exploring the possibility of a different kind of cognitive extension to that of Otto and his notebook (2009, p. 47). They consider a situation in which a person, Alexis, is given a set of facts and is asked to make a legal judgment in three different circumstances. In the first case she is given facts and evidence and is asked to make a judgment based entirely on her own sense of fairness. The evidence is considered entirely in her own head without help or interference of others. In the second case, she is given a set of facts and presented with a collection of evidence and asked to judge the legitimacy of a certain claim that is being made. But this time she is given three questions by a group of legal experts and provided with a set of possible answers from which she may choose one, but she can also decide to formulate her own answer. In the third case, Alexis is provided with the evidence and questions as before, but this time the experts inform her of a set of pre-established possible answers from which she must choose, and a set of pre-established rules she must follow in answering the questions. However, for each question she is to choose between only two answers and is not allowed to formulate her own.

Gallagher and Crisafi, shadowing Clark and Chalmers' strategy at the opening of The Extended Mind (1998, p. 7), ask how much cognitive processing or cognitive effort is present in these cases? They conclude that each case seems to require the same overall cognitive effort, but that progressively less effort was required by Alexis in the second and third cases. In these cases, they argue that the cognitive effort is distributed across a number of heads, and that the categories and concepts that Alexis would have engaged in her own head are being provided by external sources. In the third case the conceptual framework is being provided by precedent and law. "That part of the cognitive process that in the first case involves cognitive schemas that run on Alexis' brain, in the third case is replaced by cognitive schemas that are processed according to the rules of a legal institution" (2009, p. 48). Significantly for our project they write:

> The practice of law, which is highly cognitive (and communicative) is carried out via the cooperation of many people relying on external (and conventional) cognitive schemas and rules of evidence provided in part by the legal institution itself; it depends on a large and complex system, an institution, without which it could not happen. It is a cognitive practice that in principle could not happen just in the head; indeed it extends cognition through environments that are large and various. (2009, p. 48).

Is the legal institution partially constitutive of the cognitive processes involved in Alexis' judgments? This possibility is embraced explicitly in the following passage:

> It seems possible, then, to extend Clark-Chalmers' version of the extended mind, usually exemplified in terms of notebooks and such, in the direction of these larger processes where we may be able to think of social institutions as contributing to the constitution of extended cognition. (Gallagher and Crisafi, 2009, p. 49).

The worry with this passage is that it seems to threaten cognitive bloat on a large scale. Can it be the case that the legal system extends Alexis' mind? The claim sounds less dramatic translated into the terms of this thesis: is it true that Alexis and the legal system constitute a cognitive system? Could the coordinated systems approach support this claim?

We shall start modestly, by trying to identify possible coordinative processes responsible for Alexis' judgments in the third case. Let us assume for the sake of the argument that Alexis has noted down on a sheet of paper the questions, the evidence in the case, the two possible judgments for each of the three questions, and the rules that must be applied in making these judgments. Also suppose that she writes down her interim thinking - she jots down thoughts that contribute to the overall judgment, and that the writing of these thoughts triggers other thoughts relevant to the case. Perhaps she crosses out an earlier entry and writes down an amendment. At the end of the process, she writes down her judgment and the train of thought that led her there, given the input from the experts and the institution.

It is therefore entirely conceivable that the processes of writing and reading from the sheet of paper partly coordinate the whole process of coming to a judgment. Since the overall task consists in the set of questions posed to Alexis, a change of task is simply a change in one or more of the questions. The inscription of the question on the piece of paper, therefore, bears the right sort of task-sensitivity required of the coordination conditions. Writing down interim thoughts triggers new thoughts and the passage of Alexis through the process of coming to a judgment is tracked by the inscriptions on the piece of paper. It is therefore not inconceivable that they satisfy the coordination conditions – it is another case of stygmergic coordination[107]. In all, the theoretical framework that I have outlined could quite happily endorse the process of writing on paper as part of the cognitive machinery for the legal judgment tasks and that Alexis and her notepad plausibly constitute an extended system in this case.

However, Gallagher and Crisafi go further in suggesting a cognitive role for the background legal system. Does the CSA provide evidence for this view? We can agree that the legal system provides inputs into the process of Alexis' judgment, in that it supplies precedents against which the case should be judged and gives an indication of what facts are salient. However, it is not at all clear that the rest of the legal system plays a coordination role. The rules and precedents in the legal system are not sensitive to Alexis' tasks or tracking the processes of her judgment. The point of rules and precedents is precisely that they *do not* change in response to what the law wants to do with them[108]. Therefore, it seems

[107] Perhaps unsurprisingly the writing of this thesis is itself an example of sematectonic stygmergic coordination.

[108] It is possible, especially in jurisdictions based on precedent, that Alexis' judgment might indeed feed back into the norms, practices, conventions, and precedents that characterise the legal

implausible that the legal system plays a coordination role in this case. The machinery developed here does not support the further claims of Gallagher and Crisafi.

Slors argues, from a different angle, that the legal system does not constitute an extension to Alexis' cognitive system. Instead, there are good reasons to suggest that institutions such as the legal system play the role of a background set of norms, practices and conventions against which the process of judgments of Alexis may not only take place, but without which these processes would be unintelligible (Slors, 2019, forthcoming b, forthcoming a). A legal judgment is not possible if there is no legal system in place to make sense of the concept of a legal judgment, or the terms which it employs. Legal tasks depend for their significance and intelligibility on a background legal system. To borrow a term from Ed Hutchins, the legal system is a kind of cognitive ecosystem in which tasks are highly dependent upon other tasks, but that individuals are not closely functionally integrated with others (Hutchins, 2014).

Slors suggests a nice way of visualising kinds of extended cognition on a two-dimensional grid (see fig 7.6). On one axis he places the dimension of functional integration which he takes from Richard Heersmink (2015). This is an extension of the *glue and trust* conditions of Clark and Chalmers discussed in chapters 1 and 6. These conditions specify that the coupled external resource is reliably available and typically invoked (Otto always carries the notebook around with him and regularly consults it, that the information is automatically endorsed and so on). On the other axis, Slors places a scale of task dependency sketched above, conceived as the degree to which tasks depend, for their existence or intelligibility, on the external resource. For example, in the case of signing a

system. But this is a long-term feedback loop of the kind that we encountered in terms of the role of culture in habituation in Otto's notebook use. It is not sufficiently direct to count as part of the coordination machinery here-and-now.

contract, the task depends very heavily on legal notions of property, ownership and what legal powers a signature holds. This means that signing a contract is partially constituted by these legal notions - the task is not even intelligible without them. On the other hand, so the argument goes, the task of navigating to MOMA makes perfect sense even in the absence of the notebook.

Slors argues that distributed cognition of the type described by Edwin Hutchins in his description of navigation on the US navy vessel the *Palau* is an example where the task dependency is high (see 8.2.2). The tasks facing individuals in the navigation team on the bridge depend for their salience on other tasks performed by others elsewhere on the ship. This functional integration is installed and maintained through the practices of naval discipline.

Less radical positions such as embodied cognition occupy the top left square in the table where low task dependency (on external resources) combines with low functional integration. For example, embodied positions emphasise the role of bodily processes but do not consider these processes to be tightly integrated with external resources as to constitute extended systems. Moreover, friends of (non-extended) embodied cognition typically allow that many tasks encountered by bodily processes make sense without taking environmental resources into account, such as digestion.

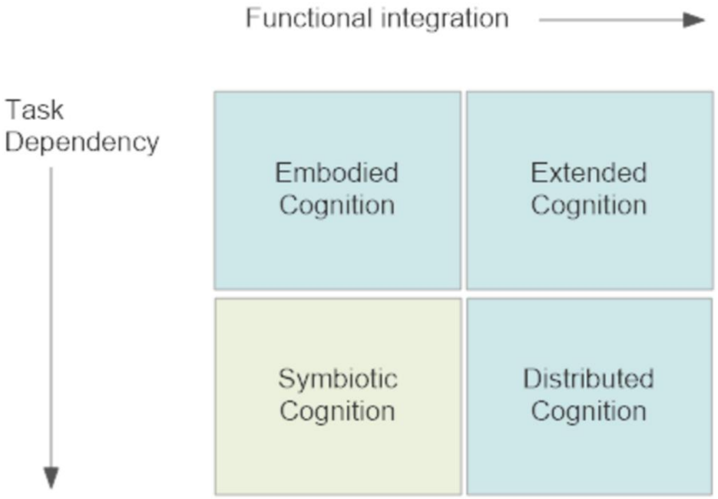

Fig. 7.6 Two dimensions of cognitive extension, (see Slors, 2019, p. 1191).

Slors argues that extended cognition and distributed cognition are different in terms of their task dependency while both possessing a high degree of functional integration. The tasks of different elements in a distributed system, according to Slors, are inter-defined and inter-dependent in the same way as tasks within an institution such as the legal system. Extended cognition such as in the Otto case requires tight functional coupling too. But “however constitutive these items are for these [cognitive] processes, the tasks of remembering or calculating are themselves intelligible in abstraction from tasks carried out by other people” (2019, p. 1192). Thus, extended cognition is placed at the top right of the table.

Slors points out that there is an empty square in the table, and an absence of discussion in the traditional debate, where low functional integration meets high task dependency. He goes on to claim that this is typical of cognition in a social setting. When one is out shopping in the supermarket, the various tasks involve a whole array of cultural conventions, norms and social roles that enable role coordination and from which they derive their meaning. For example, the cashier at the checkout engages in specific tasks shaped by the practices, conventions, concepts, and norms governing supermarket life. How else can we explain the question “have you got a bonus card, love?”. Ditto with the reciprocal roles played by the customer and the tasks she performs – there is high task dependency here. But, Slors correctly argues, the cashier and the customer are not highly functionally integrated. The cashier does not follow the customer around the shop like a mobile notebook - there are no glue and trust conditions operating here. Slors calls this fourth quadrant *symbiotic cognition* and claims that it better describes the legal system than ‘extended cognition’.

We have seen that the machinery of the CSA also provides a way of distinguishing between, say, the legal system and the classroom clearing group. In the latter, the parents formed a tightly knit group whose actions together with the state of the room were coordinative of the whole. The task of signing the document, or making legal judgments, on the other hand, did not confer coordination status upon the entire legal system. Given that the CSA seems to be sensitive to the distinction that Slors makes between extended and symbiotic cognition, it is interesting to bring the methods developed here to bear on a case of symbiotic cognition.

How can the methods of the CSA as understood in stygmergic coordination be used to understand a case of symbiotic cognition such as the supermarket example? When Barbara is on dinner duty for the family, she is required to decide a menu, shop for the required ingredients, prepare and cook these ingredients, and serve the meal. Now consider the job of going to the supermarket in the light of this complex sequence of tasks. The shopping list she carries that helps her coordinate the shopping task depends on the recipe that she has consulted and the ingredients that she has taken out of the cupboard in preparation for the meal. The shopping list coordinates the shopping task and bears a relation to the main task. It is therefore a sub-coordinator for a smaller set of subprocesses that play a part in the service of the main goal of providing the dinner[109].

But as part of the shopping process, Barbara interacts with the supermarket cashier Alan. Alan's performance is not coordinated by Barbara's meal coordinators nor her shopping list. His part in their joint behaviour bears no relevant relation to the meal task. If the meal task were changed - say Barbara changed her mind about the menu and bought different items in the supermarket, it would not significantly change the behaviour in their joint interaction. Alan would just pass the items across the scanner as usual. From the point of view of the meal task his role is plant not part of the coordination process. As in Alexis' legal case, there are no good reasons for considering the supermarket institution to be part of the meal system[110]. It is a horizon of social significance making the interactions in the supermarket possible but does not constitute the core of the supper production system.

But now there seems to be an interesting tension which is best understood in terms of stygmergic traces. It is plausible to regard the coordination of Barbara's actions in the supermarket as being at least partly stygmergic. She collects items and places them in the trolley crossing them off the list at the same time - a trace. At the checkout, she performs a series of joint actions with Alan. These are coordinated by their joint trace - the placing of the items on the belt, the advance of the belt, the scanning of the items and the putting of the items in the bag.

[109] A sub-coordinator is a coordinator for a subtask that is sensitive to the main task but only tracks and triggers the subtask.

[110] In all these examples, of course, there may be other good reasons for inclusion. But this is not the point that I want to explore in this section which is devoted to how much the co-ordinator conditions get us on their own.

At each point the trace acquires its coordinative properties from the horizon of norms and conventions. The joint trace in the supermarket coordinates both Barbara and Alan against a set of norms, for Barbara concerning customer behaviour, and for Alan for checkout personnel behaviour. At this point Barbara and Alan are briefly a joint system coordinated by the joint trace. But Barbara has a high-level goal of providing supper that evening, and Alan has a high-level goal of performing his supermarket duties well. This is puzzling. How can the joint trace serve as a sub-coordinator to both high-level goals?

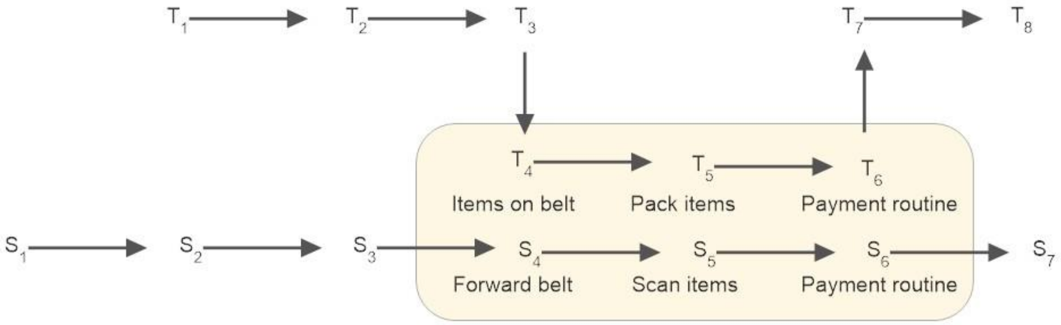

Fig. 7.7 Different aspects of the joint trace (in box) perform different coordination duties relative to higher-order tasks: supper-making (for Barbara) and supermarket serving (for Alan).

The key to this puzzle is that Barbara and Alan are sensitive to different aspects of the joint trace that bear the appropriate relation to the different kinds of task completion present in their joint behaviour, reflecting their different high-level goals. In the case of Barbara, the salient aspect of the trace consists of specific objects on the conveyor belt that bear the right relation to the supper project (via the intermediate coordinator of the shopping list). If the supper project changed in some substantial manner, the identity of these objects would also change. Alan, on the other hand is not sensitive to the identity of the objects but only interested in the location of a barcode that can be read by his scanner. What is of concern to him is that the belt keeps pace with the speed at which he is scanning the items. That fulfils his side of the joint task. Relevant to his goal is the throughput of supermarket goods and orderly management of the checkout. Despite being coordinated by different aspects of the joint trace, the interaction is nonetheless a joint task and establishes a short-lived subsystem both of the supper system and of the supermarket system that we might think of (literally) as a *trading zone*. I propose that the existence of these trading zones is characteristic of symbiotic systems[111].

A brief interjection on the term ‘trading zone’ is useful here. Peter Galison originally coined the term in philosophy of science to refer to the way in which practitioners in two different disciplines found a way to coordinate their actions in a joint project (1997). His original example involved particle physicists communicating with engineers in connection with the construction of particle accelerators. Galison writes: “Two groups can agree on rules of exchange even if they ascribe utterly different significance to the objects being exchanged; they may even disagree on the meaning of the exchange process itself” (1997, p. 783). In the paragraph above the term was used to signify a situation in which two individuals engaged in different tasks are coordinated in a joint interaction by being sensitive to different aspects of a stygmergic trace. I suggest that the two uses of the term are closely linked and that stygmergic coordination may be a way of understanding the cooperation found in Galison’s original trading zones[112].

[111] The functional processes in these zones do not fall neatly under a nice Cumminsian functional analysis but rather exist in the intersection of two such functional cascades serving two quite distinct system tasks. This is in keeping with the fact discussed in chapter 3, that sub-coordinators, as processes, do not obey a strong transitivity relation (see Seibt, 2018, p. 117).

[112] The STS concept of *boundary object* may also be useful here (Star, 2010; Star and Griesemer, 1989). Boundary objects exist in trading zones and are liable to different interpretations by different actors in the zone. A coordinated systems analysis may be used to understand these objects as being stygmergic traces whose multiple aspects coordinate the various actors in the interaction.

Marc Slors is right that social conventions play an important role in joint action-coordination in these trading zones (2021). The reason is that they provide a way in which the joint trace can possess dual aspects that are linked in a regular one-one correspondence to two quite different system tasks. The participants in the joint task are equipped, by familiarity with the convention, to ‘read’ the relevant aspect of the joint trace and to be triggered by it. If I am right in the previous section, we can remove the scare quotes. Barbara and Alan are literally reading the coordinative instructions of their joint trace, but their instruction sets are different.

The whole set of supermarket interactions should not be thought of as a large distributed system but rather an ecology of distinct but connected trading zones in which systems are soft assembled coordinated by different aspects of common system traces. The CSA and Slors’ framework seem to be complementary. Social conventions are indeed coordinative through producing stygmergic coordination for the joint action that remains, nonetheless, sensitive to the separate task structures of the individuals (or individual systems) involved. The institution of supermarket practices is not one giant system but rather a goal-directed or cognitive ecology supporting much smaller systems. In this way then the risk of massive cognitive bloat threatened by Gallagher and Crisafi’s institutionally extended cognition is avoided.

7.7 Concluding remarks

Socially extended cognition can be seen as a set of interlocking processes operating over different timescales entirely in keeping with the spirit of the CSA. What emerges from a discussion of the babysitting collective, the classroom, herding game examples and supermarket examples is the important role played by social norms. If causal infrastructure constrains the organisation of material systems, then social norms, practices, and conventions constrain define the organisation of social systems. The CSA can be applied to both kinds of system because of its more abstract functional machinery.

The second part of the chapter introduced an important feature of many cognitive systems: stygmergy - where the trace of the actions of a system in the world play a role in the coordination of those very same actions. Stygmergy gets system coordination on the cheap. Instead of building elaborate representational structures to support coordination processes, the system makes the results of its own action coordinative. It does this through developing a sensitivity to an aspect of the trace of its actions that bears a regular one-to-one relationship with the progress of the system task. In biological systems the requisite sensitivity arises through evolutionary and developmental processes, in social systems through the operation the requisite social practices governed by norms - thus explaining the pivotal role of social norms, traditions, and conventions.

Prior to this chapter in the thesis, the CSA has been applied to existing debates in the field. However, symbiotic cognition is a new development and a promising one. The CSA does have something to say here about situations where coordination happens in a trading zone involving two systems with quite different task sets. A coordinated systems framework makes sense of this more loosely bound cognitive ecology. There is further work to do here in investigating how the breakdown in transitivity in the coordinator structure of symbiotic systems parallels the breakdown of transitivity in the ontology of processes. Thinking about these situations in terms of stygmergic coordination suggests some interesting avenues for research. There are interesting links with the social cognition literature, with work in philosophy of science and STS, but also with work in situated and swarm robotics (see for example Spezzano, 2019) and multi agent system engineering (see Parunak, 2006; Ricci *et al.*, 2007). Finally, the CSA offers an alternative to standard folk-psychological approaches, mindreading and simulation, in understanding joint action. Agents need only be sensitive to the appropriate aspect of a coordinating trace. Stygmergy takes care of the rest.

