Mindedness: On the Minimal Conditions for Possessing a Mind

Bernardo Aguilera Dreyse

Submitted in partial fulfilment of the requirements for the degree of PhD

Department of Philosophy
The University of Sheffield
September 2013

Abstract

This thesis explores the grounds for justifying the ascription of mentality to non-human agents. In the first part, I set my research within the framework of scientific naturalism and the computational theory of mind. Then I argue that while the behaviour of certain agents demands a computational explanation, there is no justification for attributing mentality to them. I use these examples to backup my claim that some authors indulge in unnecessary ascription of mentality to certain animals (e.g. insects) on the main grounds that they possess computational capacities.

The second part of my thesis takes up recent literature exploring the line that divides computational agents with and without mentality. More precisely, I criticise the proposals put forward by Fodor, Dretske, Burge, Bermúdez and Carruthers. My main argument takes the form of a reductio ad absurdum by showing that their criteria apply to artefacts to which the attribution of mentality is unjustified. Overall, I conclude that even though the views advanced by the mentioned authors help to elucidate the computational grounds that could make the emergence of a mind possible, they do not offer a satisfactory criterion for the ascription of mentality to some computational agents but not others.

In the final part I develop my own proposal for grounding the attribution of mentality. My strategy consists in drawing upon the distinction between personal and subpersonal levels of explanation, according to which properly psychological descriptions have whole-agents as their subject matter, use a distinctive theoretical vocabulary, and are constrained by norms of rationality. After showing that the personal-subpersonal distinction is compatible with a naturalistic framework, I adapt the distinction so that it can be applied to non-human agents, and conclude that it imposes constraints in cognitive architecture that point in the direction of cognitive access, generality and integration.

Acknowledgements

I would like to thank:

Chile's National Commission for Scientific and Technological Research (CONICYT) for funding my PhD.

My supervisors Stephen Laurence and Dominic Gregory for their patience and intellectual advice.

Colleagues from the Philosophy Department who have given me comments and advice at some point during this long process, in particular George Botterill, Ryan Doran, Ivar Hannikainen, Kate Harrington, Bernardo Pino, Katherine Puddifoot, Philipp Rau, Paniel Reyes, Robert Stern and Naoki Usui.

Friends who have been present at different stages during the signifiant part of my life I have spent living in Sheffield, to name just a few, Carmen Anglés, Bernabé Álvarez, Raúl Berríos, Aldo Berríos, Asa Cusack, Paul Dawson, Imara Díaz, Herman Elgueta, Eszter Farkas, Claudia González, María Jesus Inostroza, Daniela Londoño, Héctor Madrid, Timea Pápai, Cristina Roadevin, Claudio Salas, Josecita Sandoval.

And of course, to my family and friends in Chile with whom I have kept in touch during my stance far from home.

Table of Contents

Word Count: 72,820

Preface

One of the most important distinctions we make in our everyday lives is that between minded and non-minded creatures. We describe and explain human behaviour by appeal to mental states, and in virtue of this we approach minded creatures in a way that is notably different from the way we approach non-minded creatures or entities (e.g. with respect to their moral status). Traditionally, having a mind has been regarded as a privilege of human beings and perhaps a few other members of the animal kingdom. Many philosophers followed Descartes's view that non-human animal behaviour can be explained in terms of the same (merely) mechanistic principles that govern mindless machines. It was normally taken as proof of these animal's lack of mentality their incapacity to do things people readily associate with human intelligence: making arithmetic calculations, showing non-associative learning, holding a conversation, etc.

But by the second half of the last century, this view began to change considerably. An important factor behind this change has to do with the advent of computation theory. Through the pioneering work of mathematicians such as Alan Turing and John Von Neumann, computing machines emerged—first in theory and later in fact—as being capable of performing intelligent behaviour[1]. These machines can fly airplanes, carry out surgical procedures, and even outperform human beings in tasks such as arithmetic calculation and chess playing (to name just two somewhat dated examples).

Computation theory has indeed contributed a great deal in our understanding of the mind. For the first time, there was a plausible theory about how intelligent behaviour could be explained in physicalist terms. It soon became widely acknowledged that important aspects of mentality—at least with regards to thinking and reasoning—could be explained by appeal to computation. This wide consensus gave rise to the so-called computational theory of mind (explained in chapter 1), which, according to

[1] I leave aside for the moment whether the intelligence of a machine is intrinsic or derived from its designer. The point here is just that some (embodied) machines can do things—by themselves—we normally qualify as intelligent.

perhaps the most influential current philosopher in the field, is "by far the best theory of cognition we've got that's worth the bother of a serious discussion" (Fodor, 2000, p. 1).

Part of the progress made by good theories is that they give rise to new and interesting questions. What kind of computer is the mind? How do we tell whether a certain computing machine can think? If we accept that the mind is a real and objective natural phenomenon (as I do in this thesis), we should expect concrete answers to these questions. And posing these questions is important, for at least two reasons. One is that technology advances fast, which means that sooner or later we will have to face the question of whether or not a certain robot can be said to have mentality. Secondly, research on animal cognition has shown that even animals (such as insects) that were previously considered simple, do complex computation. Given that we have a mind by virtue of being some kind of computer, this opens up the question of what is special about us, such that we have mentality while other computing animals do not. The main goal of this thesis is to tackle this issue, and work towards determining the minimum conditions for possessing a mind.

In this respect, one common way to proceed is by following what Lurz (2009) calls a *bottom-up approach*, which begins with taking what looks to be an intuitively plausible ascription of mentality at face value, and then proceeds with the development of a theory of behavioural explanation for non-human agents that includes psychological terms. This is the case (or so I argue in chapter 2) of philosophers who have considered it plausible to ascribe "simple minds" to animals on the basis of their possession of rather complex computational mechanisms linking their information-gathering and action systems together. However, I contend that this approach is problematic, since it rests on questionable assumptions about what distinguishes mere computational agents from genuine mental agents.

My strategy (put forward in chapter 7) is based on what we might call a *top-down approach*. Instead of furnishing computational explanations of animal behaviour with psychological notions, I take psychological explanations of human behaviour as the paradigm for judging whether other computational agents have minds. More precisely, I spell out my approach in terms of what is known as *personal-level* explanation. To avoid anthropomorphic concerns related to defining the mind in terms

of persons, I develop an *agent level* of explanation, which attempts to abstract from human-specific features by focusing just on the essential aspects of the personal-level approach, so as to adapt them to explaining the behaviour of animals and even machines. By means of this agent-level approach I attempt to justify the ascription of mentality to agents that can be properly described within this explanatory framework, and reveal the minimum conditions a computing system requires for possessing a mind.

Before going into an overview of my thesis, I find it important to make a clarification regarding the notion of mentality at stake here. In addition to thought and reason, the mind is normally understood as involving conscious states, viz. states that have a distinctive qualitative character, in the sense that there is something that "it is like" to have them. However, I set these states aside from the present inquiry. This does not mean to say that consciousness is not an important aspect of mentality, and I admit that (arguably) any complete account of the nature of the mind has to somehow address this issue. Nevertheless, following many philosophers persuaded by the computational theory of mind, I assume that consciousness is not an essential aspect of thought, and that important progress can be made on the nature of mental representation and thinking without addressing what it is like to have conscious states.

Here is how I proceed. Chapter 1 sets forth some assumptions and theoretical background relevant to this thesis. It introduces a naturalistic framework according to which we are ontologically committed to the entities described by our best scientific theories. I assume that amongst those theories is the computational theory of mind. This opening chapter also elaborates the idea that complex phenomena such as the mind can be described from the viewpoint of hierarchical levels of explanation. Following a common tripartite distinction, I characterise them as the physical, the computational, and the psychological level.

Chapter 2 develops and defends the idea that the computational level of explanation picks up an autonomous natural domain—a domain of computational agents which are not necessarily endowed with mentality. Then I take issue with some authors who defend the thesis that some animals have mentality on the basis that they possess certain complex computational abilities. I contend that these authors overlook the autonomy of the computational level, and fall into the false dilemma of assuming

that behaviour has to be explained either from the physical or the psychological level. Instead, I propose that it is possible to regard some animals and machines as mere non-mental computational agents.

Chapters 3 to 6 take up recent literature exploring the line that divides computational agents with and without mentality. Chapters 3 & 4 tackle the informational approaches advocated by Jerry Fodor and Fred Dretske, which have put forward conditions under which information coded by computing systems could become genuine mental symbols. My main arguments against them take the form of a reductio, by showing that their criteria apply to artefacts to which the attribution of mentality is unjustified. In addition to discussing the particular views of the mentioned philosophers, these two chapters provide an overview of the standard computational account of perception, to which I return in subsequent chapters.

Chapter 5 critically reviews Tyler Burge's recent proposal about the minimum conditions for having mental symbols. Burge develops a teleological approach to perceptual systems, as a way to account for their capacity of generating basic symbolic structures that demand a psychological explanation. I object to his proposal that its overall teleological picture of the mind is problematic, and that it ends up drawing the line for having mentality too low.

Chapter 6 addresses the views of José Luis Bermúdez and Peter Carruthers. They advance forms of symbolic processing and cognitive architecture that, according to them, deserve to be described in psychological terms. I contend, on different grounds, that their views do not offer a satisfactory criterion for distinguishing computational from mental symbols, and neither for telling apart mental from non-mental computational architectures.

Chapter 7 is where I present my own hypothesis on the correct way to draw the line that separates computational agents with and without mentality. It draws upon the distinction between personal and subpersonal levels of explanation, according to which properly psychological descriptions have whole-agents as their subject matter, use a distinctive theoretical vocabulary, and are constrained by norms of rationality. After showing that the personal-subpersonal distinction is compatible with a naturalistic framework, I adapt the distinction so that it can be applied to non-human agents, with

some considerations about the constraints it imposes on the ascription of mentality from a computational viewpoint.

Chapter One

Computational Psychology

1.0 Introduction

The main goal of this thesis is to explore the minimum conditions for having mentality and articulate a way to draw a line between those agents that have a mind and those that lack it. For the purposes of this thesis I shall take for granted three tenets that underlie the debate over the nature of the mind: scientific realism, the computational theory of mind, and informational approaches to representation. My goal in this opening chapter is to introduce those tenets and set up the theoretical background for the rest of this thesis.

I begin by introducing the related views of naturalism and scientific realism, as a way to establish the general metaphysical and epistemological foundations of the computational theory of mind. I then present the computational theory of mind itself, discussing first its computational and then its representational component. Finally, I elaborate the idea that explanations of behaviour can be formulated from different explanatory levels.

This chapter is mainly introductory and devoted to discussing some relevant background. In the second chapter I will present my first positive view, which is that the computational level of explanation maps onto an autonomous natural domain.

1.1 The Place of the Mind in our Scientific Worldview

Our modern understanding of the world has been deeply shaped by the emergence of science in the seventeenth century. Since then, everything in the universe began to be understood as part of a common natural order governed by deterministic laws and science became the dominant method for unveiling this natural order. Among the natural phenomena in need of explanation lies behaviour. Humans and other

animals, plants, and even robots move and perform a variety of behaviours[2]. But it is clear that not any movement counts as behaviour. It would be odd to describe the movement of a rock that sinks in water, or the motion of a planet through its orbit, as behaviour since these entities are not actually doing those actions. They are just passive respondents to external forces and nothing inside them plays an active causal role in determining their movements. Thus, I will follow Dretske (1988) in defining behaviour as the sort of activity that can be classified as the result of internal processes, given that those internal factors can be credited as primary causes of an agent's actions.

Psychology is one of the sciences that study behaviour, however it is concerned with the particular class of agents which have minds. In those agents at least some of their behaviours are caused by inner mental states such as beliefs, desires and intentions, and are governed by reasons. This sort of behaviour—often called *intentional behaviour*—is the proper domain of psychological explanation. So, if we take seriously the common slogan that science strives to "carve nature at its joints" then we could argue that psychology would correspond to the science that attempts to carve the natural domain of mental agents at its joints; the way it does so is by identifying mental states and putting forward explanations that describe them as the internal causes of their actions.