Chapter 8
Implications of the CSA

8.1 Introduction

In this penultimate chapter we take stock theoretically and situate the CSA in relation to later extended cognition accounts that are more socially oriented and not specifically parity driven. How does the CSA sit in relation to integrationist or cultural evolutionary accounts of extended cognition and enactivism?

The last two chapters discussed the CSA in connection with materially extended systems and those that were extended in the social domain. Chapter 6 was devoted to applying the CSA to questions arising from parity driven accounts of extended cognition. In engaging with these questions, the CSA naturally presents itself as an alternative to parity driven accounts of extended cognition such as that of CC. We saw in that chapter that there were significant differences in these accounts and that while the CSA seemed to solve some of the issues that threatened first wave accounts, such as the coupling-constitution question and cognitive bloat, it raised new questions of its own such as the issue of cognitive gradualism.

The CSA being a non-agent-centred approach is naturally at home in social examples and Chapter 7 showed that it was able to identify cases of extended cognition in the babysitting example, the classroom example and the virtual shepherding example. In addition, we saw how the CSA gave us a way of understanding stygmergic systems providing a powerful tool for analysing many systems both in nature and in the human social sphere.

A strength of the CSA is that it does not assume a role for representations in cognition. It is an interesting question whether representations drop out of the theory as a consequence of the coordination conditions, for example. If they did (spoiler alert) does this mean that the CSA could be used as the basis of a theory of a certain kind of representation?

At the end of the chapter, we shall have a sense of how far the CSA has taken us.

8.2 The CSA in relation to existing approaches.

John Sutton (2010, pp. 193–201) usefully grouped positions sympathetic to a version of the HEC into a number of ‘waves’. The first wave is identified with the Parity Principle and the original CC paper, and, as we saw in Chapter 6, the CSA offers a viable alternative to these theories avoiding the problems associated with parity by appealing to a basic mark of the cognitive in the form of the coordination conditions. This way causal coupling and constitution are separated because not every process that is causally coupled satisfies the coordination conditions – causal correlation does not imply similarity of function, as we discussed in Chapter 5. Moreover, cognitive bloat is avoided because the coordination conditions are rather strong conditions on what constitutes the core of the system. In its place, however, is the threat of the proliferation of cognitive systems arising from the cognitive gradualism that the coordination conditions introduce. I pointed out in Chapter 6 that this may be a feature of the account rather than a bug, because it paves the way for an argument from basic cognition to higher cognition such as that of Dan Hutto and Glenda Satne (2015, 2017).

It is important to situate the CSA in relation to second and third wave theories in the extended cognition literature given the examination of examples of socially extended systems in Chapter 7. How does it differ from existing theories given that it too emphasises the interactive, integrated, and diachronic nature of cognitive systems? It is also worth explicitly contrasting the CSA with the distributed cognition approaches of Hutchins and Kirsh.

8.2.1 Second and Third Wave extended cognition

Parity driven extended cognition is not the only game in town. Some responses to the parity argument calved promising theories of extended cognition in their own right. Here I briefly introduce Second and third Wave accounts and show how the CSA relates to them.

The second wave arises out of the observation that Otto's notebook is functionally *dissimilar* to human biological memory and it is precisely for this reason that it integrates into the Otto notebook system (see Wheeler, 2019, p. 83). The notebook complements Otto's processing rather than functionally duplicates it. This leads to two related approaches emphasising either the complementarity or the integration. John Sutton puts this nicely: "Second-wave [extended cognition] is based on a *complementarity principle:* in extended cognitive systems, external states and processes need not mimic or replicate the formats, dynamics, or functions of inner states and processes. Rather, different components of the overall (enduring or temporary) system can play quite different roles and have different properties while coupling in collective and complementary contributions to flexible thinking and acting." (2010, p. 194). As Kirchhoff and Kiverstein put it, in second wave theories, though they have different formats, dynamical properties and functions compared to internal, biological states and processes "(…) they can be brought together through agent-environment couplings to make complementary but heterogeneous contributions in the performance of a cognitive task." (2019, p. 11). One of the main streams in second wave thinking is the 'cognitive integration' framework of Richard Menary (2007, 2009, 2010a, 2010b, 2013, 2018). Kirchhoff argues that 'cognitive integration' focuses on the manipulation of artefacts and the cognitive norms that govern these manipulations (2014, p. 272). The manipulation of Tetris zoids is what might count as such a manipulation. The rotation is governed by the norm that requires the rotation to stop when the zoid is judged to fit the wall.

Sutton suggests that there is a third wave of extended cognition theories…

> [that] might be a deterritorialised cognitive science which deals with the propagation of deformed and reformatted representations, and which dissolves individuals into peculiar loci of coordination and coalescence among multiple structured media (…). Without assuming distinct inner and outer realms of engrams and exograms, the natural and the artificial, each with its own proprietary characteristics, this third wave would analyse these boundaries as hard-won and fragile developmental and cultural achievements, always open to renegotiation (2010, p. 213).

The third wave therefore brings to the table an essential social and cultural element. In Menary's terms this is *encultured cognition*: "the idea that our cognitive abilities are transformed by a cognitive species of cultural practices […]. What we are able to do is augmented and transformed by the acquisition of cognitive practices" (2012, p. 148). Kirchhoff and Kiverstein list four key tenets of the third wave (2019, p. 16):

(1) Cognitive systems lack fixed properties. I interpret this to mean that cognitive systems are not simply reducible to mechanisms with a fixed structure with properties supervening on a permanent physical and causal substrate (see also Kirchhoff, 2012)

(2) Cognitive systems have flexible and open-ended boundaries,

(3) The assembly of a cognitive system is distributed and not just a matter of being brought about by a single individual agent. Kirchhoff and Kiverstein write that the assembly of the system is distributed across a 'nexus of constraints'.

(4) Cognitive systems are diachronically constituted - meaning that the system is intrinsically temporal and dynamical, possessing processes at different timescales (see also Gallagher, 2018b).

The key to the third wave is that cognition is essentially a social phenomenon in the sense that external scaffolds are invariably public symbol systems or other structures playing a part in a culturally located set of practices, subject to what Menary calls ‘cognitive norms’. Cognitive practices and norms are established through the processes of cultural evolution and cognitive niche construction. The long-term processes by which human beings build tools for cognition and fashion their environment to expedite such systems are complemented by processes whereby neural structures in the human brain become attuned to the new tasks that they need to perform in such a system. There is a story to tell here about cultural evolution, the role of cognitive artefacts and neural plasticity (for more details, see Menary, 2007).

The CSA has more in common with the Second and Third rather than the First Wave of extended cognition. There is an emphasis on the integration of different functions within the system and that one of them is the coordination function. In the examples of stygmergic cognition the coordination function crosses the boundary between the ‘internal’ elements of the system and the environment upon which it is acting. However, the language of many Second Wave theories is often agent-centred, which is something that is not required in the CSA account.

There is a good fit between the CSA and four characteristics of third wave theories listed above. By their nature, coordinated systems are systems of processes that lack fixed properties, and whose properties evolve over time. They possess flexible and open-ended boundaries – this is awkward when dealing with the system demarcation question, but I maintain that it is nonetheless a realistic and positive feature of the account. I have not said much about cognitive assembly – how a cognitive system is put together – but the CSA is not an agent-centred account so cognitive assembly is a distributed business and not one that necessarily depends on the actions of a single agent. Finally, the emphasis on process ensures that a cognitive system is ‘diachronically constituted’. System features do not supervene on things but are properties of processes, possibly emergent ones.

Michael Kirchhoff, in the introduction to his paper on the constitution of extended cognitive systems, sounds close to being a CSA theorist:

> (…) I show that two much more promising explanations by which to ground the ontological claim of [extended cognition] are available, both starting from an exploration of the *coordination* dynamics between environmental resources and neural resources. (Kirchhoff, 2014, p. 258 emphasis added).

The two explanations he has in mind are Second Wave theories and a kind of cognitive integration supplemented by accounts of mechanistic explanation. The CSA is certainly close to the latter.

Despite its affinity to Second and third wave tendencies it does add something. It identifies coordination as the most general and arguably most important functional property of cognitive systems and sets out precise functional conditions for coordination processes that allow them to be identified in the wild. By invoking these principles and making appeals to both explanatory encapsulation of systems and the idea of a system as an emergence base for coordination, coordination becomes a sufficient condition for settling questions regarding the extent of cognitive systems. As far as I know this step has not been taken by other second or third Wave theorists. Another departure from second and third wave accounts is the recognition of the importance and ubiquity of systems whose traces in the world coordinate future actions – stygmergic systems discussed in Section 7.5, as well as adapting functionalism to systems that violate the container metaphor.

8.2.2 Distributed cognition

Now I turn to the relation between the CSA and distributed cognition (D-cog). One might expect there to be similarities between them simply because they both reject the assumption that cognitive systems should be agent centred and emphasise the centrality of coordination. In this section I show that while these similarities are deep it is the differences that matter, particularly the different relation to cognitivism. In this respect the CSA turns out to be more deflationary than traditional D-cog theories, suggesting it is representationalism or computationalism that is doing the work in these accounts not distribution or coordination.

What is distributed cognition? Traditionally it is the idea that cognition is distributed across individuals and material artefacts in the environment. But Wayne Christensen is right in pointing out that distributed cognition is always a relative concept: distributed compared to what? (2007, p. 257). The D in D-cog refers to a theory of cognition that is distributed with respect to loci of cognition in the dominant cognitive paradigm. The D-cog theorists we consider in this section, David Kirsh and Ed Hutchins both take the paradigm to be brain-based cognitivism.

> Our effective environment is a shifting coalition of resources and constraints, some physical, some social, some cultural, some computational (involving internal and external resources), When this shifting coalition of resources is appropriately *coordinated*, the tasks we set out to achieve are accomplished. (Kirsh, 1999, p. 2 emphasis added).

While Kirsh and Hutchins collectively represent an explicit inspiration for the ideas in this thesis, the CSA departs from their work in questioning their paradigm assumptions of computationalism, a more or less fixed system architecture, the container metaphor and well-defined input/output channels (see Hollan *et al.*, 2000; Hutchins, 1995b, 1995a, 2008, 2011, 2014; Kirsh, 1995, 1999, 2005, 2006, 2009). The ecological functionalism espoused by the CSA while able to accommodate these features is somewhat more general because it does not assume them.

Traditional D-cog theories are not without their critics. Christensen argues that traditional theories of D-cog such as those of Kirsh and Hutchins are not good at answering questions such as ‘what determines cognitive ability?’, or ‘how can theory explain the empirical evidence of, for example, the fundamental features of sensorimotor architecture?’, and the theory should be able to explain what it is that is selected for when cognition evolves. While it is easy to find examples of distributed cognitive organisation, it is less clear what the significance of this is. (2007, p. 258).

Christensen’s own solution is to invoke hierarchical control within a model that supposes the container metaphor and a persisting cognitive architecture (1996, 2007). I would like to suggest that the CSA offers another solution that can accommodate hierarchical models but does not presuppose them. As a theory sketch the CSA does not give definitive answers to the empirical questions above but rather points the way to further empirical lines of enquiry. For example, if we are interested in the cognition involved in ant foraging, or classroom clearing, the first question will be to examine how the system becomes sensitive to the relevant aspects of a stygmergic trace as a coordination strategy.

Perhaps the best way to illustrate the difference between the CSA and a traditional D-cog approach is to consider how the CSA casts different light on Edwin Hutchins’ detailed ethnographic case study (1995a) of navigation processes on board the US navy vessel he calls *The Palau*. I shall give a very brief sketch of the study and the main conclusions and draw a contrast with the CSA.

There are two basic problems involved in navigation: establishing one's current position on a chart and deciding which direction and at what speed to proceed in order to get to one's destination.

The first problem is solved in the following manner. On the *Palau* there were two *pelorus* operators stationed on the port and starboard wings of the pilothouse. They take sightings of distant objects on the shoreline through an instrument called an *alidade*. This is a telescope-like instrument linked to the ship's gyrocompass allowing the operator to measure the bearing of the object viewed at the crosshairs. The following (simplified) sequence of actions determines a fix:

1. The senior navigation team selects 3 promising landmarks to use to construct the fix.
2. The recorder requests that the pelorus operators identify the landmarks.
3. At a pre-arranged time, the recorder requests that the pelorus operators take their bearings.
4. The pelorus operators report their bearings to the recorder.
5. The recorder logs the bearings.
6. The plotter draws lines on the chart corresponding to the bearings using a special one-armed protractor called a *hoey*. The resulting triangle on the chart is taken to enclose the position of the ship the triangle itself being like a 3-parameter error bar. The hope is that the triangle is small. (Hutchins points out that the anxiety of the navigator is proportional to the size of the plot triangle!).

On the face of it, there are striking similarities between Hutchins' account and the CSA. Hutchins' work reveals a coordinated modular system. In CSA terms, units like the pelorus operators or the plotter are Task Specific Devices that are coordinated by instructions from the recorder on the bridge. These modules are both producers and consumers of information, which is transformed at each stage of the process. Complex tasks are broken down into a series of time-critical subtasks such as the taking of readings and their communication to the bridge. The physical set-up of the ship means that it is impossible in practice for a single individual to be in possession of all the information gathered by the whole navigation team.

Like the classroom example, the operation of the system is disciplined by a system of strict practices and norms, and these take place within a rigid system of ranks that bestows differential status on individuals. These practices define a definite task structure and the norms place adequacy conditions on their performance.

The system itself, though, is not rigid. It can reconfigure itself to carry out a specific task that is not in the normal run of things. “During an approach to an antenna-calibration buoy near the shore, Chief Richards assigned Smith to the fathometer with the instruction to report when the depth of water under the ship shoaled to less than 20 fathoms” (Hutchins, 1995, pp. 191–192). Hutchins describes this as the social construction of a *daemon* - an agent in computer science that monitors the world waiting for certain specific conditions (such as a certain input from the keyboard). In CSA terms the system has organised itself to produce a new TSD thought of as a condition-action pair that simply behaves in a certain manner when a condition is met by the local circumstances of the ship.

Can we consider the navigation processes of the *Palau* to constitute a coordinated system? What processes play coordination roles regarding the task of navigating the *Palau*? Although the role distribution is complex and mutable, I am inclined to relegate some elements merely to plant roles such as the pelorus operators. Although it is not easy to ascertain they satisfy some of the coordination conditions rather than others. As the navigation task varies, they are asked to sight different sets of landmarks, and they trigger other navigation processes such as the actions of the plotter. So, while they might be task sensitive, it is difficult to see how they track navigation actions. They provide causal input into the core of the navigational system which is the plotter and her chart.

I suspect that other ‘peripheral’ operations in the navigation team are in the same boat (sorry!) and that only the core of the navigation team on the bridge play coordination roles. This does not, of course, rule them out of the system but it does rule them out of the coordination core. More generally, it is hardly surprising in such a strictly modular and hierarchical system as the US navy, that the coordination function is localised. While the computational structure might be widespread across the navigation team, and this was Hutchins’ point, the coordinative function is relatively local and decoupled from other plant functions.

The CSA then comes to a slightly different conclusion regarding the extent at least of the core system to Hutchins, due at least in part to the difference in stance towards computationalism. Through computational eyes, the system fits nicely into the sandwich model in which cognition lies between sensory transducers and behavioural effectors described in 7.2. Pelorus operators are transducers that convert causal links to the environment into representations. These representations are transformed and manipulated and eventually converted back into behaviour in the form of the movement of the ship relative to its environment. The ship is nicely self-contained, separated from its environment, and possesses well-defined inputs and outputs. It is a perfect example to use if one has such a machine-like view of cognitive systems.

Does this mean that computationalism or informational approaches have an edge over the CSA? At a first glance this appears to be the case. The CSA still identifies a core functional unit that is smaller than the whole navigation system. But maybe the questions answered by the two approaches are different. Maybe the computational approach asks: ‘what system does the navigating?’ And the CSA asks: ‘what is responsible for making navigation behaviour goal-directed?’ If distributed *cognition* is what we want to identify, then maybe Hutchins’ system is a bit too big[113]. By this I mean that on a coordinated systems account, what matters *qua* cognition is the part of the system responsible for producing the goal-directedness of the behaviour rather than any other aspect of it. In this sense the CSA is a deflationary account of cognition in comparison, say, with a standard computational account, in which the cognitive part of the system is everything that is sandwiched between inputs and outputs, irrespective if it is responsible for goal-directedness or not.

8.3 Coordinators all the way down

At this point I want to step back and ask whether the CSA is fit for purpose for demarcating cognitive systems. As the previous section suggests, the CSA seems to be addressing a slightly different question than existing approaches. The focus has shifted from the extent of the system to the extent of the coordination process. Is this shift of focus an admission that the original question is unanswerable?

Well, yes and no. The CSA delivers only sufficient conditions for cognitive systemhood. Coordination processes are the heart of the cognitive system. This means that coordination implies systemhood. It has nothing to say about components *not* being part of such a system. But recall the Parity Principle is also only a sufficient condition (see Chapter 1). Properly speaking, the Parity Principle asserts conditions under which an external component can be taken to be cognitive. Although it is sometimes overlooked, it does not have anything to say about when an external component is *not* cognitive. With both, obtaining a negative result leaves the question of whether a process is cognitive open. If the

[113] To be fair to the CSA, computational accounts have their own problems, not least how to understand computations and their relation to representation (Beer, 1998; Churchland and Sejnowski, 1992; Fodor, 2008; Milkowski, 2013, 2017; Piccinini, 2004, 2007, 2015; Piccinini and Scarantino, 2016; Rescorla, 2013).

coordinator condition is to be faulted on these grounds, then so must the original field-defining CC paper.

What are the consequences of the shift of focus from system to coordination process? In the good cases, as we have seen, the demarcation of the coordination process turns out to be demarcation of the system itself. These are systems where the coordination function is an emergent feature of the whole system. Stygmergic systems tend to be of this kind. Insofar as all ants are involved in laying pheromone trails coordination of the foraging task is an emergent feature of the whole system. Likewise, classroom clearing is coordinated by the classroom in interaction with all parents and their tools. In both cases, a coordinative structure emerges through the complex interactions of the individual agents and their traces. There is little or no modularity to be found – meaning little localisation of coordination function inside the system[114]. The kind of system for which the coordination condition works optimally to demarcate the system typically possesses irreducibly complex intertwined feedback processes like this. Ironically, these features are exactly the opposite of the nearly-decomposable systems that are beloved of many standard cognitive science texts such as *The Sciences of the Artificial* of Herbert Simon (1996). The difficult cases for Simon are the nice cases from the point of view of the CSA.

But this works the other way around too. The easy near modular cases for Simon are awkward for the CSA. Consider a ‘Simon’ kind of system that possesses enough modularity and functional localisation so that the coordinator condition will fail to demarcate the whole system and instead demarcate just that part of it responsible for coordination – navigation on board the *Palau* is a good example

[114] I am talking here of coordination with respect to the main system tasks. There may well be localised subcoordination going on, for example, when two parents want to use the vacuum cleaner at the same time.

here. Is this in any sense a failure? I shall argue that it is not. In such a system we can divide system processes into those that realise the coordination function with respect to a set of tasks and those that do not which we have called plant. Plant components do not bear a coordinative relation to the set of tasks T that are the target of the explanation although they might bear non-coordinative relations to them, or they are coordinative with respect to other, possibly related, tasks. If we are interested in how the system *organises* itself to perform T, then plant components are not relevant. They only come into the explanation when we want to understand how the system *performs* T. In such a system, if we take ourselves to be studying the system's self-organisation then plant components are not actually directly relevant to the investigation at hand.[115]

Given the causal or even the functional porosity of systems - the original question about membership of cognitive systems now begins to look naive. It does not really matter how far the system seeps out into the environment – if by system we mean the processes in the world responsible for behaviour. What seems to be of vital importance is the extent of the coordinative core of the system – the part of the system that is responsible for the goal-directedness of its behaviour – the core that gives the system its cognitive properties.

But things might be a bit more nuanced than this. Recall from chapter 5 that a complex system may consist in a (Cummins) cascade of functional units, each being composed of other functions. In the nearly decomposable cases such a functional analysis will be hierarchical and may well correspond to a tidy functional modularity. The locomotive function of a car is decomposable into the

[115] In contrast to his early work, Wayne Christensen (2007) argues that there are advantages from an evolutionary standpoint for a basic modularity and that control processes should be functionally separated from plant processes. System robustness might be enhanced through this separation since the structure of the coordination subsystem itself will be immune to changes in the environment of the system. In these cases, the CSA will only pick out control processes – but these differ depending on the task.

functions of fuel injection, fuel ignition, exhaust removal and so on in a modular fashion. Each function or module faces its own task space and therefore can be analysed in a similar manner. For example, fuel ignition is accomplished through the electrical system via the distributor and spark plugs for example. In this kind of system tasks at a higher level in the hierarchy give rise to a set of tasks at lower levels and so on, each associated with its own coordination processes. This pyramid of coordination processes may give a better indication of the extent of the system than the extent of coordination processes at the highest system level. The driver-car system is better demarcated with respect to basic locomotion behaviour by a chain of coordination processes including the distributor, the carburettor and the camshaft, rather than just the driver’s neural processes.

In the non-decomposable cases there is not a strict hierarchical Cummins pyramid of functions and their modular implementation. Instead, the coordination function is likely to emerge from interactions involving the whole system or a large part of it, again the ants’ nest is a useful example to keep in mind here. The ant pheromone trails coordinate the ant foraging tasks and the trails themselves demarcate the ant foraging system with some accuracy. In the nice loopy systems, it really is coordinators all the way down.

8.4 Coordination and representations

There are two kinds of account of cognitive systems that are prominent in the literature – each in its own way is distinguished by its relation to a key notion – representation. The first is representationalism which we met in Chapter 1 defined by the insistence that production and manipulation of representations is necessary for cognition. The other is enactivism, especially the radical version (see Hutto and Myin, 2013), which takes the view that representations are irrelevant to cognition which is enacted through a dynamical coupling between system and world.

This section will position the CSA in relation to representations in general and the next will examine the specific relation to enactivism. Again, this is important in highlighting the new path taken by the CSA.

This fact that the CSA has arrived at a general mark of the cognitive without appealing to representations is a strength of the account. It can deal with situations such as the babysitting collective in which explicit representations play a role in coordination of the system but also the herding task and Tetris examples where they do not. How sustainable is this representation-neutral position and is the CSA actually non-representational? I shall start with the last question.

On some accounts agnosticism about the role of representation in cognition is simply not a stable position. Hutto and Myin frame it in terms of a stark choice between two opposites. The first is intellectualism or ‘content involving cognition’ (CIC) in their terms, which is the view that, as Fred Dretske puts it, “the manipulation and use of representations is the primary job of the mind” (1997, p. xiv; quoted in Hutto and Myin, 2013, p. 9). The extreme CIC view is that all cognitive processes wherever they are found require contentful representations. The second is radical enactivism that “(…) basic cognition is literally constituted by, and to be understood in terms of, concrete patterns of environmental situated organismic activity, nothing more or less” (2013, p. 11). Where does the CSA lie?

Prima facie the CSA lies more towards the enactivist side of the debate. After all it purposely avoids positing representations as basic to the theory and instead relies on the system’s interaction with its environment to constitute coordination and therefore a minimal kind of cognition. But is there a possibility that hidden in the processes of coordination is an appeal to some variety of representation?

To answer this question let us retrace the functional outlines of the coordination process. Coordination is depicted in abstract functional terms of tracking, triggering, and task-sensitivity fleshed out in certain distinctive kinds of causal process. Recall that processes are coherent spatio-temporal entities that possess temporal parts or stages. These temporal parts or stages, under certain conditions, initiate or trigger other processes in a time critical manner. Tracking requires a stage-sensitive causal connection between the tracked process and the coordination process. Upon reaching a given stage the tracked process causes some change in the coordination process (one could say that it leaves a mark). These two functions are not independent of each other in a coordinator since the states produced by tracking provided the right conditions to activate the triggering connection for some (other) process to be causally activated. Task sensitivity is couched in terms of the right kind of dependence of the coordination process upon the task faced by the system. Were the task to change, *ceteris paribus,* the coordination process would change. Indeed, I suggested in chapter 5 that, for coordination to be explanatory, this dependence relation should exhibit some counterfactual regularity; there should be a pattern in the way the coordinator changes with respect to changes of task.

The question is whether this causal-functional description of coordinators hides an implicit representational description. Certainly, some examples of coordinators possess overtly representational elements. The babysitting book for example includes names of families, dates and times of proposed babysitting turns. These inscriptions possess explicit content by virtue of their usual usage in a social setting. As it happens natural language is used in this example – but this is not necessary for the coordination conditions to be satisfied. Colours could be used to indicate which family requires the turn (and which family will supply it), and shapes could be used to code for days and times. When the babysitting turn is completed, the entry is crossed through.

The coordination function of the Book depends upon the contents of these inscriptions. They have causal power. Without getting into a big discussion here, it seems plausible that the content of these representations consists in the specifics of the babysitting turn that is requested, the agreed sitter, and the cross-through represents the successful completion of a turn. In other words, it seems reasonable to suppose that the symbols represent the actions to be performed by the system, and states of the processes involved in this performance - in line with the task-sensitivity and tracking requirements. Being both a description of the state of the world and a prescription for action makes this a special kind of representation that Ruth Millikan calls, with reference to the Dr Dolittle books, a *pushmi-pullyu* representation (1996).

As always in discussions of this sort we want that our putative representational vehicles have the possibility of misrepresenting. There are three ways in which the babysitting book can misrepresent. It can register that a babysitting turn is completed by a specific family when it is not. For example, that the bookkeeper writes down that Dorothy Bradley sat for the James family, when in fact it was Sylvia Sims. Secondly, it can send a person to babysit for a family that already has a babysitter - so it can trigger an inappropriate action. Thirdly, it can register a required babysitting turn when one is not actually required or, more likely, fail to register a turn that is required. It can misrepresent the tracking of an action, it can misrepresent the actions required to bring about a task, and it can misrepresent the task required to reach the distal goal. Misrepresentation can happen with regards to each of the tracking, triggering and task-sensitivity functions of the coordinator. It can misrepresent for the same reason that it can represent - because of the social norms governing babysitting. These are norms such as: families require a babysitter for children under a certain age when the parents go out, the babysitter needs to be fixed in advance, normally only one babysitter is required, and so on. The normative framework governing the performance of babysitting tasks serves to fix the content of the babysitting book representations. Since I have not appealed to anything beyond the coordinator functions here, there is a possibility that this argument can be generalised to non-representational coordinators.

This is all a bit quick so let us take things apart a bit more carefully. First, I want to say a bit more about pushmi-pullyu representations since the general form of Millikan's argument (but not its teleosemantic foundations) is one I want to follow in the rest of the section. A pushmi-pullyu representation is one in which both directive and descriptive aspects are irreducibly integrated. "Richard, you have not eaten your peas", that dread admonishment from the Junior School teacher in the canteen, was both a description - I had not eaten my peas (for reasons that seemed perfectly understandable) - coupled with an injunction to jolly well eat them. What is special about pushmi-pullyu representations is that they are more primitive than the conjunction of purely descriptive representations with purely prescriptive representations. As Millikan points out: "(p)urely descriptive representations must be combined with directive representations through a process of practical inference in order to be used by the cognitive system" (1996, p. 145). However, pushmi-pullyu representations do not require any inferential mediation to do their job. They are functionally integrated and evolutionarily more primitive than, say, human beliefs and desires.

On the content question (again very briefly - the full theory needs more space to do it justice) let us think of "Richard you have not eaten your peas" as being an integrated and irreducible 'conjunction' of "There are uneaten peas" and "Eat your peas". The directive aspect "eat your peas" has the function to guide the system to its conditions of satisfaction - so that would be a pea-full Richard. The descriptive part of the representation, "There are uneaten peas" played a role in ensuring that the directive part did its job - it has satisfaction conditions that are relevant to the triggering of the injunction "eat your peas!". (Note the way in which these are semantically integrated - there are no separate satisfaction conditions for the description that are independent of the directive). The content, then, of a pushmi-pullyu representation resides in its relation to the task and the function directed to accomplish it: ‘peas on the plate ought to be eaten’[116]. All one needs for this argument is that the functions that we are dealing with here are indeed normative - as I said in Chapter 5 one can be agnostic about the exact route we take to get there[117].

The importance of pushmi-pullyu representations for coordination is illustrated by Millikan’s invocation of one of her favourite examples. “The bee dance tells at once where the nectar is and where to go. Functioning properly, it produces variation in behaviour as a direct function of variation in the environment” (1996, p. 151). Millikan explicitly links pushmi-pullyu representations to social norms and emphasises their role in social coordination. She describes them as a “capacity and disposition to understand social norms in a way that is undifferentiated between descriptive and directive” (1996, p. 154). The irreducibility of the descriptive and prescriptive functions of a pushmi-pullyu representation closely parallels our insistence that the triggering and tracking functions of a coordinator are not separable. However, for Millikan these representations are all internal, while the CSA makes no such commitment.

Now I hinted that, *prima facie,* there seems no reason why this analysis should not carry over to nominally non-representational coordinators. Let us take the example of clearing the classroom. Starting with the representational version where Frida draws a map of what needs to be done and updates it when tasks are completed, it should be clear that this is similar to the babysitting case. The map is analogous to the babysitting book and Frida to the bookkeeper. We can suppose in the example, that prior to the cleaning action in the story, there is an

[116] It is important that the integrated descriptive-imperative content of PMPY representations is not read as a conditional ‘if there are peas then they should be eaten’.

[117] If pushed, I am sympathetic to some version of pragmatism here such as Hutto and Satne (2015) and Gallagher and Miyahara (2012).

earlier stage where Frida surveys the scene and constructs her map of the classroom and its associated tasks. She does this by making a comparison between messy reality and an encultured sense of what counts as clean and tidy. She then applies knowledge of cultural practices around cleaning to decide on the sequence of cleaning tasks needed to transform a messy classroom to a clean classroom. In putting together this sequence she also draws upon social norms regarding the constraints that some activities place on others - for example, vacuuming the floor presumably takes place before mopping it.

Like the babysitting example, social norms play an important role in Frida's map-making. The primary role they play is the transformation of a state of affairs in the world into a set of tasks on her map. These are tasks that are themselves heavily dependent upon cultural convention and social norms and practices. Deviation from a social norm, in this case tidiness, generates corrective tasks to restore the situation to the norm. The map-making activity combines a state of the world and a set of social norms to produce a set of tasks. This is what gives the map its distinctive descriptive and prescriptive representational functions. The map deals in pushmi-pullyu representations.

The argument is that the classroom in relation to the right set of social norms performs the same function as the map and therefore also deals in pushmi-pullyu representations. Recall the original story in which the parents distributed the tasks amongst themselves without the need for a Frida and her map. Analogous to the case of the map, social norms serve to endow aspects of the classroom itself with pushmi-pullyu representational function. Instead of 'Richard you have not eaten your peas' it is 'Danielle you have not cleared up those crumbs'. The state of the classroom, read by the parents, relative to the relevant norms serves exactly the same purpose. The relevant norms are instilled through enculturation. A well brought-up parent in Holland knows what counts as tidy and clean (let us not forget that the same Dutch word *schoon* means both 'clean' and 'beautiful'!). The previous chapter brought up Rodney Brooks' dictum that the world is its best representation and now we understand what kind of representation this is.

If the classroom is a representation, even if minimal, it should be possible for it to misrepresent. Brooks' robots interpreted the presence of a coke can as a pushmi-pullyu representation coordinating a tidying up activity in much the same way as the parents interpret cake crumbs. In Brooks' case the robots needed to be programmed for the can to take on this representational significance while in the classroom case it is society that 'programs' the parents to recognise the cake crumbs as representing the relevant cleaning task through the establishment of the appropriate normative framework. For misrepresentation to occur the coke can would not end up triggering the performance because the system did not recognise it as representing a task. The same would be the case in the classroom situation. For example, a teenager encountering the crumbs would not, I take it, see it as representing the task of applying the vacuum cleaner, but rather the task of ignoring the crumbs or kicking them under the cupboard out of sight. To use the language of stygmergy, the teenager lacks the requisite trace-sensitivity to trigger the required task.

In 7.5 I was careful to write that the trace *corresponded* to the task to avoid prejudging the representation debate, but now I propose that it is legitimate to say that the salient aspect of the trace *represents* the task. The tropistic cases are cases of misrepresentation; either because the task lies outside the task space of the system in the case of the *sphex* and the antennae-less cricket or because the one-one relation between the trace aspect and the task had been broken in the case of the *Abispa* facing a funnel with a hole, or a *sphex* presented with a displaced grasshopper for the fortieth time.

If this is correct, then coordinators do naturally possess minimal representational tendencies - expressed in terms of Millkan's pushmi-pullyu representations. These are not fully-fledged representations, if by 'fully-fledged' we mean that descriptive and imperative conditions of satisfaction are independent, for the simple reason that they are not independent - the descriptive part has content only in virtue of its triggering function for the directive part. A fully-fledged representation such as ‘the ball is on the grass’ has truth conditions that are independent of what you were going to do with the ball. In coordinator terms the content of the tracker is not independent of its role in subsequent triggerings.

The CSA then does not presuppose that cognition depends on manipulation of fully-fledged representations. But the coordination conditions, functional though they are, introduce a minimal kind of representation. Given that nowhere was representation explicitly brought into the account, the way might be open here for a functional theory of minimal representation based on the coordination function. Indeed this might ground something like Hutto and Satne’s (2015) theory of content based on minimal kinds of representation[118].

8.5 The CSA and affordances

Having discussed the role of representations in the CSA we return to a question that has been nagging away in the background since the discussion of Tolman and the enabling role of environmental constraints in Chapter 4, and sensitivity to aspects of stygmergic traces in Chapter 7. Where do affordances enter the picture and what is the relation of CSA to radical enactivism? How close are they and does the CSA add anything to an enactivist approach?

Very briefly, the key idea in enactivism is that cognition is constituted through a dynamic coupling between the organism and the environment. This insight

[118] This is work to be done in the future. The strategy would be to argue that in the past minds were without content, but that in order to cooperate in social tasks material features in the environment took on coordination roles. By doing so, and by being policed by social norms, they would become pushmi-pullyu representations. A story like Hutto and Satne’s would then be told about how fully-fledged representational content can emerge from these coordination processes.

connects enactivism to the older tradition of pragmatist philosophy, as Kiverstein notes in his citing Dewey:

> To see the organism in nature, the nervous system in the organism, the brain in the nervous system, the cortex in the brain is the answer to the problems which haunt philosophy. And when seen thus they will be seen to be in, not as marbles are in a box, but as events are in history, in a moving, growing never finished process (Dewey, 1958, p. 259; quoted in Kiverstein, 2018, p. 32).

Dewey's thought fits with the ideas in this thesis in Chapter 3 about the nature of systems - that systems are embedded in other systems and that boundaries multiply, almost, without limit. There's a Deweyan flavour to the CSA's insistence that systems are primarily distinguished by the functional role they play in larger systems. In an enlightening article on Dewey's theory of mind, Mark Johnson writes: "Dewey argues that the basic unit of experience is an integrated dynamic whole that emerges through the coordination of an active organism and its complex environment" (Johnson, 2010, p. 124). He goes on to explain that Dewey thinks of the environment in both physical and social terms and that the organism-environment system is in some sense a non-dissociable whole.

Enactive cognitive science takes this irreducible dynamical coupling as a starting point. It does not make sense to talk of the organism side of the coupling without the environment side. The two are dynamically interdependent. Kiverstein writes: "(s)uch is the degree of continuous, integrated, and coordinated mutual influence between the two systems that we can't solve the equations describing the behaviour of each system separately" (Kiverstein, 2018, p. 33). He quotes Clark to support this point:

> As a result [of the interactions between internal and external resources being highly complex and non-linear] there may, in some cases, be no viable means of understanding the behaviour and potential of the extended cognitive ensembles by piecemeal decomposition and additive reassembly" (Clark, 2011a, p. 116; quoted in Kiverstein, 2018, p. 34).

This is exactly the point made in Chapter 2 about the failure of the localisation and decomposition procedure in complex systems. The depth of the parallels here is striking and one would be forgiven for concluding that the CSA is simply a version of enactivism. But let us be a little more cautious. Dealing with this question is not easy, not least because enactivism is an umbrella term for a variety of different positions. In this section I shall concentrate on the strand that seems closest to the CSA: radical enactivism.

Radical enactivism is the view that basic cognition is non-representational - that is it does not have representational content. Radical enactivists do not deny that contentful cognition is possible. Rather they see it as an achievement that requires scaffolding by structured manipulation of environmental structures (within a social setting). Radical enactive cognition (REC) become influential in 4E cognition through the work of amongst others Tony Chemero (Chemero, 2009; Chemero and Silberstein, 2008), Dan Hutto and Erik Myin (Hutto *et al.*, 2014; Hutto and Myin, 2013, 2018), Glenda Satne (Hutto and Satne, 2015, 2018, 2018). Julian Kiverstein and Erik Rietveld (Bruineberg and Rietveld, 2014; Kiverstein, 2012, 2015, 2017, 2018, 2020b; Kiverstein and Farina, 2011; Kiverstein and Rietveld, 2018, 2020; Rietveld *et al.*, 2018; Rietveld and Kiverstein, 2014).