In order to be recognised as a scientific discipline one of the most pressing issues for psychology is to vindicate the use of mental states as part of its theoretical apparatus. After all, we cannot directly observe or measure the mental states of others, and talk about the mind has historically been linked to religious and dualistic conceptions that are of dubious scientific import. So, it will be important for our characterisation of psychology to explain how it can be compatible with two viewpoints that lie at the metaphysical and epistemological foundations of contemporary psychology: naturalism and realism. For the purposes of this thesis and following the prevalent viewpoint in philosophy of psychology, I will take both principles for granted (though in the second chapter I address some antirealist positions regarding computational states). I present those in turn.

---

[2] But as Dretske (1988) notes movement is not a necessary condition for having behaviour; even staying still can count as behaviour insofar as it is the product of inner processes. See the rest of the paragraph for a more precise definition of behaviour.

1.1.1 Naturalism

Naturalism is the view that an adequate philosophical account of the world has to be given in terms of states and processes occurring in the natural causal order. This view is normally committed to the ontological claim that everything in the world is physically constituted and that only physical entities can participate in causal relations or affect the natural world (Papineau, 2009). Most contemporary philosophers and scientists consider themselves naturalists and regard psychology as the discipline that studies the place of the mind in nature. For example, in their textbook on philosophy of psychology, Botterill and Carruthers (1999) write:

> According to *naturalism* human beings are complex biological organisms and as such are part of the natural order, being subject to the same laws of nature as everything else in the world. If we are going to stick to a naturalistic approach, then we cannot allow that there is anything to the mind which needs to be accounted for by invoking vital spirits, incorporeal souls, astral planes, or anything else that cannot be integrated with natural science. (p. 1)

As the quote suggests, naturalists also pursue the aims and methods of science to obtain knowledge of the world. They see the study of mental states as the outcome of scientific research in the same way as biology studies the structure of the cell or physics reveals the constitution of atoms, and exclude from the vocabulary of psychology any states that cannot be accounted for through the methods of science. As a consequence of its commitment to science, naturalism also recognises physics as the most fundamental of the sciences and includes the view that physics is causally complete within its domain, viz. that every physical event has a physical cause and is subject to explanation in terms of basic physics (hereafter, just "physics").

But the completeness of physics has to be distinguished from the stronger claim that physics can explain everything. It has become customary in science to accept that some sciences other than physics—often called *special sciences*—can have their own explanatory domains, which describe states and processes in a way that cannot be reduced to explanations of physics (Fodor, 1974). This is because their explanations are formulated at a higher level of abstraction, which captures generalisations that range over many different physical descriptions and therefore would otherwise be missed from the viewpoint of physics. But the special sciences are still compatible with the completeness of physics insofar as their explanations invoke states which have a

physical constitution and its processes are not in conflict with the principles of physics. Typical examples of special sciences are biology and psychology (see Sterelny, 1990; Crane, 2001).

The way psychology qua special science relates to physics is a complex issue. The basic idea, however, is that all facts described by psychology are somehow determined by facts that fall under the domain of physics. This is normally put forward by appeal to the term *supervenience*. Psychology is supposed to supervene on physics in the sense that two people cannot differ in their psychological states without also having a difference in certain relevant physical states involving them[3]. How the notion of mental state can be articulated in a way that does not refer to physical properties but at the same time supervenes on them will be explained in section 1.2.1, when presenting functionalism.

1.1.2 Scientific Realism

As noted in the previous discussion, naturalists give an authoritative role to science. This is often associated to *scientific realism* which is the metaphysical position that the states and processes described by our scientific theories do exist and that the theories themselves are at least approximately true (Fine, 1999). According to scientific realism, the authoritative role of science is also epistemological, in the sense that we are justified to adopt a positive epistemic attitude towards the theoretical elements of our best scientific theories. So provided the success of psychology as a science, through scientific realism we can justify the belief that there is an objective, observer-independent domain of mental agents.

But then it is natural to wonder how we should measure the success of psychology and be justified to choose it from alternative theories that explain behaviour. The best way scientific realism has to deal with this issue is to focus on their explanatory virtues. The idea is that explanation can provide an additional evidential

[3] It should be noted that strictly speaking supervenience only implies that mental states covary with physical states and not that the existence of the former depends on the latter, which makes supervenience compatible with property dualism (Kim, 2006). But following the standard usage of the term in the philosophy of psychology, I will understand supervenience as entailing physicalism and then assume that supervenience involves a relation of dependence of psychology on physics.

standard for choosing between alternative theories, a view that is often called *inference to the best explanation* (Day & Kinkaid, 1994; Lipton, 2004). Then the belief in psychological theories is warranted because they provide greater predictions and understanding of human behaviour, and scientists argue for the existence of mental states in the same way as with other unobservable entities such as electrons or black holes. The argument for stating that those entities are really out there in the world is that they are part of the ontology implied by our most successful theories. Returning to the context of psychology, psychological explanations involving mentalistic concepts prevailed over alternative theories precisely because it proved to a be more successful explanation of human behaviour. Let me illustrate this with an example.

From an historical viewpoint, psychology only became consolidated as a scientific discipline in the 1950s after what became known as the "cognitive revolution" (Miller, 2003). As it is normally presented in textbooks of psychology and cognitive science, before that time the dominant approach to explain behaviour was non-mentalisitic and known as behaviourism (Bechtel, Graham & Abrahamsen, 1998). Behaviouristic explanations are typically restricted to observable patterns of stimuli and behavioural responses, and deem talk about mental states unscientific due to their subjective and unverifiable source in introspective reports. For example, behavioursitic theories of language claimed that children learn their language basically through a process of operant conditioning, where their spontaneous linguistic behaviour is positively or negatively reinforced by adults.

In a famous review of Skinner's version of this theory, Chomsky (1959) pointed out to the insufficiency of this behaviouristic model to explain the children's ability to understand and produce an indefinite number of sentences on first acquaintance. Since many of the sentences children understand and produce had never been uttered or heard before, there is no way to explain their linguistic capacities by appeal to prior reinforcement. Chomsky (1965) then put forward an alternative psychological theory that appeals to inner mental mechanisms, in particular the possession of innate linguistic rules that restrict the possible grammatical structures the child could learn and produce. In this way, Chomsky was able to explain how children could acquire the ability to generate new grammatical sentences in the language they are learning. This is not the place to get into the details of Chomsky's theory of language acquisition, but to present

it as an example of how mentalistic theories prevailed over previous ones due to their explanatory virtues.

1.2 The Computational Theory of Mind

So in the course of the above mentioned cognitive revolution psychologists began to study and explain the behaviour of human agents by appeal to theories involving inner mental states and processes, and showed how this could be carried out in a scientifically respectable way. A key aspect of this cognitive turn has been the interdisciplinary approach to the study of the mind. Psychology began to absorb developments made by philosophy, computer science and information theory, among other disciplines. All this gave shape to the computational theory of mind (henceforth just "CTM") which can be regarded as the received view in current theorising about the mind and the scientific explanation of human behaviour.

According to CTM the mind is a kind of digital computer, which is to say, a discrete-state device that stores symbolic structures and manipulates them according to syntactic rules (Horst, 2009). As a preliminary description computers typically have an input layer and a memory, from where symbolic structures can be encoded or retrieved respectively, and the machinery required for performing certain fundamental operations over them. These operations can produce an output as a function of its inputs, a function that can be specified by an algorithm which is a sort of recipe that specifies step by step how to manipulate the symbols in order to obtain the desired output. For expository reasons it is useful to present CTM as having two parts, a syntactic and a representational component, however when it comes to real computers both parts are not totally independent as I will explain later. The syntactic component consists in the functional architecture of the system that determines how operations over symbols can be performed. The representational component is concerned with the content of those symbolic structures, i.e. what makes possible for them to represent or stand for other things. In the following sections I present them in turn.

An important assumption behind CTM is that a significant part of mental processes consist in carrying out inferences. This is important because it is by

implementing the logical structure of inferences[4] that computers are normally regarded as an appropriate model of how the mind works[5] as I will explain with more detail in section 1.2.2. The assumption that the mind is inferential is plain for the case of practical reasoning. For example, imagine that when Peter is about to leave his house he looks at the window and finds out that it is raining outside and then picks up an umbrella. A natural way to explain his behaviour would be to ascribe him the belief that it is raining outside, the desire to keep dry, and the capacity to carry out a practical syllogism where the belief and the desire act as premises for the practical conclusion of picking up his umbrella. Importantly, according to CTM the mind is also supposed to be inferential at the level of deeply unconscious processes such as those involved in language learning or perceptual processing. For example, according to Chomsky's theory sketched above, children are supposed to infer the right linguistic constructions they can use in their language by taking both the linguistic expressions they hear and their inner knowledge of grammar as premises.

1.2.1 The Role and Realiser Distinction

To understand how computation theory applies to psychology it is important to see how the mind can be described from a functionalist viewpoint. Broadly construed, a functionalist approach to psychology says that mental states are constituted by their function, or causal role in the overall cognitive system in which they are part, in particular their causal relations to sensory inputs, other mental states and behaviour outputs (Putnam, 1975; Fodor, 1968). Through this functionalist framework psychology can formulate causal generalisations relating mental states and their effects in behaviour, in a way compatible with the standard model of scientific explanation.

The causal role of a mental state is also called its *job description* since what individuates the state is the causal work it is supposed to do in the system, instead of

---

[4] It should be noted that CTM does not require all the inferences carried out by the mind to have a logical form, i.e. it allows them to be non-demonstrative, inductive-like, inferences. The basic point is, though, that they have a structure and that this structure can be implemented in the functional architecture of a computer.

[5] Searle (1992) calls "strong AI" the view that all that there is to having a mind is having a program that mirrors the inferential structure of the mind. CTM need not be strong AI, though, insofar as it could admit that some important part of behaviour is not mediated by inferential processes. These behaviours might be explained by non-computational explanatory frameworks, such as the physical level (see section 1.4.3).

how it is physically constituted. The fact that causal role of mental states can be formulated with relative independence from its physical constitution gives rise to the distinction between the causal role of a mental state (i.e. the cognitive job it performs) and its realiser (i.e. the physical structure that actually occupies that role, typically a state of the brain) (Levin, 2010).

To illustrate this functionalist approach to psychology, let us return to the case of Chomsky's theory sketched above. The author argued that explanations of language learning and comprehension involve the possession of inner knowledge of grammar. Among them are phrase structure rules that govern how the constituents of a sentence such as noun phrase and verb phrase are structured. They perform the job of categorising the linguistic input a child hears in order to restrict the class of possible languages the child could learn. According to the functionalist, those rules are realised by mental states defined in terms of their causal role in causally mediating between linguistic stimuli and behaviour, viz. the role of categorising the linguistic input and constraining the grammatical constructions that could be produced.

As happens with other sciences, the non-logical vocabulary of psychological theories refers to natural kinds, which correspond to the natural entities in virtue of which scientific theories describe law-like regularities. They are part of the language, so to speak, couched by scientific predicates. In many sciences, natural kinds can be identified with some physical property, such as a certain molecular or atomic configuration. However, since psychology is a functional theory its predicates are about entities individuated by their causal roles and not by their physical realisers. Therefore, psychological theories refer to functional kinds, which can in principle be realised in an indefinite number of physical substances insofar as they occupy their characteristic causal roles (Sterelny, 1990). For example, the causal role performed by the linguistic rules alluded to in the last paragraph could in principle be implemented in a computing machine distinct from the brain, insofar as the machine is capable of reproducing the complex network of input and output relations that characterises the child's cognitive system. If that is possible, then the computer would really instantiate the same mental states of the child and be capable of learning a language, even though its physical constitution is radically different from the child's brain.

The fact that the roles of mental states can be multiply realised has the advantage for psychology that makes it relatively independent from more fundamental sciences such as neuroscience or physics. Psychologists can then focus on the functional architecture of the mind and study mental states as functional kinds, defined in terms of their casual roles in psychological theories and not in terms of their intrinsic physical properties. This is a crucial step towards the vindication of psychology as a special science, with its own vocabulary and generalisations, which cannot be reduced to the vocabulary of more basic sciences.

1.2.2 The Syntactic Component

The most characteristic aspect of computer systems is that they have a syntactic structure, which basically consists in a functional description of the inferential processes that mediate between the inputs and outputs of the system. In fact, computers can be defined as automatons that can map transitions between its inputs, outputs and internal functional states. When focusing on the syntactic component of the mind CTM abstracts from the representational component of symbols and develops a purely functional or formal description of the inferential operations performed over them. There are two aspects of the syntactic component of computer systems that make them suitable for implementing the inferential architecture of the mind. One is that computers are universal machines and the other that they have actually been realised in physical devices. I explain these in turn, and then discuss some implications they have for theorising about the mind.

As mentioned in 1.2, a significant part of the mental processes that explain behaviour have the characteristic that they are inferential. Another typical aspect of mental processes (at least for the case of intelligent creatures) is that they are flexible and adaptable, in the sense that they can deploy different inferential procedures to deal with new environmental circumstances. After all, the main reason why chess-playing machines are not very intelligent is that they cannot do anything beyond playing chess. In contrast, the mind is—at least to some degree—a multi-purpose system of inference, or as is normally put, they are *universal machines*. This idea can be tracked back to Alan Turing's seminal work on computation. In his words:

> [The] special property of digital computers, that they can mimic any discrete state machine, is described by saying that they are *universal* machines. The existence of machines with this property has the important consequence that, considerations of speed apart, it is unnecessary to design various new machines to do various computing processes. They can all be done with one digital computer, suitable programmed for each case. It will be seen that as a consequence of this all digital computers are in a sense equivalent (Turing, 1950, pp. 441-442)

I shall say more about what precisely is understood by a universal machine below and in the next chapter, but for present purposes the important idea is that computers are universal machines in the sense that they can in principle be programmed to compute any algorithm (Pylyshyn, 1980; Haugeland, 1981; Copeland, 1993). This is possible because computers have the machinery required to perform a fundamental set of operations over its inner symbolic structures, such as to copy, write or delete them, and to automatically carry out series of operations as specified by a program (which is a series of instructions or algorithms stored in the memory of the computer). Turing himself is responsible of making this claim of universality plausible by creating what has become the paradigm of a computer: the *Turing machine* (Turing, 1937). Briefly, this is a mathematically characterised computing machine consisting in an input layer from where it can recognise symbolic structures, and capable to perform two types of operations depending on its initial state and the input symbol it recognises. Those operations correspond to the output of the device and are (a) move into a new state and/ or (b) erase the existing symbol (if any) and write a new one (on an imaginary tape). Which of those operations the machine actually performs will depend on the program it has stored in its memory (called the *machine table*).

It is easy to show how a Turing machine can perform simple arithmetic procedures such as addition using a binary symbolic system (see e.g. Crane, 1995). But the essential point behind Turing machines is that they can serve as proof of the universality of computer systems since any algorithmic procedure can in principle be executed by a computer, given enough time, tape and memory to it.[6] And since mental processes have the logical structure of inferences and the syntax of those structures can be specified by an algorithm (i.e. translated into a program), it becomes clear that computer systems are then capable of implementing the inferential structure that characterises the mental processes of intelligent creatures.

---

[6] This is also known as the *Turing-Church thesis*, after Alonzo Church made an equivalent proposal.

The second aspect of computer systems that make them suitable for implementing the inferential architecture of cognition is the well-established fact that their syntactic component can be realised by physical entities. For example, computing artefacts can perform the fundamental operations of a universal computer and mechanically pass through one physical state to another mirroring the steps of an algorithm. And the realisation of computers is of course not restricted to electrical circuits, since an indefinite amount of physical devices such as pipes of water or mice in traps could in principle be used for making a computer. Indeed, groups of neurones have also been described as instantiations of basic computational operations (McCulloch & Pitts, 1943). This makes plausible the claim that the syntactic or functional organisation of computers can be studied in abstraction from their physical constitution but at the same time can be implemented in concrete physical systems such as the brain.

The idea that the inferential structure of mental processes capture the core of human intelligence and that this can be decomposed into algorithms and implemented in a computer was forcibly put forward by Newell and Simon (1981) with their *symbol system hypothesis*:

> The Symbol System Hypothesis: A physical symbol system has the necessary and sufficient means for intelligent action. (p. 41).

The authors talk of "physical" symbols to emphasise the idea that they are focusing on the syntactic component of CTM, which corresponds to the shape or form of the symbols. Symbols are supposed to be manipulated purely in terms of their formal dimension, in the same sense as the shape of a key determines what lock it will open. As I will discuss in the next section, the representational properties of the symbol are left aside to focus on the logical structure of intelligence, however they are assumed to be present in real autonomous computing systems.

The modal claim implied by the symbol system hypothesis should be noted, viz. that any possible form of intelligent action has to be based on algorithmic processing over symbols. This can be challenged, however, by proposing that some intelligent human behaviour is generated by mental processes that do not have the structure of an algorithm and therefore cannot be run by a Turing machine. It can also be argued that some mental processes are actually carried out over analogue information that thus lack

the discrete digital structure typical of computational symbols. But for present purposes we do not need to deal with these controversies. We can just characterise the symbol system hypothesis as stating sufficient means for intelligent action, and thus claim that at least a significant part of the mind corresponds to the sort of symbolic processes that computers actually can implement. The point is just that a computing device, if properly programmed, can in principle reproduce the logical structures that underlie characteristic human mental processes.

Before passing to the next section, a brief note about the universality of computers. As I mentioned above, computers like Turing machines have the basic machinery required to execute any algorithm and in this sense simulate the functional architecture of any other computer machine. This makes computers universal, general-purpose, devices. But even though this is true from a theoretical viewpoint, when it comes to concrete computers instantiated in the real world their universality is constrained in many ways (Newell, 1960). An obvious constraint relates to the physical capacities of the machine itself. For example, a machine may lack the memory, velocity, and output mechanisms required for performing any task. Another is concerned with the way computers interact with their environment. Even though from a purely syntactical of formal viewpoint many computers can be regarded as performing the same algorithm, they might actually be processing different kinds of information and performing different tasks. Computers capable of behaving in the world in an autonomous way are constrained by the nature of their input-output layers. And the same happens with the brain. Its different regions are specialised to process domain-specific information and solve problems proper to that domain (Churchland & Sejnowski, 1999). Thus it is useful to distinguish between abstract, syntactically specified computers such as Turing machines, which are genuinely universal, from instances of computation in computing machines that perform tasks in the real world. I shall return to this issue in the second chapter.

1.2.3 The Representational Component

Computing machines have a syntactic component that can mirror the logical architecture of inferential mental processes. But as mentioned earlier, computation also

involves the manipulation of symbolic structures, which have a representational or semantic dimension that goes beyond their syntactic structure. The point is nicely put by Haugeland (1981) in the following passage:

> So, formal (symbol) tokens can lead two lives: *syntactical* (formal) *lives*, in which they are meaningless markers, moved according to the rules of some self-contained game; and (if the system is interpreted) *semantic lives*, in which they have meanings and significant relations to the outside world. (p. 22)

As the quote says, symbols have a semantic dimension since symbols are essentially entities that bear reference relations to something else. Imagine you are driving a car aiming to get to the airport and at some point you find a signpost in the route showing an airplane. The airplane is a symbol that is referring to an airport, and in the present context we might even say that means "airport nearby". But what the signpost means is, of course, conventional. It depends on us—participants of a symbolic community—to be interpreted as a symbol that represents anything at all. In contrast, the symbols we possess in our minds (thereafter "mental symbols") do not depend on anyone else, apart from us, to have meanings. As it is often put, they have *intrinsic content*, viz. one that is not derived from the minds of others. Mental symbols are those that have their contents intrinsically, and to explain how they get their contents is perhaps the central project in the philosophy of mind.

However CTM is sometimes labeled the "symbolic approach" to the mind, CTM theorists often abstract from the representational dimension of the mind and focus on its syntactic component, mainly for methodological reasons. Their central motivation is that the syntactic structure of mental processes can straightforwardly be implemented on computer devices and in this way the algorithmic architecture of the mind can be explored and tested empirically, without dealing with the complexities related to what symbolic structures actually represent. However, most defenders of CTM recognise that a complete account of cognition also has to deal with its representational or semantic component. This is because computational models are supposed to mirror inferences, and inferences are by definition carried out over symbols with representational properties. As Fodor (1975) notes:

> To use this sort of [CTM] model is, then, to presuppose that the agent has access to a representational system of very considerable richness. For, according to the model, deciding is a computational process; the act the agent performs is the consequence of computations

> defined over representations of possible actions. No representations, no computations. No computations, no model. (p. 31)

Therefore, apart from specifying the syntactic architecture of the mind CTM needs an account of how mental symbols could bear reference relations to things other than themselves. But this is not an easy task for CTM for the reason that computation theory itself, as can be deduced from the abstract notion of a Turing machine, takes symbols as formal structures and therefore is essentially silent or agnostic about the contents of symbols (Cummins, 1983). In other words, it presupposes that symbols do have an interpretation, but does little to explain how symbols could be endowed with intrinsic representational contents. Computation theory has to be supplemented with something else.

A convenient way of spelling out what is this "something else" is in terms of causal relations with the environment. As I mentioned at the beginning of this section, symbols are essentially entities that bear reference relations to something other than themselves. This suggests that, for example, a symbolic structure used to think about rabbits should bear some sort of relation to rabbits, or as Fodor (1990) says, for there to be a relation "something has to happen in the world" (p. 99). And the most common way to specify this relation is by saying that it is causal. This should come as no surprise at this point, given our commitment to scientific realism. Since what justifies the postulation of any theoretical entity is that it plays a role in our scientific theories, and given that scientific explanations basically consist in specifying the causal relations the govern the natural order, to be a realist about representational relations is close to saying that those relations have to be causal. In the next section, I explain a casual account of referential relations called *informational approach to representation*.

1.3 Informational Approaches to Representation

Informational approaches to representation can be regarded as stemming from the empiricist tradition in philosophy, which starts from the rather obvious premise that we obtain knowledge about the world by getting information about external objects through our senses. Information theory has attempted to refine the notion of information in the way of an objective commodity that can be generated and transmitted. It has

developed along with computation theory offering ways of understanding how computing systems such as the mind can pick up and process information about the environment (Adams, 2003).

A central claim of information theory in the present context is that information is a more fundamental notion than representation, and a precursor of semantic content. One way to see the link between information and representation is by following Grice (1957) in noting that there is a sense of "mean" used in expressions that describe a natural relation between two states. For example, in "tree rings mean the age of the tree" the tree rings are supposed to be a natural sign or indicator of the age of the tree. Grice saw that there is relation between this sense of the term "mean" and the sense normally used to describe the semantic properties of natural language, as in the expression "with 'tree' he meant the perennial woody plant". To distinguish it from the former, the author called this second sense *non-natural meaning*. For present purposes, what is important of this distinction is that natural meaning can be regarded as an objective phenomenon that does not depend on people's thoughts or conventions to exist, and thus can serve as a precursor of non-natural meaning without presupposing it.

But what is precisely natural meaning? An influential way to deal with this question has been through the notion of *environmental information* (Floridi, 2011). According to Dretske (1981) a system carries information (i.e. environmental information or natural meaning) about another when there is a nomic relation between them, in the sense that physical differences present in the former reliably covary with physical patterns of the latter[7]. So in a way analogous to natural meaning, a system is said to bear information about certain environmental property when some property of the system covaries or responds selectively to stimuli with that environmental property. Importantly, environmental information is not restricted to *natural* systems but also applies to artefacts. For example, the mercury column of a thermometer carries

[7] The causal connection is important since it makes information compatible with the naturalist constraint that everything that exist must be engaged in causal interactions with other entities of the natural order. And reliability is important since it rules out accidental correlations. It might turn out that the only person that rings my bell is the postman, but that would not make the ring about the postman. The relation between the postman and the ring is accidental, and nothing ensures its reliability. Environmental information, instead, is reliably caused by the entities it bears information about. On its strongest form, this reliable relation is nomological, i.e. mediated by a natural law, however more modest forms of reliability have been proposed.

information about the temperature in the room, and the end of a compass needle carries information about where is the north magnetic pole.