Radical enactivism shares with the CSA four main ideas: that cognition and action are essentially interlinked and not easily distinguished, that representations are not required as a premise in arguing about cognition, that environmental constraints are potentially enabling of action, and that organisms engage in *sense-making* which is the development of sensitivity to certain aspects of the environment. Hutto and Myin express the first idea in the following way: “[The radical enactive] credo “we act before we think” – is an outright denial of the [content involving cognition] thesis that “we think in order to act” (2013, p. 12). If we identify thinking with the operation of a cognitive system, then for the CSA actions are at least partially constitutive of thoughts. Regarding representations

Hutto and Myin write: “Enactivists are concerned to defend the view that our most elementary ways of engaging with the world and others (…) are mindful in the sense of being phenomenally charged and intentionally directed, despite being non-representational and content free” (2013, p. 13). They are keen to point out that this is not to say that it is never appropriate to speak of representations and that language users are capable of genuinely contentful modes of thinking and reasoning. The CSA takes something of the same view but with a caveat. We saw in the previous section that coordination involves a kind of minimal content in terms of pushmi-pullyu representations. This is not a fully-fledged content in terms of representations that have truth conditions. But the CSA, while invoking goals and tasks, does not presume that these are represented anywhere. As an outline theory, neither does it have anything to say about phenomenology of cognition although this may be something to develop in the future. The role of environmental constraints is similar in both theories, and I shall say more about this in terms of affordances in a moment. Finally, regarding sense-making, the sensitivity of an organism to aspects of its environment (see for example Colombetti, 2010), is parallel to the idea that, in stygmergic coordination, parts of the system become sensitive to specific aspects of the trace of its behaviour in the environment.

Enactivism, so conceived, implicitly, or explicitly invokes coordination. Kiverstein and Rietveld, for example, describe cognition “in terms of temporally extended activities in which the agent skilfully *coordinates* to a richly structured landscape of affordances” (2018, p. 149). However, I would argue that the CSA is alone in both making this aspect central and exploring its functional ramifications. While the two views are not identical, they are close enough to warrant further investigation.

I suggest that there are three points of difference here:

(1) Enactivism tends toward an agent-centred approach despite being willing to embrace some versions of cognitive extension. The CSA, as we have seen, is not predicated on a central agent.

(2) Affordances play a central role in the kinds of enactivism considered here, but the CSA is not committed to a Gibsonian view, although it is sympathetic to many aspects of it.

(3) While both approaches emphasise the importance of norms, the CSA deals specifically with tasks related to distal goals while enactivists explicitly rule out a goal-directed account of behaviour.

I shall expand on these differences in turn.

At first blush, enactivism seems to be agent centred – a source being its roots in 1970’s accounts of autopoiesis – the requirement that an organism is ‘organisationally closed’ in order to maintain itself as an open system in far-from-thermodynamic equilibrium (Di Paolo, 2018, p. 78). Organisational closure marks out the organism as significant in the production ‘meaning’ in terms of a system’s interaction with the environment. The CSA on the other hand does not posit a central agent but can embrace a distributed system comprising many individuals. The reader might object that the CSA draws on similar considerations as autopoietic enactivism in grounding the normative framework in which notions such as tasks make sense. This is true but the CSA is careful to avoid being specific about the origins of the normative framework and is careful to avoid assuming that the system that possesses these distal goals must have the same boundary as the cognitive system which is a central assumption of autopoietic enactivism. It *could* ultimately be based on some kind of autopoiesis, but it is completely conceivable that it the normative framework is an emergent property of social interactions and therefore does not primarily point to a fundamental agent-centricity[119]. In this respect enactivists might reply that there are distinctly social forms of enactivism such as that espoused by, amongst others, de Jaegher and di Paolo (De Jaegher *et al.*, 2010; De Jaegher and Di Paolo, 2007; Di Paolo and De Jaegher, 2017; Fuchs and De Jaegher, 2009). This difference may not then be as substantial as it first appears.

The second difference concerns the role of affordances taken to be environmental solicitations for action. There is some disagreement in the literature about whether affordances are dispositions to act or relations between the environment and the organism (see Kiverstein, 2020a for a convincing argument for the latter). The CSA also recognises, following Tolman, that the environment offers constraints that are conducive to task-performance - enabling constraints - that depend on the abilities of the system, and play a coordinative role in its future performances. Whether an environmental structure is part of coordinating system performance depends on the tasks which the system faces.

It is here concerning the role of tasks and goals in the account where the biggest difference is to be found. Should we take tasks to be equivalent to affordances and what then is the role played by norms and goals?[120]. While both accounts regard norms as being important, the central idea of task in the CSA places them centre-stage and suggests that environmental constraints help to constitute tasks only in relation to goals. Superficially they are similar, but the key difference is the relation to goals and tasks.

[119] In this connection it is helpful to think about how organisation of social systems requires highly ordered environments that are far from thermodynamic equilibrium.

[120] Julian Kiverstein suggested that they were equivalent in his talk *Skilled we-intentionality: situating joint action in the living environment* at the online workshop on the Philosophy of Coordination organised by Ric Sims and Marc Slors in connection with *Egenis: the Centre for the Study of Life Sciences* at the University of Exeter and the *Mind and Cognition group* at Radboud University Jan 2021.

This is difference between the two accounts is worth developing. For enactivist positions such as the *skilled intentionality framework* (SIF) of Rietveld and Kiverstein (Kiverstein and Rietveld, 2015, 2021; Rietveld *et al.*, 2018), affordances are normative because they are linked to the "individual's ability to distinguish correct from incorrect, better from worse, optimal from suboptimal, or adequate from inadequate activities in a specific concrete material setting" (Rietveld and Kiverstein, 2014, p. 332). Rietveld calls this *situated normativity* "because it is the concrete situation, broadly understood, that makes an individual's activity adequate or not" (Rietveld, 2008; Rietveld and Kiverstein, 2014, p. 332). But Rietveld and Kiverstein explicitly reject that 'concrete situations' come by their situated normativity through goals and tasks even though this would explain, as the CSA does, how performances can have success or adequacy conditions related to tasks. Bereft of goals and the tasks associated with environmental constraints, it is difficult to see how affordances can have normative force, or even how, in a proliferation of affordances, the relevant ones are picked out. Without the goal of the cheese at the centre there is no correct, satisfactory, or adequate way to run a maze. It is difficult to see how affordances could have success conditions on their own.

Rietveld and Kiverstein are adamant in their rejection of goals and tasks. They write: "(w)ithin the skilled intentionality framework we are careful not to presuppose goals, tasks, or aims of some mysterious origin as the source of relevance, but instead see the emergence of the soliciting character of affordances as the result of a process of self-organisation" (Rietveld *et al.*, 2018, p. 52). Nonetheless, they happily refer to the agent's 'situation', 'concerns' and even 'urges' (Rietveld and Kiverstein, 2014, p. 340 et seq).

The problem is that the notion of goals and tasks used in this thesis *is* the result of a process of self-organisation - there is not a contradiction here. In fact, it is rather strange to say that organisms self-organise and do *not* have tasks and goals, yet the environmental constraints *do* mysteriously possess normative force such that organism performances can be evaluated as better vs worse, correct vs incorrect, adequate vs inadequate and so on. If there is a question of mysterious origin it is in the normative force of environmental affordances according to the SIF.

Why do Rietveld and Kiverstein take this line on tasks, when they seem to offer a way of bridging the gap between self-maintenance and the normative requirement that the organism acquire an ‘optimum grip’ on affordances that Kiverstein refers to in recent work[121]. One possibility for this distrust of tasks might arise from assuming that the notion of a goal or a task is both represented somewhere in the system, and that it should exist prior to any relevant system behaviour.

But these are *not* requirements in our account. As we have seen, the notion of task, as we discuss it in the CSA, is not necessarily represented in the system and *is* something that can emerge through (behavioural) interaction with the environment. The task of clearing up the spilt water on the classroom floor only becomes manifest when Danielle spills her bucket trying to avoid Anne’s vacuum cleaner. At this point a task is generated - mid-activity as it were. Tasks are like this in many situations. They are not apparent beforehand and can come into existence on the fly, again as we have seen from stygmergic cases.

Furthermore, it seems that an affordance account does need something to play a role similar to a task. For example, the CSA describes how social practices and habits produce a task out of a messy classroom. For enactivists the classroom possesses affordances for clearing up, but they also possess affordances for

[121] See, for example, his talk Dissolving the Causation-Constitution Fallacy: Diachronic Constitution and the Metaphysics of 4E Cognition at the online workshop on Philosophy of Situated Cognition organised by Mark-Oliver Casper and Giuseppe Flavio Artese at University of Kassel Feb 2021.

being made untidier. Without a goal and its canalisation into tasks it is difficult to see why one set of affordances *ought* to be exploited rather than the other or why a climber is not cast hither and thither by the affordances she encounters. Hence it is difficult to see how an enactive account of this flavour has enough normative heft to explain intelligent, that is goal-directed, behaviour and therefore get a grip on cognition itself.

Both accounts have their merits – they are similar and intersect but emphasise different features. Enactivism of the kind discussed here emphasises affordances and phenomenology, while the CSA emphasises system organisation and tasks. Notwithstanding their deep affinity, there is clear blue water between them.

8.6 Concluding remarks

This chapter attempted to locate the CSA in relation to existing theoretical work. Broadly speaking it fits into the framework of Third Wave extended cognition, indeed, it ticks all Kirchhoff and Kiverstein's boxes for Third Wave theories. In not assuming agent-centredness, it is distinct from some Second Wave theories and shares much in common with distributed cognition. The distributed cognition of Kirsh and Hutchins is couched explicitly in computational terms, which is also a move that is not made by the CSA. It can accommodate such theories, but they are not built in at the start.

In this sense the CSA relates to existing accounts but makes a distinctive contribution. It makes a virtue out of functional abstraction and basic functional distinctions that fends off the coupling-constitution and sterile effects worries. It is strong enough to provide a sufficient condition for the cognitive, but not refined enough to individuate cognitive kinds. This is perhaps as it should be because there is a strong sense that cognitive systems are just too diverse to be captured by anything but the most basic functional characterisation. The job in this thesis was to demarcate the cognitive not to produce a refined taxonomy.

The question of representations is an interesting one - since the CSA does not bake them in at the start but in a minimal sense they emerge as a way of understanding stygmergic cognition. This is left largely unexplored in the thesis and coordination functions are unpacked in terms of a general description of their functional roles rather than a specific description of where representations play an active role in coordination. This is something that can be taken further.

Chapter 9

Conclusion: Coordination and Beyond

> Theories of cognition should be able to provide the operational conceptual categories with which to describe their objects of study and distinguish them from those outside their remit. They should be able to say in concrete terms what sort of system, event, or phenomenon counts as cognitive and in which cases it does not. (Di Paolo, 2018, p. 75)

The thesis started by posing the question what criteria can be used to draw a boundary around such a cognitive system and thereby provide support for a version of the HEC? In the course of the thesis, new machinery has been proposed, the Coordinated System Approach, to tackle this question. In the introduction I wrote that the test of whether this project has been successful is whether the CSA can resolve certain sticking points in the debate that were set out in part I of the thesis.

How successful is the CSA? It is a deliberately parsimonious attempt, starting from relatively modest premises, to capture the functional essence of cognition by focussing on the time-critical organisation of system processes. In this respect then there are only certain questions that it can answer. It has nothing to say about how systems do complex problem solving or how they select tasks or reconfigure themselves to tackle novel problems. Readers expecting to find out more about what underlies our standard mental vocabulary likewise will be disappointed. What it does tell us is what kind of basic functionality the core of the system possesses – the set of processes that coordinate the system in exercising these capacities. The lack of attention to specific cognitive capacities is the price paid for a basic mark of the cognitive that has the broadest reach. The coordination conditions express the most basic kind of cognitive functionality that a system must possess to deliver intelligent, that is goal-directed, behaviour. It therefore satisfies the first part of di Paolo's requirement above but perhaps not the second.

This said, then, the CSA provides conditions for identifying this basic core of cognitive systems. It can be used to assert that a process, wherever it occurs, is part of the core of the system responsible for performing a set of tasks. It is the part of the system that delivers the goal-directedness. This means that unless the whole system is responsible for these coordination functions then it does not demarcate the system.

Is this a failure of the project? No, I don't think so – for four main reasons. Firstly, the systems that are most amenable to a CSA approach are those which resist traditional localisation and decomposition methods such as those of the new mechanisms literature. In this sense then the CSA could be seen to complement these methods. Moreover, many of the systems of interest are such that the whole system supports the emergence of coordination processes, so identifying these processes also serves to identify the system which was the original aim. Secondly, the CSA supplies a sufficient condition for a process to be part of a cognitive system responsible for the performance of a set of tasks which is enough to establish HEC. It is enough to show that an 'external' process satisfies the coordination conditions with respect to a set of tasks to assert that the process belongs to the cognitive system hence assert the HEC. The CSA has been applied to standard examples in the literature to show how the HEC might be thus supported. It has also been applied to some new examples that raise interesting new questions. Thirdly, in the light of the CSA the original question about the extent of cognitive systems seems to be badly posed. If cognitive functions are divided into those responsible for coordination – that is those functions that supply the system's goal-directedness – and those that support these coordination functions but do not themselves satisfy the coordination conditions, then it strikes me that the interesting part of the system is the coordination part. 'Where is the coordination happening?' is a better first question to ask about the system simply because support functions may well end up having fuzzy boundaries. Lastly, whether readers are convinced by CSA arguments for HEC or not, this work limns the central role of coordination in the production of intelligent behaviour and hopefully highlights the importance of coordination dynamics in the research agenda to move the debate on.

There are questions that I could only touch upon in this thesis. The approach taken here relies heavily on the assumption that tasks are infused by norms. It assumes that distal system goals supply tasks with their normative force. A naturalistic account of normativity is therefore presumed by the CSA – which effectively argues from *is* to *ought*. This is not the only account that requires this move, it is a problem that crops up in teleosemantic theories, in interactionist theories, in attempts to ground semantic accounts of computation, in attempts to naturalise intentionality generally, as well as theories of agency in the philosophy of biology and elsewhere. Even if this problem is solved, there is still the question of how a basic kind of biological normativity relates, if at all, to normativity in the social world required by application of the CSA to social cognition. These questions though of vital interest are each a thesis in themselves and I have put them to one side.

Another interesting question is a problem raised by Larry Shapiro as to how effective a functional description can be in picking out causal realising mechanisms (2008). Where does the pumping function of the heart stop given that it is only has this function in relation to the whole circulation system. This is perhaps an adjunct to the grain parameter argument discussed in chapter 1. Without having worked out the details, I think I would want to emphasise the emergent or holistic basis for function while at the same time holding on to the idea that functions are implemented locally. The heart has a function because of its functional relationship to other processes in the circulation system (and indeed the whole body) yet there is an identifiable realiser of this function. I suspect that a process account is less vulnerable to this objection because of the close relationship between process and function. In any case the CSA would not be the only theory affected by this objection. Shapiro suggests that the HEC is particularly vulnerable, but I do not see that internalist theories are any better off. If the idea of location of realisers of function is under threat, then so is any theory that makes statements about location of realisers whether internal or external[122].

Much has been said about the fact that the strength of the CSA derives from its generality. But while coordination may be a sufficient condition for cognition it does not tell us much about other functions of the systems that implement it. For example, while a task is a transformation of the current world state or process to one of a category of desirable states or processes, nothing has been said how a system sorts world states into categories – surely a crucial operation within a cognitive system (see Pattee, 1979). Christensen, for example, identifies environmental categorisation as being a necessary process in the production of goal-directed behaviour (1996, p. 314). This suggests links with the concept of *meaning-making* of the enactivists[123]. Likewise, I have not said anything about how systems might make strategic choices from their behavioural repertoire, or how they can achieve a situation where environmental constraints are smoothly integrated into the behavioural repertoire via learning (Tolman's *docility*). These are not just conceptual matters but are empirical ones and lie outside the scope of this thesis; they are important lines of future inquiry.

This leads to the question where this work takes us in the future. Particularly interesting is the suggestion that a CSA perspective applied to situations involving

[122] A programme for addressing this question could start holistically in identifying functions relative to all the other functions in the system. Then go local and look for local processes that realise these functions.

[123] It might be productive to explore further the relation between the CSA and enactivism, for example, the SIF of Rietveld and Kiverstein (Kiverstein and Rietveld, 2020; Rietveld and Kiverstein, 2014). Perhaps the role of the CSA here is to show that neither goals or representations are things to fear and can be theoretically productive.

basic social action coordination might provide the basis for a pragmatic account of basic content. By this I mean that a social group of ‘contentless’ individuals may bootstrap the emergence of individual representational content by using external structures as part of coordination processes. Clark drops a broad hint along these lines: “(…) [L]anguage works its magic not (or not solely) by means of translation into appropriate expressions of neuralese or the language of thought but also by something more like coordination dynamics” (2011a, p. 53). In the terms of this thesis, traces in social stygmergic coordination have the potential to perform the transformation from pushmi-pullyu representation to fully-fledged representation. Here is a just-so story to illustrate what I mean. Keeping track of the hunting contributions of each member of the tribe may be important to root out free riders. Bringing the actual kill into the cave is a trace that coordinates hunting processes. At a certain point the group realises that drawing a pictogram on the wall is easier than dragging a whole deer into the cave. The pictogram becomes the relevant trace aspect and as such becomes a pushmi-pullyu representation (see Section 8.4). But the group are acutely aware that a new kind of free riding is now possible: a hunter may inscribe the pictogram without having produced a deer. It is then necessary for someone to check. But this act shows that the pictogram is now a fully-fledged representation equipped with truth conditions. The content has been transformed from the Neolithic version of the indivisible conjunction of description and proscription of “you have not eaten your peas” - perhaps “you have not provided a deer” – to “there is a deer outside the cave”. The deer pictogram is only ‘true’ in the case that there is a corresponding deer in the group’s larder. The story hints how a CSA analysis could provide an argument to the emergence of fully-fledged representation – a boot-strap move consistent with the ideas of Dan Hutto and Glenda Satne (2015, 2017).

The use of the CSA to study stygmergic systems is a possible future research project. Given that these systems are so widespread, from swarm robotics, through industrial systems design, economic systems, biological systems, cultural and political systems, conversation analysis, to computation theory, there is much scope for research here with practical consequences. One way in which stygmergic systems may come in handy is to help understand the ontogenesis of coordination processes. What characteristics does a complex system need to possess in order that coordination emerges? Putting this question more suggestively: what incentives are there for autonomous processes, say, to enter into coordinative interaction?

Indeed, coordination can be achieved without cooperation. Accounts of adversarial stygmergy show how coordination may develop between processes that are individually competitive (see for example Nieto-Gomez, 2016). It would be interesting to apply these ideas to social cognition. The CSA suggests that not only are shared goals not required for the elements of social cognitive systems but the meta-goal of seeking cooperation may not be necessary either.

Plant cognition is another area in which the CSA might have something to contribute, not least because it can handle situations where there is basic or minimal cognition. Paco Calvo and his lab have stimulated interest and controversy in about equal measure through his work on associative learning in plants and his ‘manifesto for plant neurobiology’ (Calvo, 2016, 2017; Calvo *et al.*, 2016, 2017, 2020; Calvo and Baluska, 2015; Calvo and Friston, 2017; Calvo and Trewavas, 2020; Frazier *et al.*, 2020; Hiernaux, 2021a, 2021b; Mediano *et al.*, 2021; Raja *et al.*, 2020; Segundo-Ortin and Calvo, 2019; Trewavas *et al.*, 2020). This is an extremely fertile area for the coordination approach precisely because it does not rely on functional similarity with animals. Parity with animal cognition has been pointed out as a source of zoocentrism in the plant cognition literature (see Yilmaz and Dupre, forthcoming).