This is a fundamentally relational definition of information, where one system can be regarded as the source that generates the information and the other as the receiver of it. This connects with information theory that traditionally defined information in quantitative terms as a measure of the reduction of possibilities, viz. when we got more information about a source we then become less uncertain about how the source is (Floridi, 2011). This is why Dretske (1981) also characterises the possession of information—such as that certain source is F—as the conditional probability that the source being F is 1, thus excluding the possibility of the source being non-F.

But as Dretske himself notes, information theory thus understood is more concerned with the transmission and quantification of information but not with its content, let alone with figuring out how information-bearing states could become symbols with representational properties. A natural suggestion that comes from CTM is that there is a flow of information from the environment to cognitive systems, and that information can become symbolic representation through a process of encoding algorithmic processing that ends up with information packed in a format suitable for computation (see Dretske, 1981; Stalnaker, 1984; Fodor, 1987, for alternative proposals of how this process could take place).

This is not the place to review the different informational approaches of representation. In this thesis I shall rather assume the basic framework of informational approaches (presented in more detail in chapters 3 & 4) and explore how different notions of symbolic representation could be distinguished by adopting different explanatory levels. An important thing to note in this respect is that not all information-processing systems need to have the capacity to compute symbolic structures. Simple thermostats do not, for example (see chapters 3 and 4 for discussion). But furthermore, I shall argue that not all computers have the capacity to compute mental symbols. There are robots which can be described as genuine computers and autonomous information processing devices, even though these lack any form of mentality. But to discuss these topics we will have to wait until the next chapter. I will now finish this chapter by

explaining how behaviour can be explained from different explanatory levels, and in this way the mind be conceived as having multiple levels of organisation.

1.4 Levels of Explanation and the Scientific Study of the Mind

According to a standard conception (Glymour, 1999), scientific theories consist in a group of predicates formulated in the vocabulary of the respective science, that describe generalisations or law-like regularities concerning certain natural phenomenon. This is normally framed in terms of the deductive-nomological model of explanation, where explanations consist in subsumption of events under natural laws, and supplemented by a causal account of explanation, which specifies the causal mechanisms that contribute in bringing about the phenomena under study (Salmon, 1989). Additionally, it has also become customary in science to study complex systems by appeal to multiple levels of analysis, each picking up natural domain. As Fodor and Pylyshyn (1988) observe:

> It seems certain that the world has causal structure at very many different levels of analysis, with the individuals recognized at the lowest levels being, in general, very small and the individuals recognized at the highest levels being, in general, very large. Thus there is a scientific story to be told about quarks; and a scientific story to be told about atoms; and a scientific story to be told about molecules .... ditto rocks and stones and rivers ... ditto galaxies. And the story that scientists tell about the causal structure that the world has at any one of these levels may be quite different from the story that they tell about its causal structure at the next level up or down. (p.5)

Each "story" corresponds to a scientific level of analysis, with its own explanatory vocabulary and laws. In principle, explanations at each level are supposed to be autonomous because they capture a genuine level or organisation in nature, which would be missed if described from lower levels. It is in this sense that psychology is an autonomous science, given that its functional characterisation of the mind can be formulated with rather independence from its physical constitution, as I explained in the previous sections. So according to the standard model of multiple levels of analysis, complex systems such as the mind are members of many natural domains, each describable from a particular explanatory level. Furthermore, levels are hierarchically structured, in the sense that the processes of each ascending level are being implemented or realised by the processes of the next level down (McClamrock, 1991).

Since the advent of cognitive revolution the mind has begun to be analysed in terms of levels of organisation (Marr, 1982; Pylyshyn, 1984; Sterelny, 1990). Though using different terminologies, theorists generally distinguish three levels of explanation for the mind. At the top there is a *psychological level* at which we describe the mental representations and reasoning processes that cause behaviour. The next level down is the *computational level*, at which we specify the data-structures and computations that underlie mental processes, and at the base we have a *physical level* which describes the mind directly in terms its physical structure (I shall describe each level with more detail below). Behind this analysis of levels is the assumption that the mind can be studied by postulating internal functional states, that mediate between perception and action. Both the psychological and the computational levels adopt a functional characterisation of the mind, and thus formulate their explanations at a higher level of abstraction that is rather independent of the physical mechanisms that implement them.

This independence should not be overestimated, though. Each level is relatively autonomous in terms of their explanatory vocabularies and generalisations, but as previously noted each describes a natural domain that depends on more fundamental domains to exist. So in the case of the mind, the domain described by the psychological level is realised by the domain of the computational level, and the computational domain is then realised by the physical domain. It is also important to note that this ontology of levels is compatible with a monist metaphysics as well as with the generality of physics as an account of the natural world. The fact that computational and psychological explanations treat their own explanatory domains by using concepts and generalisations that cannot be, or does not need to be, expressed by the vocabulary of physics, is compatible with the generality of physics insofar as they denote entities which have a physical constitution and its processes are not in conflict with the principles of physics (Crane, 2001).

1.4.1 The Psychological Level

For both commonsensical and scientific psychology there is a level of explanation that applies to creatures endowed with minds, and describes them as rational thinkers capable of engaging in purposeful behaviour. Let us return to our

previous example of intelligent human behaviour. Peter is about to leave his house, he looks at the window and finds out that it is raining outside and then picks up an umbrella. A psychological explanation makes it intelligible why Peter acts in this way by ascribing to him mental states with the form of propositional attitudes, such as beliefs and desires, constrained by principles of rationality. This typically starts by attributing to him the desire to keep dry, beliefs about the environment acquired though perception (such as the belief that it is raining) and instrumental beliefs about how the desire might be satisfied. In the present example, the psychological law that underlies Peter's behaviour would be: if someone wants *p* and believes that by doing *q* he will get *p*, he will, ceteris paribus, do *q* (Haselager, 1997).

Since this sort of psychological explanation roughy resembles the way ordinary people interpret and predict the behaviour of others it is often called *commonsense* or *folk psychology*. This association with ordinary talk has led some authors to consider psychology unscientific, by comparing it with other folk theories that are overtly false, such as astrology (Churchland, 1979). This is not the place to enter the debate about the status of commonsense psychology, but it is important to note that the pairing between commonsense psychology and (scientific) psychology is not precise, though. Psychological explanations need not be strictly matched with the theorising ordinary people use to explain behaviour. The central idea appears to be that psychology shares with commonsense psychology some of its fundamental vocabulary and principles. As Haselager (1997) says:

> At a minimum, folk psychology is characterised by the use of a vocabulary in which mentalistic concepts like "belief", "desire", "fear", "hope", etc., might play a major part. As such, folk psychology plays a major part in scientific psychology. (p. 9).

Therefore the way psychology uses mentalistic notions to explain behaviour does not need to conform the same formulas used by the folk. The point is that mentalistic vocabulary is not eliminable if we want to capture generalisations. An example of mentalistic explanations grounded in scientific methodology has been put forward by Rey (1997). He argues that there are some "standardised regularities" that can objectively be found in the results exhibited by common standardised tests of mental abilities. Those regularities correspond to patterns of response (e.g. marks in a paper) common among millions of answer sheets for those tests. Rey argues that there is

no alternative way to account for the objective fact of those regularities in the way people respond to those tests if our explanations are not couched in terms of people's *seeing* and *understanding* the questions in the sheets, *representing* the sentences in a certain way, *believing* that some answers are correct, having the *desire* to respond them correctly, and so on. Therefore, and following scientific realism, the use of mentalistic explanations in psychological level explanations is then vindicated due to their explanatory and predictive power.

As must be apparent at this point, I am taking for granted that psychology—understood as the scientific study of the mind and behaviour of minded agents—is possible. So in the following I shall assume that there is a higher-level explanation of the behaviour of minded agents, which is characteristically psychological in the sense described above, and that can support law-like, counterfactually supporting, generalisations.

1.4.2 The Computational Level

Explanations framed at a computational level are at the very heart of the scientific study of the mind after the cognitive revolution (see 1.1.2 for this term). The basic idea is that beneath psychological-level explanations, there is an intricate level of symbolic structures and computational processing that implements the mentalistic capacities described by the level above. For example, when we describe Peter as looking through the window and generating the belief that it is raining outside, we are omitting many details about how the cognitive capacities perform that job. Let us focus on visual perception. When someone perceives an object, psychological explanations say that she can represent it, often consciously, and think about its properties and relations. But at a deeper level of description, the starting point is not the perception of the object as such but a complex process that starts from the encoding of information about the light reflected by the object on the retina, and a series of transformations carried out over that information. Marr (1982) was one of the first researches to focus on the computational processes that underlie perception. He offers a glimpse to his account in the following passage:

> First, suitable representations are obtained of the changes and structures in the image. This involves things like the detection of intensity changes, the representation and analysis of local geometrical structure, and the detection of illuminating effects like light sources, highlights, and transparency. The result of this first stage is a representation called the *primal sketch*. Second, a number of processes operate on the primal sketch to derive a representation —still retinocentric—of the geometry of the visible surfaces. This second representation, that of the visible surfaces, is called the *2½ dimensional (2½-D) sketch*. Both the primal sketch and the 2½-D sketch are constructed in a viewer-centered coordinate frame, and this is the aspect of their structures denoted by the term sketch. (p. 42)

The transition between one of the representational stages to another is mediated by computational processes, that transform informational structures into increasingly complex representations of aspects of the visual array such as *blobs* (closed curves), *zero-crossings* (changes in light intensity), *boundaries*, etc. At the end of this process the system is able to integrate this information and construct a *3-D* representation of an object in the environment. Marr proposed that in order to understand these complex processes it was important to focus on a computational level of description devoted to specifying the representational structures and the algorithms involved (he in fact calls this level *algorithmic*).

An important aspect of computational explanations is that they usually work by decomposing mental capacities into a series of subsystems or interconnected components that carry out more specific algorithmic processes and contribute to the functioning of the system as a whole (Dennett, 1979; Lycan, 1995). According to CTM those subsystems are also often regarded as modular, in the sense that their computational operations run in relative isolation from other subsystems (Fodor, 1983). These characteristics of the computational level show why it constitutes a distinctive level of description situated below explanations about the behaviour of a person as a whole, which is why it is sometimes called the *subpersonal level* (see chapter 7 for a detailed account of this term). So for example, instead of focusing on an object a person perceives and the impact it has on its behaviour, a computational explanation goes deep into a functional analysis of the different computational stages that make possible the perception of an object in the first place. Illustrative of the distinctiveness of the computational level is the nature of its symbolic structures. While psychological mental symbols normally denote objects or properties that are familiar to commonsense (e.g. an apple, or the colour red), the symbols that figure in computational level explanations are

rather disparate. For example, symbols processed at the stage of primal sketch are what Marr calls zero-crossings, which carry information about discontinuity in image brightness. These symbols are part of the coding of multiple information about the visual array that in further stages will give form to a mental symbol with (what we might call) full-fledged representational content[8].

But this last point can also help us to see that the computational and psychological levels also have a common inferential and symbolic nature. In both cases, computational explanations can be formulated, and the appeal to symbolic structures is mandatory. For this reason, it seems appropriate to call both levels "cognitive". However, as suggested in the previous paragraphs, it is scientifically useful to characterise the computational level as a level of description distinct form the psychological level, given its analysis in terms of subsystems and the disparate nature of its symbolic structures (I will delve into the distinction between both levels in the last chapter of this thesis). Furthermore, in the following chapter I will argue that an additional reason for keeping computational explanations separate is that they map onto a distinctive natural domain which is independent from the psychological.

1.4.3 The Physical Level

Physical-level explanations of behaviour are formulated in terms of the physical sciences broadly conceived, encompassing not just physics but also sciences such as biology and neuroscience. They characteristically describe the causal events that underlie behaviour by appeal to their physical properties. It is important to note that in principle any behaviour can be described by appeal to the physical level, insofar as every causal event in nature involves physical structures. However, as noted above when describing the behaviour of some complex agents there are generalisations that correspond to arbitrarily large disjunctions of physical structures. They are thus couched in a functional vocabulary that quantifies over multiple physical descriptions, as

[8] Some authors have found it appropriate to characterise the content of some computational states as *non-conceptual* as a way to account for their difference with the content of paradigmatic mental states such as beliefs (Bermúdez, 1995; Stalnaker, 2003). Even though I prefer to stay neutral about the conceptual/non-conceptual debate, it should be noted that what motivates those authors to make that distinction is the same as mine for singling out the symbolic structures described by the computational level, namely, to determine whether (at least some of) those structures have a different nature than mental symbols.

happens with computational and psychological level explanations. Overall, we have complex agents such as human beings whose behaviour can be analysed in terms of different levels of organisation, each adequate for capturing generalisations that would otherwise be missed from the viewpoint of the other levels.

In some agents, though, their behaviours can be straightforwardly explained by the physical level without there being any justification for ascribing computational or psychological states of any sort. A straightforward example can be seen in reflexive behaviours. For example, when some animals are exposed to an abrupt, intense sound, they respond with a contraction of skeletal and facial muscles known as *startle reflex*. The basic neural circuit underlying this reflex has been well studied in rats, and consists of four synapses running from the auditory nerve to a spinal motor neurone (Swerdlow, Caine, Braff & Geyer, 1992). A physical explanation attempts to describe this sequence of causal interactions with more or less detail about the physicochemical processes involved. It is worth noting that reflexive behaviours can be quite elaborated and comprise a chain of automatic responses rather than a single reflexive reaction. They are sometimes called *fixed action patterns* and like simple reflexes are characteristically innately constrained and invariable.

A second form of behavioural explanation that does not require going beyond the physical level corresponds to *associative conditioning*, whether classical or instrumental. Take for example a simple form of conditioning called habituation that has been extensively studied in a sea slug called *Aplysia*. A tactile stimulus to the siphon of this animal (a tubular structure on the dorsal surface of the animal) normally causes a withdrawal effect of the siphon and gill. After repeated stimulation of the siphon, this response shows a decrement known as habituation. Significant progress has been made in elucidating the neural and molecular mechanisms underlying the habituated response of *Aplysia* (Byrne, 1990). After repeated stimulation of the siphon, its sensory neurones release progressively less neurotransmitter to their presynaptic terminals, due to internal molecular changes. This results in a less activation of the motor neurones and a diminished motor response. But beyond the case of habituation, the neuronal

mechanisms that underlie associative conditioning are well understood and can be spelled out in proper physical level terms[9] (see e.g. Hawkins & Kandel, 1984).

An important thing to remark is that physical level explanations sometimes suffice by themselves to account for behaviour, and no better explanation couched in computational or psychological levels is justified. For instance, the habituation response of *Aplysia* consists on single cells that diminish their neurotransmitter release to motor neurones, and there is no place for a mapping of abstract properties of the environment to generate symbolic structures, and less for the instantiation of fundamental computational operations such as storing of transforming symbols. As I will explain with more detail in the next chapter, computational-level explanations (and so psychological-level explanations as well) are typically flexible and adaptable, and cannot be spelled out in terms of reactions to physical properties of the stimulus or direct associations between stimuli and responses.

[9] Gallistel has made a case against explanations of learning based on associative conditioning by arguing that they can be better formulated by appeal to symbolic structures and computation (see e.g. Gallistel & Gibbon, 2001). This is because some learning behaviour exhibits complex features (such as time scale invariance) that do not fit into the framework of associative conditioning but are naturally described from a computational viewpoint. For present purposes, it should be noted that Gallistel's proposal is compatible with the ideas discussed in this section. All it would imply is that some behaviours previously explained just by the physical level (because they involve associative learning) should be explained by the computational level.

Chapter Two

The Autonomy of the Computational Domain

2.0 Introduction

In the present chapter I argue that the computational level of explanation picks up an autonomous natural domain: a domain of computation and information processing that is common to many computing systems but is autonomous from the domain of psychology. This has important consequences for the purposes of drawing a line between minded and non-minded agents. One is that many authors who defend the mental capacities of animals indulge in unnecessary ascription of mentality to some animals (e.g. insects) on the main grounds that they possess certain complex computational capacities. I contend that these authors overlook the autonomy of the computational domain, and fall into the false dilemma of assuming that behaviour has to be explained either from the physical or the psychological level. Instead, I propose that it is possible to conceive some animals as, say, marvelous biological computers without having to ascribe mentality to them.

A second consequence I draw from the claim that the computational domain is autonomous will be matter for the remainder of this thesis, but it is worth mentioning at this point. It is that given that we have a mind by virtue of being some kind of computer, while it is possible to conceive some animals as non-minded computers, we can wonder what is special about us, such that we have mentality while other computing animals do not. This opens the negative side of my thesis, where I critically review—through chapters 3 to 6—recent literature exploring the line that divides computational agents with and without mentality, until we get to the last chapter where I put forward my own proposal.

2.1 The Reality of the Computational Domain

Cognitive science normally takes the computational level to be central for understanding the behaviour of mental creatures. However, to some extent it is the least intuitive of the three levels of behavioural explanation I presented in the previous chapter. The physical level describes behaviour by appeal to the physical constitution of the brain in a way that is not substantially different from physiological explanations of the functioning of other parts of the body. The psychological level, for its part, explains behaviour in terms that do not differ much from the vocabulary of commonsense psychology. But with regard to computational explanations, they are couched in a vocabulary and describe generations in a way far less familiar to commonsense. This is partly due to the fact that they generally describe subpersonal processes not accessible to introspection and because they involve functional states that cannot be identified with physical structures of the brain. Moreover, sometimes it is not clear under what circumstances a machine implements a computer, or what we have in common with them on this respect. For these reasons I devote most of this chapter to clarifying the nature of the computational domain.

When we say that an entity's behaviour can be explained by the computational level we basically mean that it is a computer. But what precisely is a computer? On first consideration, a computation is just a functional mapping between two domains (e.g. inputs and outputs). From this basic definition, however, computations are being realised in every entity whose behaviour can be described by appeal to mathematical function. For example, a planet would be computing its orbit, or an enzyme the chemical reaction it is catalysing. It soon becomes clear that in this broad sense a computation is nothing beyond physical-level descriptions that use mathematical models to explain their phenomena. As Crane (1995) points out, in cases like this we might describe the planet and the enzyme as instantiating mathematical functions, but not as computing them. To be a computer is, then, a more demanding notion.

As I mentioned in the last chapter, a more precise definition of computing system was put forward by Turing (1937) and later on synthesised by Newell & Simon (1981). They claim that (digital) computers are symbolic and universal. They are symbolic because computations act upon symbolic structures that can be interpreted as

bearing reference relations to something else. And computers are universal machines in the sense that they perform some basic set of fundamental operations[10] (such as encoding, storing, deleting and transforming symbolic structures) required for computing any algorithm (see 1.2.2). This is not to say that any actual computer has to be a multi-purpose system, though. The point is just that it must have an internal structure that is flexible enough to be in principle programmed to run different algorithmic procedures (cf. Haugeland, 1981; Copeland, 1993; Pylyshyn, 1984).

Some authors have challenged the idea that computational level explanations map onto a real and objective domain we can find in nature, and claimed that computation is essentially an observer-dependent phenomenon. I address two ways of posing this critique and present replies to them. In this way I hope to strengthen my positive claim that computing systems constitute a genuine and autonomous natural phenomenon.

2.1.1. Objection 1: Computation is not a Real Property of Entities

In a series of writings, John Searle (1990, 1992) has attacked the metaphysical status of the computational domain. He starts by stating that computational patterns can in principle be found in almost any physical entity. For example, Searle (1992) writes:

> Thus for example the wall behind my back is right now implementing the Wordstar program, because there is some pattern of molecule movements that is isomorphic with the formal structure of Wordstar. But if the wall is implementing Worldstar, then if it is a big enough wall it is implementing any program, including any program implemented in the brain. (p. 208).

The argument takes the form of a reductio ad absurdum: if something as simple and crude as a wall has the capacity to implement a computational program, then the ascription of computational states becomes trivial and uninteresting. In some passages Searle (1990) makes the stronger claim that even the programs (if any) run by the brain could be implemented in any physical entity given enough "free hand" for selecting the

[10] Note that the list of fundamental operations can be flexible, given that different combinations of fundamental operations can be equivalent in terms of their capability to compute any algorithm. So, for example, two computing machines might carry out the same task by employing alternative sets of fundamental operations. See Copeland (1993, ch. 4) for discussion.

events in that entity that match the corresponding computational patterns stated by the program, since “such patterns are everywhere” (p. 636). Clearly if any program is actually being implemented in any physical entity, the claim that computation is an objective natural phenomenon becomes implausible.

Thus formulated, Searle’s argument is unconvincing for many reasons. First, he certainly overestimates the capacity of a wall to implement a computational program. Consider Searle’s own example of Wordstar. This is a program—popular in the eighties—that in some versions consists in around 137.000 lines of code. Now we might ask, on what grounds does Searle make his claim that an ordinary wall, even a very big one, could possibly be capable of implementing 137.000 lines of a program? As Haugeland (2003) points out, to plausibly claim that some entity has certain program encoded or inscribed in it there must be some way, at least in principle, of reading or extracting the program from that entity. For example, if Wordstar is “inscribed” in a magnetic disc or a compact disc, there would be technological devices capable of recovering the program from those discs. But could any device, a “wall-reader” say, possibly recover precisely the 137.000 lines of code of Wordstar from Searle’s wall? It is certainly implausible to say that this is so and therefore the claim that a wall or any entity can implement any program is false.

A second reason why Searle’s argument is flawed is that a computing system is not just a program. To see this point recall the definition of computation I gave in the last section. A genuine computing system has to be able to instantiate a universal machine and perform computational operations over symbolic structures. Therefore, a program instantiated by a computer is not a static script of instructions but a causal mapping of transition states that run according to those instructions. It is not clear how the molecules of a wall, even a very big one, could ever have the causal structure required for mirroring the dynamics of the transition states of a program.

A possible rejoinder to this critique can be extracted from Putnam (1988) who, while putting forward an argument similar to Searle, argues that an ordinary object such as a wall is an open system in the sense of being affected by external forces and therefore its internal structures are going through a series of changes over time. Since the evolution of such complex structures is dynamically changing over time, they could

then be described as the formal mapping between state transitions specified by a computational program such as Wordstar or any other. But as Chalmers (1996) points out, this is still a deflationary account of what is to implement a computational program. The causal sequences performed by a computer are not rigid but flexible. Computers can perform operations such as storing, retrieving, transforming and generating new combinations of its inner components, and so have the potential to function in more than one way, something that the physical structure of a wall just lacks. Put in more precise terms, the causal sequence of state transitions run by a computer must support counterfactuals, that is, their inner structure must have a casual organisation that allows us to say that if the system had been in a different state, it would had functioned in an interestingly different way. As Chalmers notes, only a small fraction of physical systems actually have the structural complexity required for implementing a computational program.

Thus taken at face value, Searle's critique rests on an overly simple notion of computation. But to be fair, Searle (1992) is aware of this and recognises that after "tightening up our definition of computation" the problem of multiple realisability could vanish. He even outlines how that definition might be, pointing out that it would emphasise the causal structure of the system, its programmability and situatedness in the real world, therefore anticipating the response to his argument presented above. But Searle (1992) also unfolds a second argument:

> But these further restrictions on the definition of computation are no help in the present discussion because the really deep problem is that syntax is essentially an observer-relative notion. The multiple realizability of computationally equivalent processes in different media is not just a sign that the processes are abstract, but that they are not intrinsic to the system at all. They depend on an interpretation from outside. (p. 209, emphasis removed)

The core of this second argument is that computation is not intrinsic to physics, in the sense that it is a property that finds no parallel in the physical structure of the world. Any attempt to characterise a physical entity as having a computational structure is an observer-relative description that says nothing about what the entity intrinsically is. Searle (1992) compares, for example, the expression "computational program" with "chair" or "weed", which arguably "do not name intrinsic features of reality" but rather "name objects by specifying some feature that has been assigned to them, some feature that is relative to observers and users" (p. 211).