A large class of stygmergic systems that was one of the original motivations for this thesis can be found in scientific investigations centred on the use of models.

The model is a result of previous scientific work and in many cases coordinates future investigations. The CSA may well support Ronald Giere's claim that groups of scientists form distributed cognitive systems (Giere, 2002a, 2002b, 2010, 2013; Knorr-Cetina, 1999; Poirier and Chicoisne, 2006; Vaesen, 2011). If a scientific model were shown to satisfy coordination conditions with respect to a set of scientific tasks then this view would be supported and might suggest a rapprochement between cognitive and social views of science (see Magnus, 2007; Magnus and McClamrock, 2015; Toon, 2014a, 2014b, 2015). In particular the work on symbiotic cognition, trading zones and boundary objects suggests a link between cognitive science and philosophy of science or STS (Galison, 1997; Gallagher and Crisafi, 2009; Slors, 2019; Star, 2010; Star and Griesemer, 1989, 1989). This is a particularly interesting project and, although it became too big for a single PhD, might be something that is worth taking further in the future.

In what way then does this work contribute to our understanding of extended cognitive systems? It points out the centrality of coordination and characterises this in terms of functional organisation of systems. By doing so it provides a criterion that is broad enough to cover a wide range of actual and possible cases while at the same time being specific enough to address the question whether a given set of processes is a coordinator for a given task set. In doing so it brings together insights from many different perspectives on the debate; from principal protagonists CC, AA and Rob Rupert, from the Third Wave theorists, from computationalists, from dynamical systems theorists and enactivists. It has also integrated aspects of neo-cybernetics and complex system theorists with early work by Tolman and Gibson. Moreover, it clarifies why the debate has reached a stalemate and produced a recipe for moving it forward partly by changing the kind of question that should be asked, partly by using a different machinery and partly by putting to one side prior commitments (such as (anti-) representationalism).

More speculatively, the work on pushmi-pullyu representations suggests a move away from a representationalist belief-desire psychology towards a pushmi-pullyu psychology in which beliefs and desires are intimately integrated and not separable, and, foundationally only minimally representational. The work done by folk-psychology in accounting for interpersonal interaction in the performance of joint tasks could be accounted for through the CSA by positing coordination processes for the joint task, for example, through stygmergy, without requiring that individuals in the system attribute beliefs and desires to their collaborators.

This project also contributes in potentially calving a diverse range of future research projects in disparate but related fields: philosophy of science – models as part of coordination processes in distributed cognitive systems, plant biology – coordinative processes in plants as a key to plant cognition, social cognition – investigating cognitive ecologies and symbiotic cognition, theories of content – social coordination as a key to getting a pragmatic theory of the content of mental representations off the ground.

Finally does the CSA support the HEC? The discussion in this thesis suggests that some of the standard examples do support an HEC interpretation. Moreover, there are many new examples, especially those based on stygmergic systems, that show that extended cognitive systems are widespread. That is a positive note on which to end.

References:

Abramova, E., and Slors, M. V. P. (2015) Social cognition in simple action coordination: A case for direct perception. *Consciousness and Cognition* 36: 519–531.

Abramova, E., and Slors, M. V. P. (2019) Mechanistic explanation for enactive sociality. *Phenomenology and the Cognitive Sciences* 18: 401–424.

Adams, D. (1996) *The ultimate hitchhiker's guide*. New York: Wings Books.

Adams, F. (2018) Cognition wars. *Studies in History and Philosophy of Science* 68: 20–30.

Adams, F., and Aizawa, K. (2001) The bounds of cognition. *Philosophical Psychology* 14(1): 43–64.

Adams, F., and Aizawa, K. (2008) *The bounds of cognition*. Oxford: Bleckwell.

Adams, F., and Aizawa, K. (2009) Why the mind is still in the head. In Robbins, P. and Aydede, M. (Eds.), *The Cambridge Handbook of Situated Cognition*. Cambridge University Press.

Adams, F., and Aizawa, K. (2010a) The value of cognitivism in thinking about extended cognition. *Phenomenology and the Cognitive Sciences* 9(4): 579–603.

Adams, F., and Aizawa, K. (2010b) Defending the bounds of cognition. In Menary, R. (Ed.), *The Extended Mind*. Cambridge, Mass.: M.I.T.Press.

Aizawa, K., and Adams, F. (2005) Defending non-derived content. *Philosophical Psychology* 18(6): 661–669.

Anderson, M. L. (2015) Beyond componential constitution in the brain: starburst amacrine cells and enabling constraints. *Open MIND* 1: 1–13.

Anderson, M. L., and Chemero, A. (2009) Affordances and Intentionality: Reply to Roberts. *Journal of Mind and Behavior* 30(4): 301–312.

Anderson, M. L., and Chemero, A. (2019) The world well gained: on the epistemic implications of ecological information. In Colombo, M., Irvine, E., and Stapleton, M. (Eds.), *Andy Clark and His Critics*. Oxford: Oxford University Press.

Anderson, P. W. (1972) More Is Different - Broken Symmetry and Nature of Hierarchical Structure. *Science* 177(4047): 393–396.

Anscombe, E. (1981) Causality and determination. In *Metaphysics and the Philosophy of Mind, the Collected Philosophical Papers of G.E.M. Anscombe* (Vol. 2). Minneapolis: University of Minneapolis Press.

Aristotle (1941) *Physics*. (Hardie, R. P. and Gaye, R. J., Trans., McKeon, R., Ed.) (Vol. II). New York: Random House.

Ashby, W. R. (1956) *An introduction to cybernetics*. London: Chapman & Hall.

Ashby, W. R. (1960) *Design for a brain: the origin of adaptive behaviour*. (2nd ed.revised.). Chapman & Hall.

Ballard, D., Hayhoe, M., and Pelz, J. (1995) Memory Representations in Natural Tasks. *Journal of Cognitive Neuroscience* 7(1): 66–80.

Barab, S. A., and Plucker, J. A. (2002) Smart people or smart contexts? Cognition, ability, and talent development in an age of situated approaches to knowing and learning. *Educational Psychologist* 37(3): 165–182.

Barandiaran, X. E. (2017) Autonomy and Enactivism: Towards a Theory of Sensorimotor Autonomous Agency. *Topoi-an International Review of Philosophy* 36(3): 409–430.

Barandiaran, X. E., Di Paolo, E. A., and Rohde, M. (2009) Defining Agency: Individuality, Normativity, Asymmetry, and Spatio-temporality in Action. *Adaptive Behavior* 17(5): 367–386.

Bateson, G. (2000) *Steps to an ecology of mind*. Chicago: University of Chicago Press.

Baumgartner, M., and Casini, L. (2017) An Abductive Theory of Constitution. *Philosophy of Science* 84(2): 214–233.

Baumgartner, M., Casini, L., and Krickel, B. (2018) *Horizontal Surgicality and Mechanistic Constitution*. Article in Press doi:10.1007/s10670-018-0033-5.

Baumgartner, M., and Gebharter, A. (2016) Constitutive Relevance, Mutual Manipulability, and Fat-Handedness. *British Journal for the Philosophy of Science* 67(3): 731–756.

Baumgartner, M., and Wilutzky, W. (2017) Is it possible to experimentally determine the extension of cognition? *Philosophical Psychology* 30(8): 1104–1125.

Bechtel, W. (2007a) Biological mechanisms: organised to maintain autonomy. In Boogerd, F. C. (Ed.), *Systems Biology: Philosophical Foundations*. Amsterdam ; Boston: Elsevier.

Bechtel, W. (2007b) Reducing psychology while maintaining its autonomy via mechanistic explanations. In Schouten, M. and Looren de Jong, H. (Eds.), *The Matter of the Mind: Philosophical Essays on Psychology, Neuroscience, and Reduction*. Oxford: Blackwell.

Bechtel, W. (2008) *Mental mechanisms: philosophical perspectives on cognitive neuroscience*. Hove, East Sussex: Psychology Press.

Bechtel, W. (2009a) Mechanism, modularity, and situated cognition. In Robbins, P. and Aydede, M. (Eds.), *The Cambridge Handbook of Situated Cognition*. Cambridge University Press.

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

Bechtel, W. (2011) Mechanism and Biological Explanation. *Philosophy of Science* 78(4): 533–557.

Bechtel, W. (2017) Explicating Top-Down Causation Using Networks and Dynamics. *Philosophy of Science* 84(2): 253–274.

Bechtel, W. (2019) Resituating cognitive mechanisms within heterarchical networks controlling physiology and behavior. *Theory & Psychology* 29(5): 620–639.

Bechtel, W., and 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 SPEC. ISS.): 421–441.

Bechtel, W., and Richardson, R. (2010) *Discovering complexity: decomposition and localization as strategies in scientific research*. (3rd ed.). Cambridge, Mass.: M.I.T. Press.

Bedau, M. (1991) Can Biological Teleology Be Naturalized. *Journal of Philosophy* 88(11): 647–655.

Beer, R. D. (1995) A Dynamical-Systems Perspective on Agent Environment Interaction. *Artificial Intelligence* 72(1–2): 173–215.

Beer, R. D. (1998) Framing the debate between computational and dynamical approaches to cognitive science. *Behavioral and Brain Sciences* 21(5): 630-+.

Beer, R. D. (2000) Dynamical approaches to cognitive science. *Trends in Cognitive Sciences* 4(3): 91–99.

Beer, R. D., and Williams, P. L. (2015) Information Processing and Dynamics in Minimally Cognitive Agents. *Cognitive Science* 39(1): 1–38.

Bennett, M. R., and Hacker, P. M. S. (2003) *Philosophical foundations of neuroscience*. Oxford: Blackwell.

Bickhard, M. H. (2000) Autonomy, function and representation. *Communication and cognition: Artificial Intelligence* 17((3-4)): 111–131.

Bickhard, M. H. (2004) Process and emergence: Normative function and representation. *Axiomathes* 14(1–2): 121–155.

Bickhard, M. H. (2009a) The interactivist model. *Synthese* 166(3): 547–591.

Bickhard, M. H. (2009b) Interactivism: A manifesto. *New Ideas in Psychology* 27(1): 85–95.

Bickhard, M. H. (2011) Systems and process metaphysics. In Hooker, C. (Ed.), *Philosophy of Complex Systems* (Vol. 10). Amsterdam ; Boston: Elsevier.

Bingham, G. (1988) Task-Specific Devices and the Perceptual Bottleneck. *Human Movement Science* 7(2–4): 225–264.

Block, N. (1981) Psychologism and Behaviorism. *Philosophical Review* 90(1): 5–43.

Bolici, F., Howison, J., and Crowston, K. (2016) Stigmergic coordination in FLOSS development teams: Integrating explicit and implicit mechanisms. *Cognitive Systems Research* 38: 14–22.

Bosga, J., and Meulenbroek, R. G. J. (2007) Joint-action coordination of redundant force contributions in a virtual lifting task. *Motor Control* 11(3): 235–258.

Boyd, R., and Richerson, P. J. (1985) *Culture and the evolutionary process*. Chicago : London: University of Chicago Press.

Bratman, M. (1992) Shared Cooperative Activity. *Philosophical Review* 101(2): 327–340.

Bratman, M. (2014) *Shared agency: a planning theory of acting together*. Oxford: Oxford University Press.

Broad, C. D. (1925) *The mind and its place in nature*.

Brooks, R. (1991) Intelligence Without Representation. *Artificial Intelligence* 47(1–3): 139–159.

Bruineberg, J., and Rietveld, E. (2014) Self-organization, free energy minimization, and optimal grip on a field of affordances. *Frontiers in Human Neuroscience* 8: 599.

Burge, T. (1986) Individualism and Psychology. *Philosophical Review* 95(1): 3–45.

Burghes, D., and Downs, A. (1975) *A modern introduction to classical mechanics and control*. Chichester: Ellis Horwood.

Butterfill, S. (2018) Coordinating joint action. In Jankovich, M. and Ludwig, K. (Eds.), *The Routledge Handbook of Collective Intentionality*. Abingdon: Routledge.

Calvo, P. (2016) The philosophy of plant neurobiology: a manifesto. *Synthese* 193(5): 1323–1343.

Calvo, P. (2017) What Is It Like to Be a Plant? *Journal of Consciousness Studies* 24(9–10): 205–227.

Calvo, P., and Baluska, F. (2015) Conditions for minimal intelligence across eukaryota: a cognitive science perspective. *Frontiers in Psychology* 6: 1329.

Calvo, P., Baluska, F., and Sims, A. (2016) ‘Feature Detection’ vs. ‘Predictive Coding’ Models of Plant Behavior. *Frontiers in Psychology* 7: 1505.

Calvo, P., and Friston, K. (2017) Predicting green: really radical (plant) predictive processing. *Journal of the Royal Society Interface* 14(131): 20170096.

Calvo, P., Gagliano, M., Souza, G. M., and Trewavas, A. (2020) Plants are intelligent, here's how. *Annals of Botany* 125(1): 11–28.

Calvo, P., Sahi, V. P., and Trewavas, A. (2017) Are plants sentient? *Plant Cell and Environment* 40(11): 2858–2869.

Calvo, P., and Trewavas, A. (2020) Physiology and the (Neuro)biology of Plant Behavior: A Farewell to Arms. *Trends in Plant Science* 25(3): 214–216.

Carandini, M., and Heeger, D. J. (2012) Normalization as a canonical neural computation. *Nature Reviews Neuroscience* 13(1): 51–62.

Carlson, J. M., and Doyle, J. (2002) Complexity and robustness. *Proceedings of the National Academy of Sciences of the United States of America* 99((suppl 1)): 2538–2545.

Chalmers, D. (2011) Foreword. In *Supersizing the Mind*. Oxford: Oxford University Press.

Chemero, A. (2009) *Radical embodied cognitive science*. Cambridge, Mass.: M.I.T. Press.

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

Chirimuuta, M. (2014) Minimal models and canonical neural computations: the distinctness of computational explanation in neuroscience. *Synthese* 191(2): 127–153.

Chirimuuta, M. (2018a) Explanation in Computational Neuroscience: Causal and Non-causal. *British Journal for the Philosophy of Science* 69(3): 849–880.

Chirimuuta, M. (2018b) Marr, Mayr, and MR: What functionalism should now be about. *Philosophical Psychology* 31(3): 403–418.

Christensen, W. (1996) A complex systems theory of teleology. *Biology & Philosophy* 11(3): 301–320.

Christensen, W. (2004) Self-directedness, integration and higher cognition. *Language Sciences* 26(6): 661–692.

Christensen, W. (2007) The evolutionary origins of volition. In Ross, D., Spurrett, D., Kincaid, H., and Stephens, G. L. (Eds.), *Distributed Cognition and the Will: Individual Volition and Social Context*. Cambridge, Mass.: M.I.T. Press.

Christensen, W. (2012) Natural sources of normativity. *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences* 43(1): 104–112.

Christensen, W., and Bickhard, M. H. (2002) The process dynamics of normative function. *Monist* 85(1): 3–28.

Chromophone 2 physical object modelling synthesizer VST plugin (2014). Montreal: Applied Acoustic systems.

Chu, M., and Kita, S. (2011) The Nature of Gestures' Beneficial Role in Spatial Problem Solving. *Journal of Experimental Psychology-General* 140(1): 102–116.

Churchland, P. S. (1989) *Neurophilosophy*. Cambridge, Mass.: M.I.T. Press.

Churchland, P. S., and Sejnowski, T. J. (1992) *Computational brain*. Cambridge, Mass: M.I.T. Press.

Clark, A. (1997) *Being there: putting brain, body and world together again*. Cambridge, Mass.: M.I.T.Press.

Clark, A. (1999) An embodied cognitive science? *Trends in Cognitive Sciences* 3(9): 345–351.

Clark, A. (2005) Intrinsic content, active memory and the extended mind. *Analysis* 65(1): 1–11.

Clark, A. (2010a) Mementos revenge. In Menary, R. (Ed.), *The Extended Mind*. Cambridge, Mass. ; London: M.I.T. Press.

Clark, A. (2010b) Coupling, constitution, and the cognitive kind: a reply to Adams and Aizawa. In Menary, R. (Ed.), *The Extended Mind*. Cambridge, Mass. ; London: M.I.T. Press.

Clark, A. (2011a) *Supersizing the mind: embodiment, action and cognitive extension*. Oxford: Oxford University Press.

Clark, A. (2011b) Finding the Mind. *Philosophical Studies* 152(3): 447–461.

Clark, A. (18/03/2017) Andy Clark on the Extended Mind. *Philosophy Bites*. Accessed: 26th February 2019 <https://hwcdn.libsyn.com/p/f/e/2/fe24a935c35bf9f2/Andy_Clark_on_The_Extended_Mind.mp3?c_id=14566655&cs_id=14566655&expiration=1551183348&hwt=d871367182296d8a6136b505bb508139. >.

Clark, A., and Chalmers, D. (1998) The extended mind (Active externalism). *Analysis* 58(1): 7–19.

Colombetti, G. (2010) Enaction, sense-making, and emotion. In Stewart, J., Gapenne, O., and Di Paolo, E. A. (Eds.), *Enaction: Toward a New Paradigm for Cognitive Science*. Cambridge, Mass.: M.I.T. Press.

Colombetti, G., and Roberts, T. (2015) Extending the extended mind: the case for extended affectivity. *Philosophical Studies* 172(5): 1243–1263.

Conant, R., and Ashby, W. R. (1970) Every good regulator of a system must be a model of that system. *International Journal of Systems Science* 1(2): 89–97.

Consoli, A. (2016) AC(3)M: The Agent Coordination and Cooperation Cognitive Model. In Tweedale, J. W., NevesSilva, R., Jain, L. C., PhillipsWren, G., Watada, J., and Howlett, R. J. (Eds.), *Intelligent Decision Technology Support in Practice* (Vol. 42). Berlin: Springer-Verlag Berlin.

Corradini, A., and O'Connor, T. eds. (2010) *Emergence in science and philosophy*. Abingdon: Routledge.

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

Craver, C. F. (2006) When mechanistic models explain. *Synthese* 153(3): 355–376.

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

Craver, C. F. (2015) Mechanisms and emergence. *Open MIND* 1: 1–5.

Craver, C. F., and Bechtel, W. (2007) Top-down causation without top-down causes. *Biology & Philosophy* 22(4): 547–563.

Craver, C. F., and Kaplan, D. M. (2014) Towards a mechanistic philosophy of neuroscience. In French, S. and Saatsi, J. (Eds.), *The Bloomsbury Companion to the Philosophy of Science*. London and New York: Bloomsbury.

Craver, C. F., and Kaplan, D. M. (2020) Are more details better? On the norms of completeness for mechanistic explanations. *British Journal for the Philosophy of Science* 71(1): 287–319.

Cummins, R. (1975) Functional-Analysis. *Journal of Philosophy* 72(20): 741–765.

Cummins, R. (1983) *The nature of psychological explanation*. Cambridge, Mass.: M.I.T. Press.