This point seems misguided, though, if we recall the notion of functional kind explained in the first chapter. Contrary to chairs or pictures drawn in a paper, functional kinds correspond to abstract states that play a causal role in behaviour. Chairs and pictures do not enter into causal explanations by virtue of being chairs and pictures, at least not in a way independent of there being people capable of thinking about them. But more importantly, functional kinds are supposed to capture generalisations that are of scientific import and according to scientific realism this is what vindicates them as genuinely existing. Chairs and pictures on the other hand, do not take part in scientific predicates and the law-like regularities they describe.

But Searle makes a further distinction that might evade that critique. For him, computers are functional systems in a way different from other functionally characterisable entities such as carburettors or digestive systems, since the latter but not the former are supposed to be linked with physical causes and effects. Computational processes, the argument goes, consist of purely abstract algorithms, similar to instructions written in a paper, which can only cause something as far as someone thinks of them as carrying out a certain job. Therefore, the same point is restated: computation is not an intrinsic property instantiated in the real world but an observer-relative property that depends on people's purposes and intentions.

This argument works, however, only for a purely syntactical view of computation, one that characterises computing systems just in terms of the formal structure of the algorithms they perform (and which I will address later in 2.3.1). But as I explained in the previous chapter, this is only part of what makes something a computing system. Computers are embodied entities capable to enter into genuine causal commerce with their environments, and thus the internal symbolic structures that explain their behaviour have a physical dimension linked with causes and effects, which exist independently from observer-relative considerations. Therefore in this sense the functional description of a computational system is not so different from the description of a carburettor. In both cases we could have a functional description that abstracts from the particular context in which the system works, but that description would not exhaust the nature of the system. I shall return to these ideas in section 2.3.

2.1.1 Objection 2: The Computational Level is just an Interpretationist Stance

A second objection to a realistic approach to the computational domain is also based on the assumption that the attribution of computational states is always observer-dependent. However, the present objection recognises that computational states can play an important role in our explanations of behaviour and justifies their use for pragmatic reasons such as their predictive success. This view is often associated to *interpretationist* approaches regarding behavioural explanations and its most illustrious defender has been Daniel Dennett[11].

For Dennett (1979, 1987, 1991) what I call computational level corresponds to a *system-design approach* (or *stance*) to behavioural explanation that focuses on the computational mechanisms that underlie psychological capacities. This strategy, carried out mainly by artificial intelligence researchers, consists basically in successively breaking-down complex cognitive capacities into a set of simpler subcapacities until the algorithms performed by these subcapacities can be determined and then simulated in a computing machine (Dennett, 1979, p. 113). Dennett recommends the design stance as an insightful research programme capable of improving our understanding of how the mind works. But as I shall explain in more detail later, he regards this approach as an interpretative exercise where what is true about the system is the interpretational scheme in general but not its details.

Thus Dennett's critique to a realistic approach to mental states can be also considered a critique to realistic approaches to the computational domain. But what distinguishes Dennett's interpretativism from Searle's scepticism about computation is that Dennett believes that the good predictive results obtained by computational explanations indicate that there is something true and objective about them, however he remains agnostic about the ultimate structure and contents of the computational states involved. Dennett (1991) defines his peculiar mixture of realism and interpretativism as a form of "mild realism". Before going into his arguments, it is pertinent to note that the present view is not uncommon among cognitive scientists. For instance, in the

[11] This view should not be confounded with general scepticism about scientific truth. Interpretativist authors such as Dennett, do believe that science can reveal to us what is out there. His view can be better understood as a moderate scepticism about scientific explanations based on the ascription of abstract functional states and processes.

introduction to their textbook on computational neuroscience, Churchland and Sejnowski (1999) point out:

> [W]hether something is a computer has an interest-relative component, in the sense that it depends on whether someone has an interest in the device's abstract properties and in interpreting its states as representing states or something else. (p. 48)

But at the same time, they recognise that a computational approach to the functions performed by the brain has an objective side, at least to the extent that

> discovering what the function [performed by a system] is reveals something important and perhaps unexpected about the real nature of the device and how it works. (p. 49)

The idea that a computational description reveals something about the inner mechanisms that underlie behaviour makes this view less radical than the eliminativism about computation championed by Searle. In other words, the interpretativist finds somewhat justified the ascription of computational states in an "as if" fashion for the purposes of interpreting behaviour.

At this point it is important to note that Dennett's concerns about realistic ascriptions of cognitive states are most of the time directed to what I call psychological level (for Dennett *intentional stance*) explanations. He claims that even though belief-desire explanations are in some sense real, they do not neatly map onto specific inner mental states or mechanisms. But as commented earlier, he extends his concerns to the system-design stance (viz. computational-level explanations) and there are common motivations behind his scepticism towards both computational and psychological ascriptions. More precisely, according to Dennett when we ascribe a mental or computational state to some entity we are saying that it is in certain sort of functionally characterised inner state, both in the sense of being defined by its causal role within the system and in having teleology, i.e. a role in the performance of certain cognitive capacity. As such, mental and computational states involve the ascription of whole sets of background information, rules and goals. As Dennett (1979) notes:

> One predicts behaviour in such a case by ascribing to the system the possession of certain information and supposing it to be directed by certain goals, and then by working out the most reasonable or appropriate action on the basis of these ascriptions and suppositions. It is a small step to calling the information possessed the computer's beliefs, its goals and subgoals its desires. What I mean by saying a small step is that the notion of information or misinformation is just as intentional a notion as that of belief. (pp. 6-7).

Then similar principles apply to both psychological (i.e. intentional) and computational level explanations, and moreover both imply "interpreting an entity by adopting the presupposition that it is an approximation of the ideal of an optimally designed (i.e. rational) self-regarding agent" (Dennett, 1994, p. 239). This is because for Dennett the only way to make intelligible a functional capacity is by assuming that the system has been optimally designed or disposed to perform that capacity. And this appeal to optimality amounts to idealisation: we have to assume that the system has been optimally designed to fulfil its role and that it is operating under optimal conditions. The main reason for adopting this approach is, according to Dennett (1987), pragmatic. In his words:

> The fact that an object can be reliably expected to approximate optimality (or rationality) may be a deeper and more valuable fact than any obtainable from a standpoint of greater realism and detail. (p. 79)

But this appeal to optimality is very puzzling. Certainly the assumption of optimal design would reveal deeper and valuable information about human cognitive capacities if they really were optimal, but as Stich (1981) notes, this is just not the case. Breakdowns and shortcomings in human cognitive performances are common. People often perform below standards of full rationality on psychological tasks, and this is not attributable to a lack of cognitive capacities to perform those tasks, but to the fact that human capacities are sub-optimal. Let me explain with an example from computational-level explanations.

Wasps have complex navigational abilities that make them capable of foraging over long distances and then finding their way home. In order to orientate themselves, wasps can memorise visual properties of landmarks present near their nests and during their displacements and rely on them to determine the direction of their flight back. In addition, wasps can also navigate by keeping record of the distance and direction travelled, through a process known as *path integration*. They appear to shift to this second mode of navigation when landmarks are not available, as for example when flying in unfamiliar terrain (Healy, 1998). If we adopt a design stance, we could describe the wasp's path integration system as a computational device with the function of orienting the insect towards food and then back to its hive. Now, suppose we are adopting Dennett's interpretationist approach to explain the behaviour of a wasp that

soon after departing starts to fly in apparently random directions and never gets back to the hive. Since we would be assuming that the navigation system of the wasp has an optimal design, we would be unable to keep ascribing that system to the insect and have instead to find some other optimal system that could explain its behaviour or just assume that it is random and deserves no explanation at all. But it certainly would be more natural to suppose that the wasp does have a navigation system, but that it simply got lost. The navigational system might have malfunctioned, or perhaps have a less than optimal design, but this is no reason for denying its existence.

The same idea can be extended to almost all cognitive functions, since it seems likely that they are almost never optimally designed, in particular if we look at their evolutionary origins. The selective processes that explain their design show that in order to be evolved by natural selection functional systems need not necessarily be optimal, even not very efficient or precise. All they need to get passed through generations is that they make their possessors better adapted to their environment than their local competitors. Therefore Dennett's strategy of assuming that cognitive functions are (approximately) optimal would lead us to either deny that entities have functions or to impose multiple post-hoc amendments to ensure that our interpretation of the behaviour fits some notion of optimal function. But even if we take the design stance as a kind of heuristic strategy for guiding our study into the sub-systems or an entity, it would clearly lead us nowhere in our attempt to reveal the real nature of the functional systems under study.

However, Dennett (1987) puts forward another reason for sticking to optimality which derives from an arguably unavoidable degree of indeterminacy in our behavioural explanations. According to him, when scientific theories deal with abstract theoretical states and processes, possible explanations are underdetermined by the observational data. In the case of explaining behaviour, an indefinite number of alternative computational or intentional interpretative schemes may be equally compatible with the behaviour under study, but without sufficient empirical grounds to prevail over the others. And note that this lack of objective evidence becomes more critical for the case of animals, were we do not have the data from introspection that, arguably, provides humans with additional evidence about their internal workings.

Nonetheless, it is not clear that indeterminacy is unavoidable. As it is customary in any science that deals with abstract entities, it is possible to formulate hypotheses about possible states and processes that cause behaviour and to test them empirically. Indeterminacy can then in principle be reduced in so far as we collect adequate and enough empirical data. To illustrate this idea let us return to the example of insect navigation, this time focused on honeybees. There is strong evidence that honeybees are capable of navigating long distances and rarely get lost in their way, and also that they can communicate the others about the location of rich sources of food (Menzel et al., 2000). Imagine we are studying their navigational capacities and thus we are interested in the information and algorithms involved in their computational operations. From the viewpoint of the interpretativist, though, due to her characteristic pessimism about finding out reliable details about inner states we would have to stick to a broad functional characterisation, perhaps adding some intentional qualities such as beliefs about landmarks in the environment and desires about reaching sources of pollen, etc. But at the end we would not be able to say much about the inner states of the insect.

However, careful observation of the behaviour of honeybees has provided researchers with enough evidence for supporting hypothesis about the inner states of those insects. For example, it is well known which sorts of information of the environment are transmitted by the different movements of their dance, and thanks to the use of harmonic radar it has become possible to track the flying-paths of individual bees (Menzel et al., 2005). Then it would come as no surprise that the kinds of information being processed and the algorithms that control the navigation of the honeybees are gradually discovered and that progress is made in the understanding of their navigational systems. On the contrary, it is hard to see how the instrumentalist could make a good job predicting behaviour just remaining neutral about which computational processes are going on inside the insect.

These examples illustrate how the internal states and processes the govern behaviour are relevant and that the more precise we get into its details, the more accurate predictions we will obtain (cf. Rey, 1997; Carruthers, 2004). And even if it we concede to the instrumentalist that sometimes a certain degree of indeterminacy cannot be completely ruled out, this would be preferable over the massive amount of indeterminacy proclaimed by the interpretativist, indeterminacy that leads them to

refrain from ascribing any particular internal state to organisms. This would make impossible to identify patterns of internal states that could be generalised among individuals, challenging the very possibility of a scientific explanation of complex behaviour. And this is, arguably, too high a price to pay for embracing this position (cf. Margolis & Laurence, 2007).

2.2 The False Dilemma

Before the advent of cognitive science, standard forms of explanation for animal behaviour came from what I have called the physical level, viz. reflexes, fixed action patterns and associative conditioning. But since cognitive science emerged philosophers and cognitive ethologists have become optimistic about the prospects of ascribing internal functional states to animals (e.g. Fodor, 1975; Gallistel, 1990). This has also been encouraged by ethological studies that show that many animals appear to possess complex computational capacities.

A good case-study of this shift of perspective comes again from the behaviour of the wasp. As many insect behaviour, the wasp's was generally considered the result of fixed action patterns or at best basic forms of associative conditioning. For example, when it comes to lay its eggs the digger wasp *Sphex* brings food to its burrow nest, leaves the food near its opening, and proceeds to check inside the burrow for the presence of intruders. If nothing is found disturbed, then the wasp emerges from its burrow and drags in the food. Interestingly, if an experimenter moves the food while the wasp is still inside the burrow, it will invariably repeat the same routine of leaving the food near the burrow and going into it for inspection. This procedure can be repeated again and again, without the wasp altering its behaviour (Wooldridge, 1971).

The example of the digger wasp is often put forward as a case of a rigid, stereotyped behaviour, that does not deserve to be explained in psychological terms (e.g. Dennett, 1984; Sterelny, 1990). However, recently some philosophers have argued that some insects such as bees and ants do have mentality, on the basis that they exhibit behaviours that are much more complex and flexible than the one observed in the digger wasp (Carruthers, 2004a, 2006; Fitzpatrick. 2008). Indeed, as I mentioned in the last

section even the wasp itself has been described as possessing complex navigation systems such as path integration that demand a computational explanation. In sum, it has been argued that some behaviours of insects have to be explained by appeal to capacities of computation and information processing that far outstrip fixed innate patterns or associative conditioning. Therefore, the argument goes, we are justified in shifting to psychological explanations and in ascribing mentality to those insects.

As a consequence, it has become common for cognitive ethologists to use the psychological level to account for those complex behaviours, often behind the assumption that without the availability of mental vocabulary they would be left without alternative ways of explaining them (Jamieson & Beckoff, 1996). Staying with the case of insects, the argument put forward to ascribe mentality to them normally takes the following form:

P1: We are justified in placing animals within the psychological domain iff their behaviours cannot be best explained in terms of other explanatory domains

P2: Certain behaviours of insects cannot be best explained in terms of the physical domain

---

C: Insects that exhibit those behaviours can be placed within the psychological domain

Let me briefly comment the premises. P1 is an assumption about the epistemic grounds that justify the description of animal behaviour in terms of certain explanatory domain. It states the common idea that the same animal behaviour can normally be explained at different levels, but that we should only adopt a positive epistemic attitude towards those explanations that are better than the others. How to precisely spell out what makes an explanation better than another is not an easy task, but most cognitive ethologists agree in that best explanations are those that have more explanatory coverage and predictive power (Allen & Beckoff, 1997).

P2 is a consequence of the well-established fact that some insect behaviour is far too complex to be explained at only the physical level. The most common example are their already mentioned navigational abilities to which I shall return in section 2.4.

Others are communication and non-associative forms of learning (Gallistel, 1990; Carruthers, 2006). But even if we take both P1 and P2 as true, the argument is not sound insofar as it has a hidden premise, which I shall argue is false. It is the following:

> P3: The physical domain is the only alternative to the psychological domain to explain animal behaviour

By omitting this premise theorists have implicitly assumed that the behaviour of animals has to be described by using either of two explanatory levels: the physical or the psychological. But this is a false dilemma, since it ignores one alternative explanatory level, viz. the one that describes the computational domain. To support this claim, though, it has to be argued that computational level explanations are autonomous, which will be the task for the remainder of this chapter.

2.3 The Autonomy of the Computational Domain

So far I have argued that the computational domain is a real, objective part of the natural order. As a consequence, it should be considered an alternative explanatory domain, on pain of falling into the false dilemma explained above. A possible rejoinder to this, however, is that the computational domain is only instantiated when the psychological domain appears. That is, genuine computation is a mental phenomenon, and therefore any computation in the world either happens in entities with a mind or is derivative from them. In this section I will argue that this is wrong, and that computational level explanations pick up a natural domain that supervenes on the physical domain, but is not dependent on the psychological domain. The key argument is that it is at least possible to conceive entities whose behaviour can be satisfactorily explained in terms of computational processing over symbolic structures, while there is no reason for adding psychological notions to explain them. Therefore, the computational level can be regarded as autonomous from other explanatory levels. But before arguing for the autonomy of the computational domain some words are in order on what is been understood by autonomy and computer. Below I address them in turn.

Autonomy is certainly a thorny philosophical notion and a full discussion of its meaning goes beyond the purposes of this thesis, however it shall be useful to set forth a basic idea of what is an *autonomous agent*. First, an autonomous agent is one that is self-governed in the sense that its behaviour results from the intrinsic character of the system[12]. This is a sense of autonomy commonly used in artificial intelligence, where typical examples of autonomous agents are mobile robots that can navigate and execute their own actions, as opposed to teleoperated robots that are remotely controlled by a human operator (e.g. Brooks, 1991). Given that we are interested in entities that behave as part of the natural order, another important aspect of autonomy is *embodiment*. Embodied organisms are supposed to cope with, and perform tasks effectively in, physical (real) environments. A more precise definition of embodied autonomy is given by Franklin and Graesser (1996) who claim that "an autonomous agent is a system situated within and as part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future" (p. 31). Therefore disembodied entities, such as virtual agents that figure in computational simulations, or physical entities the inputs and outputs of which are highly dependent on human operators, would not count as autonomous in the present context.

As Franklin and Graesser note, this definition of an embodied autonomous agent is not very restrictive since it would include simple artefacts such as a thermostat. Indeed, some authors have characterised autonomy more strongly by adding requirements that are characteristic of living creatures, such as self-organisation and self-regulation. For instance, Smithers (1997) claims that "autonomous systems must have some means of forming their own laws of regulation as well as the means to regulate their behavior with respect to them" (p. 94). This capacity of "self law-making" can be found in creatures that can adjust their responses according to their interaction with their environment, as is the case of learning[13]. A complex issue that emerges at this point is whether life is a necessary condition for autonomy. Boden (2001) convincingly contends that this is not the case, for even though autonomy appears to be essential for life, it is at the same time a more fundamental notion. Nothing in principle prevents

[12] The point can indeed be viewed as a restatement of the notion of behaviour defined in section 1.1, viz. as the sort of activity that is primarily caused by some internal factors of the agent who is behaving.

[13] The proposal of Dretske discussed in chapter 4 below can be interpreted along these lines.

some non-biological creatures from being autonomous insofar as they govern and regulate themselves, and even evolve.

Regarding the notion of computer (see 1.2), a crucial aspect of a computer is that it is a universal machine in the sense explained in 1.2.2 and 2.1. This universality constraint rules out all those entities that process information in a fixed functional architecture that lacks the computational resources for performing the set of fundamental computational operations required to run a program. For example, the first electronic calculators developed in the 1960s have in-built logic circuits that can perform basic arithmetic operations, however they are incapable of manipulating inner symbolic structures in a flexible way or of transforming them. As a consequence, even though those calculators can perform computation in the sense of doing functional mappings between their inputs and outputs, they are not computers since they cannot be programmed to run a different algorithm. Genuine computers have the capacity to perform a fundamental set of operations such as having branching points in their sequence of states that allow them to jump onto alternative sequences of symbol manipulation.

So if we agree that there is a more or less clear distinction between agents that are computers and those that are not, then we have that the computational level maps onto a natural domain, a domain of computation and information processing that is common to some computers. But can we say that this domain is also independent from the domain of mental agents? I believe the response is yes, and that the most straightforward way of arguing for this is by showing that some entities can be considered genuine computers and that at the same time there is no justification for ascribing minds to them. Let me elaborate this idea with an example.

In the last chapter I explained that computers can mirror the inferential structure of mental processes and in this way exhibit what is one of the hallmarks of mentality. Besides, computers can be embodied creatures, equipped with transducers to pick up information from their environments and the algorithms required for processing that information in a way that leads to intelligent action. In those cases, I argue, it is plausible to ascribe to a computer at least basic forms of symbolic algorithmic processing. A straightforward example of this are robotic computing machines that can

behave in natural environments without any help. Take the case of the two robotic rovers that are now exploring the surface of Mars (i.e. Opportunity and Curiosity). Among their multiple capacities are autonomous navigation, thermoregulation, and the capacity to collect detailed geological information from the planet and transmit it to the earth[14]. It appears highly intuitive to say that these vehicles are fairly intelligent computing machines, even though they lack mentality. But in any case, and this is the main point, they exemplify how computational-level explanations can be autonomous, for in order to explain how the robot behaves in such a desolated place we have to appeal to the syntactic and symbolic components of its computational architecture. I will return to the case of these robotic rovers in the following sections when addressing objections to the idea that the computational domain is autonomous.

2.3.1 Objection 1: The Computational Level is just Syntax

This first objection directly challenges the autonomy of the computational domain. It claims that when cognitive scientists describe the mind from the computational level they are not picking out any domain distinct from the psychological domain. Instead, computational explanations are regarded as nothing beyond a formal characterisation, or a syntactic description, of the computational operations carried out by mental agents (Fodor, 1980; Egan, 1995). According to this objection the computational level describes the manipulation of symbolic structures, but from a perspective that focuses on their formal or syntactic properties, not their representational or informational properties. Therefore, the objection goes, the computational level does not describe any particular ontological level of organisation but just "provides a formal, environment-independent, characterisation of a process". (Egan, 1995, p. 199).

[14] Of course, the Mars rovers receive instructions from earth. But since those instructions take some time to reach the rover, they are designed to carry out many tasks in a rather autonomous way, such as self-monitoring, navigating and making some decisions without human intervention.

One motivation[15] behind this position is the *formality condition*, the view that computational explanations of the behaviour of mental agents can only advert to formal (nonrepresentational) properties of symbolic structures (Fodor, 1980). Given that psychological explanations do deploy mental symbols with representational properties, a consequence of the formality condition is that the computational domain can only be partially realised by the psychological domain, in particular that "syntactic processes at the computational level implement causal laws [situated] at the intentional [psychological] level" (Fodor, 1991, p. 280), while the symbolic structures at the computational level just cannot implement the representational properties of the psychological level. If this is the case, then the computational level cannot map onto a separate—autonomous—domain in a hierarchy of supervenient levels of explanation.

I believe the formality condition is too restrictive, though, and that computational-level explanations should do without it. First, this condition rests on an incomplete characterisation of computing systems, one that focuses on their syntactic architecture but abstracts from their behaviour as embodied agents in the real world. This distinction can be made clearer by reflecting on the idea that computers are universal machines. On the one hand, a computer has the potential to, in principle, run any algorithm, and on the other, a single algorithm can be used for performing many tasks (in the same sense as a single algebraic operation can be used for different purposes). However, when an algorithm is implemented in concrete computers instantiated in the real world, computational-level explanations do not describe them as abstract, universal formulas, but as inferential procedures engaged in genuine causal commerce with the environment. This factual, situated dimension of computational explanations makes algorithms something more than a formal abstraction. Even if two computers are running the same algorithm considered from a formal viewpoint, computational explanations of their behaviour might differ insofar as the information they process and the tasks they perform are different.

Secondly, it should be noted that computational-level explanations do not lack the resources to account for the representational dimension of symbolic structures. As

---

[15] A second motivation can be *internalism*, the view that representational notions cannot play any genuine role in a scientific psychology (Kim, 1982; Stich, 1983). But for the purposes of this thesis—and in accordance to the background I set forth in the previous chapter—I assume *externalism*, according to which it is plausible to formulate psychological explanations that advert to representational contents.

explained in the previous chapter, informational approaches to representation can provide a framework for doing so. The basic idea is that computational symbols are information-bearing structures, and computational-level explanations normally deal with the coding and transformation of those structures. These computational processes are typically subpersonal, that is, they underlie psychological (belief-desire) explanations. So if information is an objective commodity that can be picked up from the environment, coded and transmitted through subpersonal—computational-level—explanations, then there is nothing mysterious in assuming that computational symbols do possess contents (although distinct from the contents of mental symbols couched at the next level "up", viz. the psychological level). In chapter 7, when spelling out the distinction between personal and subpersonal levels of explanation, I shall return to this point.