Davidson, D. (1987) Knowing one's own mind. *Proceedings and addresses of the American Philosophical Association* 60: 441–458.

De Caro, M., and Macarthur, D. eds. (2010) *Naturalism and normativity*. New York ; Chichester: Columbia University Press.

De Jaegher, H., and Di Paolo, E. A. (2007) Participatory sense-making: An enactive approach to social cognition. *Phenomenology and the Cognitive Sciences* 6(4): 485–507.

De Jaegher, H., Di Paolo, E. A., and Gallagher, S. (2010) Can social interaction constitute social cognition? *Trends in Cognitive Sciences* 14(10): 441–447.

Deacon, T. W. (2013) *Incomplete nature: how mind emerged from matter*. New York: W.W.Norton.

Deacon, T. W., Srivastava, A., and Bacigalupi, J. A. (2014) The transition from constraint to regulation at the origin of life. *Frontiers in Bioscience-Landmark* 19: 945–957.

Dennett, D. C. (1981) *Brainstorms*. Brighton: Harvester Press.

Dennett, D. C. (1987) *The intentional stance*. Cambridge, Mass.: M.I.T. Press.

Dennett, D. C. (1998) Reflections on language and mind. In Carruthers, P. and Boucher, J. (Eds.), *Language and Thought: Interdisciplinary Themes*. Cambridge: Cambridge University Press.

Dennett, D. C. (2003) *Freedom evolves*. London: Penguin Books.

Dewey, J. (1938) *Logic: the theory of inquiry*. New York: Holt Rinehart & Winston.

Dewey, J. (1958) *Experience and nature*. New York: Dover.

Dewey, J. (1998) Nature, life and body-mind. In Hickman, L. A. and Alexander, T. M. (Eds.), *The Essential Dewey: Vol1: Pragmatism, Education, Democracy* (Vols 1-2, Vol. 1). Bloomington and Indianapolis: Indiana University Press.

Di Paolo, E. A. (2018) The enactive conception of life. In Newen, A., de Bruin, L., and Gallagher, S. (Eds.), *The Oxford Handbook of 4E Cognition*. Oxford: Oxford University Press.

Di Paolo, E. A., and De Jaegher, H. (2017) Neither individualistic nor interactionist. In Durt, C., Fuchs, T., and Tewes, C. (Eds.), *Embodiment, Enaction, and Culture*. Cambridge, Mass: M.I.T. Press.

Donald, M. (1991) *Origins of the modern mind: three stages in the evolution of culture and cognition*. Cambridge, Mass. ; London: Harvard University Press.

Dretske, F. (1997) *Naturalising the mind*. Cambridge, Mass.: M.I.T.Press.

Dreyfus, H. L. (1992) *What computers still can't do*. Cambridge, Mass.: M.I.T. Press.

Dreyfus, H. L. (2014) *Skillful Coping: Essays on the phenomenology of everyday perception and action*. Oxford: Oxford University Press doi:10.1093/acprof:oso/9780199654703.001.0001.

Dupre, J., and Nicholson, D. J. (2018) A manifesto for a processual philosophy of biology. In Dupre, J. and Nicholson, D. J. (Eds.), *Everything Flows*. Oxford: Oxford University Press.

Eisenberg, I. W., Bissett, P. G., Canning, J. R., Dallery, J., Enkavi, A. Z., Whitfield-Gabrieli, S., et al. (2018) Applying novel technologies and methods to inform the ontology of self regulation. *Behaviour Research and Therapy* 101: 46–57.

Eisenberg, I. W., Bissett, P. G., Enkavi, A. Z., Li, J., MacKinnon, D. P., Marsch, L. A., and Poldrack, R. A. (2019) Uncovering the structure of self-regulation through data-driven ontology discovery. *Nature Communications* 10: 2319.

Elman, J. (1991) Distributed Representations, Simple Recurrent Networks, and Grammatical Structure. *Machine Learning* 7(2–3): 195–225.

Estany, A., and Martínez, S. (2014) “Scaffolding” and “affordance” as integrative concepts in the cognitive sciences. *Philosophical Psychology* 27(1): 98–111.

Fabre, J. H. (1916) *The hunting of wasps*. Hodder & Stoughton.

Fodor, J. (1974) Special sciences (or: The disunity of science as a working hypothesis). *Synthese* 28(2): 97–115.

Fodor, J. (1981) *Representations*. Cambridge, Mass.: M.I.T. Press.

Fodor, J. (1983) *The modularity of mind*. Cambridge, Mass.: M.I.T. Press.

Fodor, J. (1985) Fodor’s guide to mental representation: The intelligent auntie’s vade-mecum. *Mind* 94(373): 76–100.

Fodor, J. (1997) Special sciences: Still autonomous after all these years. *Nous* : 149–163.

Fodor, J. (2000) Precis of The Modularity of Mind. In Cummins, R. and Cummins, D. (Eds.), *Minds, Brains and Computers: The Foundations of Cognitive Science*. Oxford: Blackwell.

Fodor, J. (2008) *LOT2: The language of thought revisited*. Oxford: Oxford University Press.

Frazier, P. A., Jamone, L., Althoefer, K., and Calvo, P. (2020) Plant Bioinspired Ecological Robotics. *Frontiers in Robotics and Ai* 7: 79.

Fuchs, T., and De Jaegher, H. (2009) Enactive intersubjectivity: Participatory sense-making and mutual incorporation. *Phenomenology and the Cognitive Sciences* 8(4): 465–486.

Fuller, T. (2016) The Extended Scientific Mind. *Cognitive Systems Research* 40: 75–85.

Gabor, D. (1946) Theory of communication. Part 1: The analysis of information. *Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering* 93(26): 429–441. Presented at the Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering.

Galison, P. (1997) *Image and logic: a material culture of microphysics*. Chicago, Ill. ; London: University of Chicago Press.

Gallagher, S. (2018a) The Extended Mind: State of the Question. *Southern Journal of Philosophy* 56(4): 421–447.

Gallagher, S. (2018b) New mechanisms and the enactivist concept of constitution. In Guta, M. P. (Ed.), *The Metaphysics of Consciousness*. London: Routledge.

Gallagher, S., and Crisafi, A. (2009) Mental Institutions. *Topoi-an International Review of Philosophy* 28(1): 45–51.

Gertler, B. (2007) Overextending the mind? In Gertler, B. and Shapiro, L. (Eds.), *Arguing about the Mind*. New York, London: Routledge.

Gibb, S. C., Hendry, R. F., and Lancaster, T. eds. (2019) *The Routledge handbook of emergence*. London: Routledge.

Gibson, J. J. (2014) *The ecological approach to visual perception*. New York: Psychology Press.

Giere, R. N. (2002a) Scientific cognition as distributed cognition. In *The Cognitive Basis of Science*. Cambridge: Cambridge University Press.

Giere, R. N. (2002b) Discussion note: Distributed cognition in epistemic cultures. *Philosophy of Science* 69(4): 637–644.

Giere, R. N. (2010) An agent-based conception of models and scientific representation. *Synthese* 172(2): 269–281.

Giere, R. N. (2013) Distributed cognition without distributed knowing.

Gilbert, M. (2014) *Joint commitment: how we make the social world*. Oxford: Oxford University Press.

Gillett, C. (2010) Moving beyond the subset model of realization: The problem of qualitative distinctness in the metaphysics of science. *Synthese* 177(2): 165–192.

Glennan, S. (1996) Mechanisms and the nature of causation. *Erkenntnis* 44(1): 49–71.

Glennan, S. (2010) Mechanisms, Causes, and the Layered Model of the World. *Philosophy and Phenomenological Research* 81(2): 362–381.

Glennan, S. (2017) *The new mechanical philosophy*. Oxford: Oxford University Press.

Goldstone, R. L., and Roberts, M. E. (2006) Self-organized trail systems in groups of humans. *Complexity* 11(6): 43–50.

Goldstone, R. L., and Theiner, G. (2017) The multiple, interacting levels of cognitive systems (MILCS) perspective on group cognition. *Philosophical Psychology* 30(3): 334–368.

Goupil, L., Wolf, T., Saint-Germier, P., Aucouturier, J.-J., and Canonne, C. (2021) Emergent Shared Intentions Support Coordination During Collective Musical Improvisations. *Cognitive Science* 45(1): e12932.

Gray, W. D., and Fu, W. T. (2004) Soft constraints in interactive behavior: the case of ignoring perfect knowledge in-the-world for imperfect knowledge in-the-head. *Cognitive Science* 28(3): 359–382.

Grush, R. (2003) In Defense of some 'Cartesian' assumptions concerning the brain and its operation. *Biology & Philosophy* 18(1): 53–93.

Hardcastle, V. G. (2002) On the normativity of functions. In Ariew, A., Cummins, R., and Perlman, M. (Eds.), *Functions: New Essays in the Philosophy of Psychology and Biology*. Oxford: Oxford University Press.

Haugeland, J. (1998) *Having thought: essays in the metaphysics of mind*. Cambridge, Mass. ; London: Harvard University Press.

Hayek, F. A. von (2003) *Law, legislation and liberty : a new statement of the liberal principles of justice and political economy*. Routledge.

Heersmink, R. (2015) Dimensions of integration in embedded and extended cognitive systems. *Phenomenology and the Cognitive Sciences* 14(3): 577–598.

Heidegger, M. (1962) *Being and time*. (Macquarrie, J. and Robinson, E., Trans.). Oxford: Blackwell.

Heider, F., and Simmel, M. (1944) An experimental study of apparent behavior. *American Journal of Psychology* 57: 243–259.

Helbing, D., Keltsch, J., and Molnar, P. (1997) Modelling the evolution of human trail systems. *Nature* 388(6637): 47–50.

Helbing, D., Molnar, P., Farkas, I. J., and Bolay, K. (2001) Self-organizing pedestrian movement. *Environment and Planning B-Planning & Design* 28(3): 361–383.

Helbing, D., Schweitzer, F., Keltsch, J., and Molnar, P. (1997) Active walker model for the formation of human and animal trail systems. *Physical Review E* 56(3): 2527–2539.

Henrich, J. P. (2016) *The secret of our success: how culture is driving human evolution, domesticating our species, and making us smarter*. Princeton: Princeton University Press.

Heylighen, F. (2013) Self-organisation in communicating groups: the emergence of coordination, shared references, and collective intelligence. In Massip-Bonet, A. and Bastardas-Boada, A. (Eds.), *Complexity Perspectives on Language, Communication, and Society*. Berlin: Springer-Verlag.

Heylighen, F. (2016) Stigmergy as a universal coordination mechanism I: Definition and components. *Cognitive Systems Research* 38: 4–13.

Hiernaux, Q. (2021a) Differentiating Behaviour, Cognition, and Consciousness in Plants. *Journal of Consciousness Studies* 28(1–2): 106–135.

Hiernaux, Q. (2021b) History and epistemology of plant behaviour: a pluralistic view? *Synthese* 198(4): 3625–3650.

Hill, M. A. (n.d.) Neutrophil chasing bacteria. *Embryology Movie - Neutrophyl Chasing Bacteria*. Accessed: 12th August 2020 <https://embryology.med.unsw.edu.au/embryology/index.php/Movie_-_Neutrophil_chasing_bacteria. >.

Høffding, S., and Satne, G. (2019) Interactive expertise in solo and joint musical performance. *Synthese*. doi:10.1007/s11229-019-02339-x.

Hofkirchner, W., and Schafranek, M. (2011) General system theory. In Hooker, C. (Ed.), *Philosophy of Complex Systems* (Vol. 10). Amsterdam ; Boston: Elsevier.

Hofstadter, D. (1985) *Metamagical themas*. New York: Basic Books.

Hollan, J., Hutchins, E., and Kirsh, D. (2000) Distributed cognition: Towards a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction (TOCHI)* 7(2): 174.

Hooke, R. (1665) *Micrographia: Some Physiological Descriptions of Minute Bodies made by Magnifying Glasses with Observations and Inquiries Thereupon*. London: Royal Society.

Hooker, C. (2011) Introduction to philosophy of complex systems: A. In Hooker, C. (Ed.), *Philosophy of Complex Systems* (Vol. 10). Amsterdam ; Boston: Elsevier.

Huebner, B. (2014) *Macrocognition: a theory of distributed minds and collective intentionality*. New York: Oxford University Press.

Humphreys, P. (1997) How properties emerge. *Philosophy of Science* 64(1): 1–17.

Humphreys, P. (2016) *Emergence: a philosophical account*. New York, NY: Oxford University Press.

Hurley, S. (1998) *Consciousness in action*. Cambridge, Mass: Harvard University Press.

Hurley, S. (2010) The varieties of externalism. In Menary, R. (Ed.), *The Extended Mind*. Cambridge, Mass: M.I.T. Press.

Hutchins, E. (1991) The social organisation of distributed cognition. In Resnick, L. B., Levine, J. M., and Teasley, S. D. (Eds.), *Perspectives on Socially Shared Cognition*. Washington DC: American Psychological Association.

Hutchins, E. (1995a) *Cognition in the wild*. Cambridge, Mass. ; London: MIT Press.

Hutchins, E. (1995b) How a Cockpit Remembers Its Speeds. *Cognitive Science* 19(3): 265–288.

Hutchins, E. (2008) The role of cultural practices in the emergence of modern human intelligence. *Philosophical Transactions of the Royal Society B-Biological Sciences* 363(1499): 2011–2019.

Hutchins, E. (2011) Enculturating the Supersized Mind. *Philosophical Studies* 152(3): 437–446.

Hutchins, E. (2014) The cultural ecosystem of human cognition. *Philosophical Psychology* 27(1): 34–49.

Hutto, D. D. (2008) *Folk psychological narratives: the sociocultural basis of understanding readings*. Cambridge, Mass.: M.I.T. Press.

Hutto, D. D. (2013) Fictionalism About Folk Psychology. *Monist* 96(4): 582–604.

Hutto, D. D. (2021) A brickhouse defence for folk psychology: how to defeat 'Big Bad Wolf' eliminativism.

Hutto, D. D., Kirchhoff, M., and Myin, E. (2014) Extensive enactivism: why keep it all in? *Frontiers in Human Neuroscience* 8: 706.

Hutto, D. D., and Myin, E. (2013) *Radicalizing enactivism: basic minds without content*. Cambridge, Mass: MIT Press.

Hutto, D. D., and Myin, E. (2018) Going radical. In Newen, A., de Bruin, L., and Gallagher, S. (Eds.), *The Oxford Handbook of 4E Cognition*. Oxford: Oxford University Press.

Hutto, D. D., and Satne, G. (2015) The Natural Origins of Content. *Philosophia* 43(3): 521–536.

Hutto, D. D., and Satne, G. (2017) Continuity skepticism in doubt. In Durt, C., Fuchs, T., and Tewes, C. (Eds.), *Embodiment, Enaction, and Culture*. Cambridge, Mass: M.I.T. Press.

Hutto, D. D., and Satne, G. (2018) Wittgenstein's Inspiring View of Nature: On Connecting Philosophy and Science Aright. *Philosophical Investigations* 41(2): 141–160.

Hyvarinen, A., and Hoyer, P. O. (2001) A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Research* 41(18): 2413–2423.

Illari, P. M., and Williamson, J. (2012) What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for Philosophy of Science* 2(1): 119–135.

Janssen, A., Klein, C., and Slors, M. V. P. (2017) What is a cognitive ontology, anyway? *Philosophical Explorations* 20(2): 123–128.

Johnson, M. (2010) Cognitive science and Dewey's theory of mind, thought, and language. In Cochran, M. (Ed.), *The Cambridge Companion to Dewey*. Cambridge: Cambridge University Press.

Kaiser, M. I., and Krickel, B. (2017) The Metaphysics of Constitutive Mechanistic Phenomena. *British Journal for the Philosophy of Science* 68(3): 745–779.

Kaplan, D. M. (2011) Explanation and description in computational neuroscience. *Synthese* 183(3): 339–373.

Kaplan, D. M. (2012) How to demarcate the boundaries of cognition. *Biology & Philosophy* 27(4): 545–570.

Keijzer, F. (2013) The Sphex story: How the cognitive sciences kept repeating an old and questionable anecdote. *Philosophical Psychology* 26(4): 502–519.

Kirchhoff, M. (2012) Extended cognition and fixed properties: steps to a third-wave version of extended cognition. *Phenomenology and the Cognitive Sciences* 11(2): 287–308.

Kirchhoff, M. (2014) Extended cognition & constitution: Re-evaluating the constitutive claim of extended cognition. *Philosophical Psychology* 27(2): 258–283.

Kirchhoff, M. (2015) Cognitive assembly: towards a diachronic conception of composition. *Phenomenology and the Cognitive Sciences* 14(1): 33–53.

Kirchhoff, M. (2017) From mutual manipulation to cognitive extension: challenges and implications. *Phenomenology and the Cognitive Sciences* 16(5): 863–878.

Kirchhoff, M., and Kiverstein, J. (2019) *Extended consciousness and predictive processing*. Abingdon: Routledge.

Kirsh, D. (1995) The Intelligent Use of Space. *Artificial Intelligence* 73(1–2): 31–68.

Kirsh, D. (1999) Distributed cognition, coordination and environment design. *Proceedings of the European congerence on cognitive sience* : 1–11.

Kirsh, D. (2005) *Metacognition, distributed cognition, and visual design*. (Gardenfors, P. and Johansson, P., Eds.). Mahwah: Lawrence Erlbaum Assoc Publ.

Kirsh, D. (2006) Distributed cognition: A methodological note. *Pragmatics and Cognition* 14(2): 249–262.

Kirsh, D. (2009) Problem solving and situated cognition. In Robbins, P. and Aydede, M. (Eds.), *The Cambridge Handbook of Situated Cognition*. Cambridge University Press.

Kirsh, D., and Maglio, P. (1994) On Distinguishing Epistemic from Pragmatic Action. *Cognitive Science* 18(4): 513–549.

Kistler, M. (2009) Mechanisms and downward causation. *Philosophical Psychology* 22(5): 595–609.

Kitano, H. (2004) Biological robustness. *Nature Reviews Genetics* 5(11): 826–837.

Kitano, H. (2007) Towards a theory of biological robustness. *Molecular Systems Biology* 3: 137.

Kiverstein, J. (2012) What is Heideggerian cognitive science? In Kiverstein, J. and Wheeler, M. (Eds.), *Heidegger and Cognitive Science*. Basingstoke: Palgrave Macmillan.

Kiverstein, J. (2015) Empathy and the responsiveness to social affordances. *Consciousness and Cognition* 36: 532–542.

Kiverstein, J. (2017) Sociality and the human mind. In Kiverstein, J. (Ed.), *The Routledge Handbook of Philosophy of the Social Mind*. Abingdon: Routledge.

Kiverstein, J. (2018) Extended cognition. In Newen, A., de Bruin, L., and Gallagher, S. (Eds.), *The Oxford Handbook of 4E Cognition*. Oxford: Oxford University Press.

Kiverstein, J. (2020a) In Defence of a Relational Ontology of Affordances. *Constructivist Foundations* 15(3): 226–229.

Kiverstein, J. (2020b) Free Energy and the Self: An Ecological-Enactive Interpretation. *Topoi-an International Review of Philosophy* 39(3): 559–574.

Kiverstein, J., and Clark, A. (2009) Introduction: Mind Embodied, Embedded, Enacted: One Church or Many? *Topoi* 28(1): 1–7.

Kiverstein, J., and Farina, M. (2011) Embraining Culture: Leaky Minds and Spongy Brains. *Teorema* 30(2): 35–53.

Kiverstein, J., and Rietveld, E. (2015) The Primacy of Skilled Intentionality: on Hutto & Satne's the Natural Origins of Content. *Philosophia* 43(3): 701–721.

Kiverstein, J., and Rietveld, E. (2018) Reconceiving representation-hungry cognition: an ecological-enactive proposal. *Adaptive Behavior* 26(4): 147–163.

Kiverstein, J., and Rietveld, E. (2020) Skill-based engagement with a rich landscape of affordances as an alternative to thinking through other minds. *Behavioral and Brain Sciences* 43: e106.

Kiverstein, J., and Rietveld, E. (2021) Scaling-up skilled intentionality to linguistic thought. *Synthese*. doi:10.1007/s11229-020-02540-3.

Knorr-Cetina, K. (1999) *Epistemic cultures: how the sciences make knowledge*. Cambridge, Mass: Harvard University Press.

Krickel, B. (2017) Making Sense of Interlevel Causation in Mechanisms from a Metaphysical Perspective. *Journal for General Philosophy of Science* 48(3): 453–468.

Krickel, B. (2018a) Saving the mutual manipulability account of constitutive relevance. *Studies in History and Philosophy of Science* 68: 58–67.

Krickel, B. (2018b) *The mechanical world: the metaphysical commitments of the new mechanistic approach*. Cham: Springer Nature Switzerland AG.

Krickel, B. (2019a) *Extended cognition: the new mechanists' mutual manipulability criterion and the challenge of trivial extendedness*. forthcoming in Mind and Language, Bochum.

Krickel, B. (2019b) Extended cognition, the new mechanists' mutual manipulability criterion, and the challenge of trivial extendedness. *Mind & Language* : 1–23. doi:10.1111/mila.12262.

Krueger, J. (2012) Seeing mind in action. *Phenomenology and the Cognitive Sciences* 11(2): 149–173.

Krueger, J. (2018) Direct social perception. In Newen, A., de Bruin, L., and Gallagher, S. (Eds.), *The Oxford Handbook of 4E Cognition*. Oxford: Oxford University Press.

Ladyman, J., and Wiesner, K. (2020) *What is a complex system?* New Haven, London: Yale University Press.

Lakoff, G., and Johnson, M. (1980) *Metaphors we live by*. Chicago : London: University of Chicago Press.

Laughlin, R. (2005) *A different universe: remaking physics from the bottom down*. New York: Basic Books.

Leijnen, S., Heskes, T., and Deacon, T. W. (2016) *Exploring Constraint: Simulating Self-Organization and Autogenesis in the Autogenic Automaton*. (Gershenson, C., Froese, T., Siqueiros, J. M., Aguilar, W., Izquierdo, E., and Sayama, H., Eds.)*Alife 2016, the Fifteenth International Conference on the Synthesis and Simulation of Living Systems*. Cambridge: Mit Press.

Leuridan, B. (2012) Three Problems for the Mutual Manipulability Account of Constitutive Relevance in Mechanisms. *British Journal for the Philosophy of Science* 63(2): 399–427.

Linson, A., and Calvo, P. (2020) Zoocentrism in the weeds? Cultivating plant models for cognitive yield. *Biology & Philosophy* 35(5): 49.

List, C., and Pettit, P. (2011) *Group agency: the possibility, design, and status of corporate agents*. Oxford ; New York: Oxford University Press.

Liu, L., Aerbajinai, W., Ahmed, S. M., Rodgers, G. P., Angers, S., and Parent, C. A. (2012) Radil controls neutrophil adhesion and motility through beta 2-integrin activation. *Molecular Biology of the Cell* 23(24): 4751–4765.

MacDonald, G., and Papineau, D. (2006) Prospects and problems for teleosemantics. In MacDonald, G. and Papineau, D. (Eds.), *Teleosemantics*. Oxford: Oxford University Press.

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

Magnus, P. D. (2007) Distributed cognition and the task of science. *Social Studies of Science* 37(2): 297–310.

Magnus, P. D., and McClamrock, R. (2015) Friends with benefits! Distributed cognition hooks up cognitive and social conceptions of science. *Philosophical Psychology* 28(8): 1114–1127.

Maguire, E. A., Gadian, D. G., Johnsrude, I. S., Good, C. D., Ashburner, J., Frackowiak, R. S. J., and Frith, C. D. (2000) Navigation-related structural change in the hippocampi of taxi drivers. *Proceedings of the National Academy of Sciences of the United States of America* 97(8): 4398–4403.

Maguire, E. A., Nannery, R., and Spiers, H. J. (2006) Navigation around London by a taxi driver with bilateral hippocampal lesions. *Brain* 129: 2894–2907.

Maley, C. J., and Piccinini, G. (2017) A unified mechanistic account of teleological functions for psychology and neuroscience. In Kaplan, D. M. (Ed.), *Explanation and Integration in Mind and Brain Science*. Oxford: Oxford University Press.

Marsh, L., and Onof, C. (2008) Stigmergic epistemology, stigmergic cognition. *Cognitive Systems Research* 9(1–2): 136–149.

Matthews, R., and Matthews, J. (2009) Nesting Behavior of Abispa ephippium (Fabricius) (Hymenoptera: Vespidae: Eumeninae): Extended Parental Care in an Australian Mason Wasp. *Psyche* : 1–15. doi:doi:10.1155/2009/851694.

Maturana, H. R., and Varela, F. J. (1980) *Autopoiesis and cognition: the realization of the living*. Dordrecht ; London: D. Reidel.

Mayr, E. (1961) Cause and Effect in Biology - Kinds of Causes, Predictability, and Teleology Are Viewed by a Practicing Biologist. *Science* 134(348): 1501-.

McClamrock, R. (1991) Marr's three levels: A re-evaluation. *Minds & Machines* 1(2): 185.

Mediano, P. A. M., Trewavas, A., and Calvo, P. (2021) Information and Integration in Plants Towards a Quantitative Search for Plant Sentience. *Journal of Consciousness Studies* 28(1–2): 80–105.

Menary, R. (2006) Attacking the bounds of cognition. *Philosophical Psychology* 19(3): 329–344.

Menary, R. (2007) *Cognitive integration: mind and cognition unbounded*. Australia, Australia/Oceania: Research Online.

Menary, R. (2009) Intentionality, cognitive integration and the continuity thesis. *Topoi* 28(1): 31–43.

Menary, R. (2010a) Cognitive integration and the extended mind. In Menary, R. (Ed.), *The Extended Mind*. Cambridge, Mass.: M.I.T. Press.

Menary, R. (2010b) Dimensions of mind. *Phenomenology and the Cognitive Sciences* 9(4): 561–578.

Menary, R. (2012) Cognitive practices and cognitive character. *Philosophical Explorations* 15(2): 147–164.

Menary, R. (2013) Cognitive integration, enculturated cognition and the socially extended mind. *Cognitive Systems Research* 25–26: 26–34.

Menary, R. (2015) Mathematical cognition: a case of enculturation. In Metzinger, T. and Windt, J. (Eds.), *Open MIND* (Vol. 25). MIND Group.

Menary, R. (2018) Cognitive integration: how culture transforms us and extends our cognitive capabilities. In Newen, A., de Bruin, L., and Gallagher, S. (Eds.), *The Oxford Handbook of 4E Cognition*. Oxford: Oxford University Press.

Milkowski, M. (2013) *Explaining the computational mind*. Cambridge, Mass.: M.I.T. Press.

Milkowski, M. (2017) The false dichotomy between causal realisation and semantic computation. *Hybris* 38: 1–20.

Millikan, R. G. (1984) *Language, thought and other biological categories: new foundations for realism*. Cambridge, Mass. ; London: MIT Press.

Millikan, R. G. (1989) In Defense of Proper Functions. *Philosophy of Science* 56(2): 288–302.

Millikan, R. G. (1996) Pushmi-pullyu representations. In May, L., Friedman, M., and Clark, A. (Eds.), *Mind and Morals: Essays on Cognitive Science and Ethics*. Cambridge, Mass ; London: MIT Press.

Millikan, R. G. (2002) Biofunctions: two paradigms. In Ariew, Andre, Cummins, R., and Perlman, M. (Eds.), *Functions: New Essays in the Philosophy of Psychology and Biology*. Oxford: Oxford University Press.

Miyazono, K. (2017) Does functionalism entail extended mind? *Synthese* 194(9): 3523–3541.

Moreno, A., Etxeberria, A., and Umerez, J. (2008) The autonomy of biological individuals and artificial models. *Biosystems* 91(2): 309–319.

Moreno, A., and Mossio, M. (2015) *Biological autonomy: a philosophical and theoretical enquiry*. Dordrecht: Springer.

Moreno, A., Ruiz-Mirazo, K., and Barandiaran, X. (2011) The impact of the paradigm of complexity on the foundational frameworks of biology and cognitive science. In Hooker, C. (Ed.), *Philosophy of Complex Systems* (Vol. 10). Amsterdam ; Boston: Elsevier.

Morton, A. (2003) *The importance of being understood*. London: Routledge.

Moss, L., and Nicholson, D. J. (2012) On nature and normativity: Normativity, teleology, and mechanism in biological explanation. *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences* 43(1): 88–91.

Nalepka, P., Kallen, R. W., Chemero, A., Saltzman, E., and Richardson, M. J. (2017) Herd Those Sheep: Emergent Multiagent Coordination and Behavioral-Mode Switching. *Psychological Science* 28(5): 630–650.

Nalepka, P., Kallen, R. W., Lamb, M., and Richardson, M. J. (2018) *Emergence of efficient, coordinated solutions despite differences in agent ability during human-machine interaction Demonstration using a multiagent 'shepherding' task*. *18th Acm International Conference on Intelligent Virtual Agents (Iva'18)*. New York: Assoc Computing Machinery doi:10.1145/3267851.3267879.

Neander, K. (1991) The Teleological Notion of Function. *Australasian Journal of Philosophy* 69(4): 454–468.

Neander, K. (1995) Misrepresenting and Malfunctioning. *Philosophical Studies* 79(2): 109–141.

Neander, K. (2017) Functional analysis and the species design. *Synthese* 194(4): 1147–1168.

Newen, A., de Bruin, L., and Gallagher, S. (2018) 4E cognition: historical roots, key concepts, and central issues. In Newen, A., de Bruin, L., and Gallagher, S. (Eds.), *The Oxford Handbook of 4E Cognition*. Oxford: Oxford University Press.

Nicholson, D. J. (2012) The concept of mechanism in biology. *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences* 43(1): 152–163.

Nicholson, D. J. (2013) Organisms <> machines. *Studies in history and philosophy of biological and Biomedical Sciences* 44: 669–678.

Nicholson, D. J. (2014) The machine conception of the organism in development andevolution: A critical analysis. *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences* 48(PB): 162–174.

Nieto-Gomez, R. (2016) Stigmergy at the edge: Adversarial stigmergy in the war on drugs. *Cognitive Systems Research* 38: 31–40.

Oliver, A. (1996) The Metaphysics of Properties. *Mind* 105(417): 1–80.

O'Regan, J. K., and Noe, A. (2001) A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences* 24(5): 939-+.

Palermos, S. O. (2014) Loops, constitution, and cognitive extension. *Cognitive Systems Research* 27: 25–41.

Parunak, H. V. D. (2006) A survey of environments and mechanisms for human-human stigmergy. In Weyns, D., Parunak, H. V. D., and Michel, F. (Eds.), *Environments for Multi-Agent Systems Ii* (Vol. 3830). Berlin: Springer-Verlag Berlin.

Pattee, H. (1969) Primitive conditions for physical functional hierachies. In Whyte, L. L., Wilson, A. G., and Wilson, D. (Eds.), *Hierarchical Structures*. Presented at the Hierarchical structures symposium held at Douglas Advanced Research Laboratories Huntington Beach CaliforniaNov 18 - 19 1968 New York: American Elsevier.

Pattee, H. (1971) Physical theories of biological coordination. *Quarterly Reviews of Biophysics* 4(2 and 3): 255–276.

Pattee, H. (1973) The physical basis and origin of hierarchical control. In Pattee, H. (Ed.), *Hierarchy Theory: The Challenge of Complex Systems*. New York: George Braziller.

Pattee, H. (1979) Complementarity Principle and the Origin of Macromolecular Information. *Biosystems* 11(2–3): 217–226.

Pattee, H. (1983) Comment by H.H. Pattee. *Journal of Social and Biological Structures* 6(2): 146–147.

Pattee, H. (2013) Epistemic, Evolutionary, and Physical Conditions for Biological Information. *Biosemiotics* 6(1): 9–31.

Perrow, C. (1999) *Normal accidents: living with high-risk technologies*. Princeton: Princeton University Press.

Perry, R. B. (1921) A behavioristic view of purpose. *Journal of Philosophy* 18: 85–105.

Pezzulo, G. (2011) Shared Representations as Coordination Tools for Interaction. *Review of Philosophy and Psychology* 2(2): 303–333.

Pezzulo, G. (2015) Shared Action Spaces: (How) does the brain merge the co-actors' spatial representations during social interactions? *Cognitive Processing* 16: S25–S25.

Piccinini, G. (2004) Functionalism, computationalism, and mental contents. *Canadian Journal of Philosophy* 34(3): 375–410.

Piccinini, G. (2007) Computing Mechanisms. *Philosophy of Science* 74(4): 501–526.

Piccinini, G. (2015) *Physical computation: a mechanistic account*. Oxford: Oxford University Press.

Piccinini, G., and Scarantino, A. (2016) Computation and information. In Floridi, L. (Ed.), *The Routledge Handbook of Philosophy of Information*. London and New York: Routledge.

Pickering, A. (2010) *The cybernetic brain: sketches of another future*. Chicago, Ill: University of Chicago Press.

Piredda, G. (2017) The Mark of the Cognitive and the Coupling-Constitution Fallacy: A Defense of the Extended Mind Hypothesis. *Frontiers in Psychology* 8: 2061.

Poirier, P., and Chicoisne, G. (2006) A framework for Thinking about Distributed Cognition.

Poldrack, R. A., and Yarkoni, T. (2016) From Brain Maps to Cognitive Ontologies: Informatics and the Search for Mental Structure. In Fiske, S. T. (Ed.), *Annual Review of Psychology, Vol 67* (Vol. 67). Palo Alto: Annual Reviews.

Powell, A., and Dupré, J. (2009) From molecules to systems: the importance of looking both ways. *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences* 40(1): 54–64.

Prigogine, I., and Stengers, I. (1984) *Order out of chaos: man's new dialogue with nature*. London: Verso.

Productions, L. S. (2014/14/November) Applying the Knowledge. *The New York Times*.

Psujek, S., Ames, J., and Beer, R. D. (2006) Connection and coordination: The interplay between architecture and dynamics in evolved model pattern generators. *Neural Computation* 18(3): 729–747.

Putnam, H. (1973) Meaning and reference. *Journal of Philosophy* 70(19): 699–711.

Putnam, H. (1988) *Representation and reality*. Cambridge, Mass. ; London: MIT.

Putnam, H. (2004) Psychological predicates. In Heil, J. (Ed.), *Philosophy of Mind: A Guide and Anthology*. Oxford: Oxford University Press.

Raja, V., and Anderson, M. L. (2020) Behavior considered as an enabling constraint. In Calzavarini, F. and Viola, M. (Eds.), *Neural Mechanisms*. Switzerland: Springer Nature Switzerland AG.

Raja, V., Silva, P. L., Holghoomi, R., and Calvo, P. (2020) The dynamics of plant nutation. *Scientific Reports* 10(1): 19465.

Rapoport, A. (1986) *General system theory*. Tunbridge Wells: Abacus Press.

Rawls, J. (1971) *A theory of justice*. Cambridge, Mass.: Harvard University Press.

Rescher, N. (2000) *Process philosophy*. Pittsburgh: Unbiversity of Pittsburgh Press.

Rescorla, M. (2013) Against Structuralist Theories of Computational Implementation. *British Journal for the Philosophy of Science* 64(4): 681–707.

Resnick, M. (1997) *Turtles, termites, and traffic jams: explorations in massively parallel microworlds*. Cambridge, Mass. ; London: MIT.

Resnick, M., and Klopfer, E. (2018) *Starlogo*. Java, C, Cambridge, Mass.: M.I.T. Media Lab.

Ricci, A., Omicini, A., Viroli, M., Gardelli, L., and Oliva, E. (2007) Cognitive stigmergy: Towards a framework based on agents and artifacts. In Weyns, D., Parunak, H. V. D., and Michel, F. (Eds.), *Environments for Multi-Agent Systems Iii* (Vol. 4389). Berlin: Springer-Verlag Berlin.

Richardson, M. J., Kallen, R. W., Nalepka, P., Harrison, S. J., Lamb, M., Chemero, A., Saltzman, E., and Schmidt, R. C. (2016) *Modeling Embedded Interpersonal and Multiagent Coordination*. (Munoz, V. M., Gusikhin, O., and Chang, V., Eds.). Setubal: Scitepress.

Rietveld, E. (2008) Situated Normativity: The Normative Aspect of Embodied Cognition in Unreflective Action. *Mind* 117(468): 973–1001.

Rietveld, E., Denys, D., and van Westen, M. (2018) Ecological-enactive cognition as engaging with a field of relevant affordances: the skilled intentionality framework. In Newen, A., de Bruin, L., and Gallagher, S. (Eds.), *The Oxford Handbook of 4E Cognition*. Oxford: Oxford University Press.

Rietveld, E., and Kiverstein, J. (2014) A Rich Landscape of Affordances. *Ecological Psychology* 26(4): 325–352.

Romero, F. (2015) Why there isn't inter-level causation in mechanisms. *Synthese* 192(11): 3731–3755.

Rosenblueth, A., Wiener, N., and Bigelow, J. (2017) Behavior, purpose, and teleology. In *Systems Research for Behavioral Science: A Sourcebook*.

Ross, D., and Ladyman, J. (2010) The alleged coupling-constitution fallacy. In Menary, R. (Ed.), *The Extended Mind*. Cambridge, Mass. ; London: M.I.T. Press.

Roth, M., and Cummins, R. (2014) Two tales of functional explanation. *Philosophical Psychology* 27(6): 773–788.

Roth, M., and Cummins, R. (2017) Neuroscience, psychology, reduction, and functional analysis. In Kaplan, D. M. (Ed.), *Explanation and Integration in Mind an Brain Science*. Oxford: Oxford University Press.

Rowlands, M. (2003) *Externalism: putting mind and world back together again.* Montreal: McGill - Queens University Press.

Rowlands, M. (2009) Enactivism and the Extended Mind. *Topoi* 28(1): 53–62.

Rowlands, M. (2010) *The new science of the mind: from extended mind to embodied phenomenology*. Cambridge, Mass. [u.a.: MIT Press.

Roy, A. E. (1973) *Orbital motion*. Boca Raton: CRC Press.

Rumelhart, D. E., and McClelland, J. L. (1986) *Parallel distributed processing: explorations in the microstructure of cognition*. (Vol. 1). Cambridge, Mass. ; London: MIT Press.

Rupert, R. D. (2004) Challenges to the hypothesis of extended cognition. *Journal of Philosophy* 101(8): 389–428.

Rupert, R. D. (2009a) *Cognitive systems and the extended mind*. New York ; Oxford: Oxford University Press.

Rupert, R. D. (2009b) Innateness and the situated mind. In Robbins, Philip and Aydede, M. (Eds.), *The Cambridge Handbook of Situated Cognition*. Cambridge: Cambridge University Press.

Rupert, R. D. (2010a) Systems, functions, and intrinsic natures: On Adams and Aizawa's The Bounds of Cognition. *Philosophical Psychology* 23(1): 113–123.

Rupert, R. D. (2010b) Extended cognition and the priority of cognitive systems. *Cognitive Systems Research* 11(4): 343–356.

Rupert, R. D. (2011) Cognitive systems and the supersized mind. *Philosophical Studies* 152(3): 427–436.

Rupert, R. D. (2012) Supersizing the Mind: Embodiment, Action, and Cognitive Extension. *Philosophical Review* 121(2): 304–308.

Rupert, R. D. (2013) Memory, Natural Kinds, and Cognitive Extension; or, Martians Don't Remember, and Cognitive Science Is Not about Cognition. *Review of Philosophy and Psychology* 4(1): 25–47.

Rupert, R. D. (2014) Against group cognitive states. In Chant, S. R., Hindriks, F., and Preyer, G. (Eds.), *From Individual to Collective Intentionality: New Essays*. Oxford: Oxford University Press.

Rupert, R. D. (2016) Review: Mark Rowlands The New Science of the Mind: From Extended Mind to Embodied Phenomenology. *Mind* 125(497): 206–226.

Rutherford, M. D., Pennington, B. F., and Rogers, S. J. (2006) The perception of animacy in young children with autism. *Journal of Autism and Developmental Disorders* 36(8): 983–992.

Salmon, W. (1994) Causality Without Counterfactuals. *Philosophy of Science* 61(2): 297–312.

Salmon, W. C. (1998) *Causality and explanation*. New York ; Oxford: Oxford University Press.

Satne, G. (2015) The social roots of normativity. *Phenomenology and the Cognitive Sciences* 14(4): 673–682.

Schlicht, T. (2018) Critical note: cognitive systems and the dynamics of representing-in-the-world. In Newen, A., de Bruin, L., and Gallagher, S. (Eds.), *The Oxford Handbook of 4E Cognition*. Oxford: Oxford University Press.

Searle, J. R. (1984) *Minds, brains and science*. Cambridge, Mass.: Harvard University Press.

Searle, J. R. (1998) *Mind, language and society: philosophy in the real world*. New York: Basic Books.

Searle, J. R. (2004) Minds, brains, and programs. In Heil, J. (Ed.), *Philosophy of Mind a Guide and Anthology*. Oxford: Oxford University Press.

Segundo-Ortin, M., and Calvo, P. (2019) Are plants cognitive? A reply to Adams. *Studies in History and Philosophy of Science* 73: 64–71.

Seibt, J. (2004) Free Process Theory: Towards a Typology of Occurrings. *Axiomathes* 14(1): 23–55.

Seibt, J. (2009) Forms of emergent interaction in General Process Theory. *Synthese* 166(3): 479–512.

Seibt, J. (2015) Non-Transitive Parthood, Leveled Mereology, and the Representation of Emergent Parts of Processes. *Grazer Philosophische Studien* 91(1): 165–190.

Seibt, J. (2018) Ontological tools for the process turn in biology. In Dupre, J. and Nicholson, D. J. (Eds.), *Everything Flows*. Oxford: Oxford University Press.

Seibt, J., and Rescher, N. (2018) Process philosophy. In (Zalta, E., Ed.)*Stanford Encyclopedia of Philosophy*.

Shapiro, L. A. (2008) Functionalism and mental boundaries. *Cognitive Systems Research* 9(1–2): 5–14.

Shapiro, L. A. (2011) *Embodied cognition*. London: Routledge.

Silberstein, M., and McGeever, J. (1999) The search for ontological emergence. *Philosophical Quarterly* 49(195): 182–200.

Simon, H. (1996) *The sciences of the artificial*. (Third Edition.). Cambridge, Mass.: M.I.T. Press.

Slors, M. V. P. (2019) Symbiotic cognition as an alternative for socially extended cognition. *Philosophical Psychology*.

Slors, M. V. P. (forthcoming a) *Cultural intelligence as symbiotic cognition*.

Slors, M. V. P. (forthcoming b) *Cultural conventions as group-makers*.

Smortchkova, J. (2020) Seeing goal-directedness: a case for social perception. *British Journal for the Philosophy of Science* 71: 855–879.

Southgate, V., Johnson, M. H., and Csibra, G. (2008) Infants attribute goals even to biomechanically impossible actions. *Cognition* 107(3): 1059–1069.

Spaulding, S. (2013) Mirror Neurons and Social Cognition. *Mind & Language* 28(2): 233–257.

Spaulding, S. (2015) On Direct Social Perception. *Consciousness and Cognition* 36: 472–482.

Spezzano, G. ed. (2019) *Swarm robotics*. Basel: MDPI.

Sprevak, M. (2009) Extended Cognition and Functionalism. *Journal of Philosophy* 106(9): 503–527.

Sprevak, M. (2010) Inference to the hypothesis of extended cognition. *Studies in History and Philosophy of Science* 41(4): 353–362.

Star, S. L. (2010) This is Not a Boundary Object: Reflections on the Origin of a Concept. *Science Technology & Human Values* 35(5): 601–617.

Star, S. L., and Griesemer, J. (1989) Institutional Ecology, Translations and Boundary Objects - Amateurs and Professionals in Berkeleys-Museum-of-Vertebrate-Zoology, 1907-39. *Social Studies of Science* 19(3): 387–420.

Stephan, A. (2018) 3E's are sufficient but dont forget the D. In Newen, A., de Bruin, L., and Gallagher, S. (Eds.), *The Oxford Handbook of 4E Cognition*. Oxford: Oxford University Press.

Stewart, J., Gapenne, O., and Di Paolo, E. A. (2010) *Enaction: toward a new paradigm for cognitive science*. Cambridge, Mass.: M.I.T. Press.

Stout, R. (1996) *Things that happen because they should: a teleological approach to action*. Oxford: Oxford University Press.

Susi, T. (2016) Social cognition, artefacts, and stigmergy revisited: Concepts of coordination. *Cognitive Systems Research* 38: 41–49.

Sutton, J. (2006) Distributed cognition: Domains and dimensions. *Pragmatics and Cognition* 14(2): 235–247.

Sutton, J. (2010) Exograms and interdisciplinarity: history, the extended mind, and the civilizing process. In Menary, R. (Ed.), *The Extended Mind*. Cambridge, Mass.: M.I.T. Press.

Sutton, J., Harris, C. B., Keil, P. G., and Barnier, A. J. (2010) The psychology of memory, extended cognition, and socially distributed remembering. *Phenomenology and the Cognitive Sciences* 9(4): 521–560.

Talbot, W., Dariansm.i, Kornhuber, H., and Mountcastle, V. (1968) Sense of Flutter-Vibration - Comparison of Human Capacity with Response Patterns of Mechanoreceptive Afferents from Monkey Hand. *Journal of Neurophysiology* 31(2): 301-+.

Theiner, G. (2009) Making sense of group cognition: the curious case of transactive memory systems. In Christensen, W., Schier, E., and Sutton, J. (Eds.), *ASCS09 Proceedings of the 9th Conference of the Australasian Society for Cognitive Science*. Presented at the 9th conference of the Australasian Society for Cognitive Science Sydney: Sydney: Macquarie Centre for Cognitive Science.

Theiner, G. (2011) *Res cogitans extensa: a philosophical defence of the extended mind thesis*. Frankfurt am Main: Peter Lang.

Theiner, G. (2018) Groups as distributed cognitive systems. In Jankovich, M. and Ludwig, K. (Eds.), *The Routledge Handbook of Collective Intentionality*. Abingdon: Routledge.

Theraulaz, G., and Bonabeau, E. (1999) A brief history of stigmergy. *Artificial Life* 5(2): 97–116.

Thom, R. (1975) *Structural stability and morphogenesis*. Reading, Mass.: Benjamin/Cummings.

Thompson, E. (2007) *Mind in life: biology, phenomenology, and the sciences of mind*. Cambridge, Mass: Belknap Press of Harvard University Press.

Thorndike, L., and Bruce, D. (2017) *Animal Intelligence: Experimental Studies*. New York: Routledge doi:10.4324/9781351321044.

Thurner, S., Hanel, R., and Klimek, P. (2018) *Introduction to the theory of complex systems*. Oxford: Oxford University Press.

Tolman, E. C. (1948) Cognitive Maps in Rats and Men. *Psychological Review* 55(4): 189–208.

Tolman, E. C. (1967) *Purposive behavior in animals and men*. (Second edition.). New York: Meredith.

Tomasello, M. (1999) *The cultural origins of human cognition*. Cambridge, Mass.: Harvard University Press.

Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005) Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences* 28(5): 675-+.

Toon, A. (2014a) Friends at last? Distributed cognition and the cognitive/social divide. *Philosophical Psychology* 27(1): 112–125.

Toon, A. (2014b) Empiricism for Cyborgs. *Philosophical Issues* 24(1): 409–425.

Toon, A. (2015) Where is the understanding? *Synthese* 192(12): 3859–3875.

Toon, A. (2016) Fictionalism and the Folk. *Monist* 99(3): 280–295.

Toon, A. (2021) Minds, materials and metaphors. *Philosophy* 96(2): 181–203.

Trewavas, A., Baluska, F., Mancuso, S., and Calvo, P. (2020) Consciousness Facilitates Plant Behavior. *Trends in Plant Science* 25(3): 216-+.

Vaesen, K. (2011) Giere's (In)Appropriation of Distributed Cognition. *Social Epistemology* 25(4): 379–391.

van Gelder, T. (1998) The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences* 21(5): 615-+.

van Gelder, T. (1999) What might cognition be, if not computation. In Lycan, W. (Ed.), *Mind and Cognition* (2nd ed.). Oxford: Blackwell.

Varela, F., Thompson, E., and Rosch, E. (1993) *The embodied mind: cognitive science and human experience*. Cambridge, Mass.: M.I.T. Press.

Varoquaux, G., Schwartz, Y., Poldrack, R. A., Gauthier, B., Bzdok, D., Poline, J.-B., and Thirion, B. (2018) Atlases of cognition with large-scale human brain mapping. *Plos Computational Biology* 14(11): e1006565.

Vernon, D. (2014) *Artificial cognitive systems: a primer*. Cambridge, Massachusetts: The MIT Press.

Viola, M. (2016) The Ontological Agenda of Cognitive Neuroscience: Neuroscience as an 'Arbiter' for Psychological Categories (and viceversa). *Rivista Internazionale Di Filosofia E Psicologia* 7(2): 144–165.

Viola, M., and Zanin, E. (2017) The standard ontological framework of cognitive neuroscience: Some lessons from Broca's area. *Philosophical Psychology* 30(7): 945–969.

von Bertalanffy, L. (1969) *General system theory: foundations, development, applications*. New York: Braziller.

von Uexkull, J., and O'Neil, J. D. (2010) *A Foray into the Worlds of Animals and Humans: With A Theory of Meaning*. Jackson: University of Minnesota Press.

Vygotsky, L. S. (1978) *Mind in society: the development of higher psychological processes*. (Cole, M., John-Steiner, V., Scribner, S., and Souberman, E., Eds.). Cambridge, Mass. ; London: Harvard University Press.

Wahn, B., Karlinsky, A., Schmitz, L., and Koenig, P. (2018) Let's Move It Together: A Review of Group Benefits in Joint Object Control. *Frontiers in Psychology* 9: 918.

Wahn, B., Kingstone, A., and Koenig, P. (2017) Two Trackers Are Better than One: Information about the Co-actor's Actions and Performance Scores Contribute to the Collective Benefit in a Joint Visuospatial Task. *Frontiers in Psychology* 8: 669.

Wahn, B., Kingstone, A., and Koenig, P. (2018) Group benefits in joint perceptual tasks-a review. *Annals of the New York Academy of Sciences* 1426(1): 166–178.

Walsh, D. M. (1996) Fitness and function. *British Journal for the Philosophy of Science* 47(4): 553–574.

Walsh, D. M. (2002) Brentano's chestnuts. In Ariew, André, Cummins, R., and Perlman, M. (Eds.), *Functions*. Oxford: Oxford University Press.

Walsh, D. M. (2008) Teleology. In *The Oxford Handbook of Philosophy of Biology*. doi:10.1093/oxfordhb/9780195182057.003.0006.

Walsh, D. M. (2012) Mechanism and purpose: A case for natural teleology. *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences* 43(1): 173–181.

Walsh, D. M. (2013) Mechanism, emergence, and miscibility: The autonomy of evo-devo. In Huneman, P. (Ed.), *Functions: Selection and Mechanisms*. doi:10.1007/978-94-007-5304-4_3.

Walsh, D. M. (2015a) *Organisms, agency, and evolution*. Cambridge: Cambridge University Press.

Walsh, D. M. (2015b) Variance, Invariance and Statistical Explanation. *Erkenntnis* 80: 469–489.

Walsh, D. M., and Ariew, A. (1996) A taxonomy of functions. *Canadian Journal of Philosophy* 26(4): 493–514.

Walter, S. (2010) Cognitive extension: the parity argument, functionalism, and the mark of the cognitive. *Synthese* 177(2): 285–300.

Walter, S., and Kaestner, L. (2012) The where and what of cognition: The untenability of cognitive agnosticism and the limits of the Motley Crew Argument. *Cognitive Systems Research* 13(1): 12–23.

Walter, W. G. (1950) An imitation of life. *Scientific American* 182(5): 42–45.

Walton, A. E., Richardson, M. J., Langland-Hassan, P., and Chemero, A. (2015) Improvisation and the self-organization of multiple musical bodies. *Frontiers in Psychology* 6: 313.

Walton, A. E., Washburn, A., Langland-Hassan, P., Chemero, A., Kloos, H., and Richardson, M. J. (2018) Creating Time: Social Collaboration in Music Improvisation. *Topics in Cognitive Science* 10(1): 95–119.

Washburn, A., Roman, I., Huberth, M., Gang, N., Dauer, T., Reid, W., Nanou, C., Wright, M., and Fujioka, T. (2019) Musical Role Asymmetries in Piano Duet Performance Influence Alpha-Band Neural Oscillation and Behavioral Synchronization. *Frontiers in Neuroscience* 13: 1088.

Weiskopf, D. A. (2008) Patrolling the mind's boundaries. *Erkenntnis* 68(2): 265–276.

Weiskopf, D. A. (2010) The Goldilocks problem and extended cognition. *Cognitive Systems Research* 11(4): 313–323.

Weiskopf, D. A. (2011) The Functional Unity of Special Science Kinds. *British Journal for the Philosophy of Science* 62(2): 233–258.

Weiskopf, D. A. (2017) The explanatory autonomy of cognitive models. In Kaplan, D. M. (Ed.), *Explanation and Integration in Mind an Brain Science*. Oxford: Oxford University Press.

Wheeler, M. (2010) In defence of extended functionalism. In Menary, R. (Ed.), *The Extended Mind*. Cambridge, Mass.: M.I.T. Press.

Wheeler, M. (2011) In search of clarity about parity. *Philosophical Studies* 152(3): 417–425.

Wheeler, M. (2012) Minds, things, and materiality. In Schulkin, J. (Ed.), *Action, Perception and the Brain: Adaptation and Cephalic Expression*. Houndsmill, Basingstoke, Hampshire ; New York, New York: Palgrave Macmillan.

Wheeler, M. (2017) The Revolution will not be Optimised: Radical Enactivism, Extended Functionalism and the Extensive Mind. *Topoi-an International Review of Philosophy* 36(3): 457–472.

Wheeler, M. (2019) Breaking the waves: beyond parity and complementarity in the arguments for extended cognition. In Colombo, M., Irvine, E., and Stapleton, M. (Eds.), *Andy Clark and His Critics*. Oxford: Oxford University Press.

Whitehead, A. N. (1978) *Process and reality*. New York: The Free Press.

Wiener, N. (1961) *Cybernetics, or Control and Communication in the Animal and the Machine*. (2nd Ed.). M.I.T. ; Wiley.

Williams, P. L., and Beer, R. D. (2010) Information Dynamics of Evolved Agents. In Doncieux, S., Girard, B., Guillot, A., Hallam, J., Meyer, J. A., and Mouret, J. B. (Eds.), *From Animals to Animats 11* (Vol. 6226). Berlin: Springer-Verlag Berlin.

Wilson, M. A., and McNaughton, B. (1993) Dynamics of the hippocampal ensemble code for space. *Science* 261: 1055–1058.

Wilson, R. A. (2004) *Boundaries of the mind: the individual in the fragile sciences: cognition*. Cambridge: Cambridge University Press.

Winning, J., and Bechtel, W. (2016) Biological Autonomy. *Philosophy of Science* 83(3): 446–452.

Winning, J., and Bechtel, W. (2018) Rethinking Causality in Biological and Neural Mechanisms: Constraints and Control. *Minds and Machines* 28(2): 287–310.

Winning, J., and Bechtel, W. (2019) Being emergence vs pattern emergence: complexity, control and goal-directedness in biological systems. In Gibb, S. C., Hendry, R. F., and Lancaster, T. (Eds.), *Routledge Handbook of Emergence*. London: Routledge.

Wolf, T., Sebanz, N., and Knoblich, G. (2020) Adaptation to unstable coordination patterns in individual and joint actions. *Plos One* 15(5): e0232667.

Wolf, T., Vesper, C., Sebanz, N., Keller, P. E., and Knoblich, G. (2019) Combining Phase Advancement and Period Correction Explains Rushing during Joint Rhythmic Activities. *Scientific Reports* 9: 9350.

Woodward, J. (2003) *Making things happen: a theory of causal explanation*. Oxford: Oxford University Press.

Woodward, J. (2017) Explanation in neurobiology: an interventionist perspective. In Kaplan, D. M. (Ed.), *Explanation and Integration in Mind and Brain Science*. Oxford: Oxford University Press.

Woodward, J., and Hitchcock, C. (2003) Explanatory generalization, part 1, A counterfactual account. *Nous* 37(1): 1–24.

Wooldridge, D. E. (1963) *The machinery of the brain*. New York [etc.] ; London: McGraw-Hill.

Wright, L. (1973) Functions. *The Philosophical Review* 82(2): 139–168.

Yilmaz, O., and Dupre, J. (forthcoming) *Plant individuality: a physiological approach*.

Zamm, A., Pfordresher, P. Q., and Palmer, C. (2015) Temporal coordination in joint music performance: effects of endogenous rhythms and auditory feedback. *Experimental Brain Research* 233(2): 607–615.