Another motivation for conceiving the computational level as a distinct, autonomous level of description, is that without it we would be left with no theoretical resources to explain the behaviour of non-mental computing agents. The existence of autonomous computing robots capable of engaging in informational transactions with their environment and behaving effectively on it, is undeniable. But if, as the view presented above contends, the computational level is just a formal description of the syntax of psychological processes, then how could we account for those robots' successful negotiation with their environment?

By way of example, imagine that we want to study one of the Mars robotic rovers mentioned in the previous section, in particular its behaviour related with searching and examining a Martian stone. If we adopt a purely syntactic approach we would be able to describe the algorithms implemented by the robot throughout the process. But of course, this computational account would have to recognise that these algorithms are not just formal abstractions but effective procedures that actually manipulate symbolic structures that carry information from Mars. What could we do to account for these symbolic structures then? One way could be to shift to the psychological level and call those structures mental symbols. However, this would imply ascribing mentality to the robot, something that seems implausible. The only alternative appears to be, then, to simply ascribe the robot with symbolic structures from an autonomous computational domain. This would suffice to explain how information

picked up from Mars is relevant to account from its behaviour, and also to account for counterfactual situations such as what algorithms would had been implemented if the stone it examined had a different composition. Therefore there seems to be no reason for treating computational level explanations as purely formal, or as dependent on there being mentality in those agents that implement a computer. They might just be non-mental computing agents in their own right.

2.3.2 Objection 2: The Computational Level of Artefacts has Derived Representations

The second objection to the autonomy of the computational domain can be formulated as a rejoinder to my reply to the first objection. According to this second objection, the case of the Mars rovers does not count as a example of autonomous computation because they are human-designed machines. The general idea behind this argument is clearly stated in the following quote from Haugeland (1981):

> [symbolic structures] only have meaning because we give it to them; their intentionality, like that of smoke signals and writing, is essentially borrowed, hence *derivative*. To put it bluntly: computers themselves don't mean anything by their tokens (any more than books do)—they only mean what we say they do. Genuine understanding, on the other hand, is intentional "in its own right" and not derivatively from something else. (pp. 32-33).

The author uses the term *intentionality* to refer to the semantic or representational properties of symbolic structures. For present purposes the point is that any computational artefact, insofar as it is the product of the purposeful design of a human being, inherits its intelligence from the purposes and intentions of its creator. So, the objection goes, robots such as the Mars rovers have intelligence and other capacities only in a derived, non-original sense. This objection can be linked to theorists who defend teleological approaches to the mind and adopt a historical approach to cognitive functions. They also make the positive claim that the only way an entity could be endowed with intelligence or any kind of function is by have been designed by non-purposeful mechanisms such as natural selection (Millikan, 1984; Papineau, 1987).

One problem with this teleological objection is that it rests on controversial assumptions about how history determines the nature of functions and cognitive

capacities (Crane, 1995; Fodor, 2000). For instance, a creature that happens to lack an evolutionary history of selection would be incapable of instantiating any function, regardless of how complex its current behaviour is. But even more problematic is the fact that a teleological approach would have to rule out, as a matter of conceptual analysis, the possibility that an artefact of the right sort could ever be capable of developing intelligent behaviour. On the contrary, it seems likely that even if we have not created genuinely intelligent or cognitive machines so far, things might change as computational technology develops (Copeland, 1993).

This is not the place to settle this issue, however. The idea that artefacts have intelligence or process information in a way that does not depend on human beings might still be resisted by some authors. In any case, I believe that the objection that artefacts cannot count as instantiations of an autonomous computational domain can be overcome in a different way: by arguing that there are biological systems that instantiate computational capacities, but even though do not deserve to be ascribed with mentality. Note that since biological systems are the product of evolutionary processes, they are invulnerable to the second objection. I elaborate this point with examples of real biological computers in the final section of this chapter.

2.4 The Case of Biological Computers

So far in these first two chapters I have sketched the minimal requirements for instantiating a computing system. They basically consist in having symbolic structures capable of engaging in informational relations with the environment and capable of performing the basic operations that characterise a universal computer. In the present section I present two examples of biological entities that appear to meet these requirements for being a computing system, without there being a clear justification of ascribing mentality to them. With this, I wish to finally make the point that the computational domain is autonomous and independent of the psychological domain, and that theorists that ascribe mentality to animals often make the mistake of ignoring the former as an alternative explanation of behaviour.

My first example is concerned again with the digger wasp. As I mentioned in section 2.2 the food-dragging behaviour of this insect has become a commonplace for showing how rigid and stereotypical insect behaviour could be. But as also noted, the wasp also has some more clever facets such as sophisticated navigational abilities. Wasps can orientate themselves by using a range of navigational cues and also by path integration, which involves the capacity to estimate their position by using information about their own speed and direction travelled (Healy, 1998). The navigational capacities of the wasp exhibit many of the hallmarks of a computational system. They can encode and store information about environmental cues and carry out computational operations over that information, and those operations can be quite complex, notably when doing path integration. In those cases the insect has to monitor its angular and linear displacements and continually update the current distance and direction from their present location to their starting point, and sometimes recalculate their vector flight, for instance when they find themselves lost. That involves basic computational operations such as storing, deleting and transforming data-structures, as well as manipulating them through algorithmic steps that branch into alternative courses of action.

An important thing to note is that the navigational capacities of the wasp are probably modular and specific for that task (Carruthers, 2006). That explains why the remarkable intelligence exhibited when flying cannot be used for a different task, such as altering their food-dragging behaviour when fooled by an experimenter. So the wasp does exhibit intelligent behaviour, however it is restricted to its navigational capacities. Should we ascribe psychological states to the wasp then? It is at this point when some authors have fallen into the false dilemma of regarding the computational capacities of the wasp as suitable for psychological explanation. But why should we do this? Would it not be more plausible to just describe the wasp's behaviour in terms of the computational domain and remain neutral about whether they have minds or not? After all, they might just be, if you like, marvellous biological robots, that instantiate computation but not mentality (in chapter 7 I return to this case).

It might be objected, however, that my scepticism about the ascription of mentality to the wasp is question-begging since I am just assuming that the computational processes that control its navigational behaviour lack any form of mentality. Someone might claim, for instance, that any creature that instantiates

computational processes in its brain deserves to be explained by the psychological level. I believe this objection does not work since it draws the line for instantiating mentality too low, reaching biological systems to which the ascription of mentality appears as totally implausible. I will present a final example to illustrate how implausible it is to ascribe mentality solely on the grounds of biological computation.

My final example relates to the *enteric nervous system*. It consists in a large network of a hundred million neurones (ten times more than the wasp's brain!) located in the wall of the human gastrointestinal tract. It accomplishes a variety of functions such as regulating processes of secretion and absorption, blood flow, and controlling the motility of the intestine (Wood, 2011). This motility control involves coordinated patterns of contraction and relaxation at different parts of the intestine, related with segmentation, peristalsis and motility cycles. The enteric nervous system has been shown to be quite complex and to operate rather autonomously from the brain, to the extent that it has become known as "the gut brain". A remarkable aspect of this system is that it can receive and integrate information coming from different sources; in particular inputs from the brain and information about the mechanical and chemical conditions of the intestine. Then, the enteric nervous system can produce organised motor patterns that are generated, coordinated and modulated by the own system (Thomas, Sjövall & Bornstein, 2004).

Arguably, capacities of the enteric nervous system such as encoding and integrating information, and modulating complex patterns of intestinal motility, make it a good candidate for instantiating computational processes. But whether the enteric nervous system should be regarded as a genuine digital computer or not is beyond the scope of this chapter. My purpose with this example is just to illustrate how plausible is to conceive biological systems that even though they lack mentality, can instantiate the computational domain. If there is no impediment for the implementation of computational capacities by non-mental creatures, then the claim that insects have minds cannot be grounded just on the fact that they are computing systems. Something else should be said about what makes them part of the selected group of computing systems that possess minds.

Chapter Three

Informational Approaches: Fodor on Drawing the Line

3.0 Introduction

In the first two chapters I put forward a distinction between computational and mental symbols, in the sense that their extension can be different, as happens when computational symbols are instantiated in non-mental agents. But what makes some animals or computing agents in general capable of developing mental symbols? If we were to draw a line between computer agents with and without mentality, where should we do so?

The overarching goal of the remainder of this thesis will be to address these questions, and in chapters 3 to 6 I take up recent literature exploring what makes a computing agent capable of developing mental symbols. In this chapter and the next I will deal with informational approaches to mental symbols, starting in the present chapter with the work of Jerry Fodor. He develops an informational account of mental symbols, claiming that what is special about them is the nature of their contents, which is determined by their capacity to bear certain informational relations with entities in the environment. Roughly speaking, the author claims that what matters for drawing a line between mental and merely computing systems is the way in which those relations are fixed and the entities they can bear relations to.

After critically presenting Fodor's views at two moments of his work, I conclude that even though his account presents some problems it also constitutes a promising approach to the issue of telling what makes for mental symbols. However, it still needs to say more about the computational architecture that is required for those purposes, something that shall be explored in further chapters of this thesis.

3.1 From the False Dilemma to the Slippery Slope

In the previous chapter I argued that some defenders of animal cognition were at risk of falling into a false dilemma when justifying the attribution of mentality to animals on the grounds that the physical level is not enough for explaining some of their behaviours. I suggested that in order to avoid this fallacy, authors dealing with behavioural explanation should take the computational level as mapping onto an autonomous natural domain that is independent from the domain of mental creatures. Put in a different way, animals that lack mental symbols might be possessors of computational symbols, and therefore should be considered as more sophisticated than non-computational creatures but at the same time below the domain of creatures with mentality.

But if we follow this suggestion and incorporate the computational domain to our metaphysical inventory, then we would have to face an additional problem when dealing with non-human cognition. It is that if we try to justify the attribution of mentality to certain agents by appeal to the framework of CTM, then it would appear as legitimate to ascribe mentality "all the way down" to entities that do computation but do not appear to have a mind at all, such as the human digestive system or the Mars rovers. The risk is real given that most proponents of CTM regard as likely that the difference between us and other computer machines is, at least from the viewpoint of computation, largely quantitative. For example, Fodor (2003) writes:

> If ... some sort of inferentialism is likely to work for our minds, isn't the least hypothesis that it is also likely to work for the minds of other kinds of creatures? Surely it's reasonable, absent contrary evidence, to suppose the differences between our minds and theirs are largely quantitative. (p.4)

If what distinguishes us from other entities can be measured in terms of the quantity of inferential (i.e. computational) capacities, then where should we stop ascribing inferential thought to computer machines less complex than us? Not surprisingly, Fodor (1986) is aware of this sort of slippery slope objection and, as the following passage shows, he also hints at a possible way out:

> [The] slippery slope argument that gets you from us to paramecia can also be made to get you from computers to thermostats or, for that matter, from *us* to thermostats. It will not, in short, do to take Descartes's route and get off the slippery slope by postulating that lower

> organisms are machines. For one would then need an argument for not attributing mental representations to machines; and if to any machines, why not to all of them? (p.4)

As suggested at the end of the quote, on pain of falling in a slippery slope CTM needs to put forward some argument for not attributing mental representations (i.e. mental symbols[16]) to entities to which it is implausible to do so. It should be noted, though, that according to the conceptual framework advanced in the previous chapters the problem here is not with machines in general, but with computing machines in particular. For machines that do not instantiate a computer, i.e. not capable of running a program, are not candidates for mentality insofar as a minimum condition for having mental symbols is to have computational processes at its subvenient base. This certainly rules out paramecia and ordinary thermostats as candidates for mentality.

But there is a vast space of computing agents that might still fall into the slippery slope, for example the human digestive system or the Mars Rovers discussed in the previous chapters, and we might even include programmable thermostats equipped with complex input-output systems. So for present purposes we will have to adjust Fodor's proposal by claiming that what CTM needs is not just a principled way for not attributing mentality to every machine, but to every computing machine. In this chapter I examine how Fodor has attempted to develop this idea and how it could be used to draw a line between computers that do and do not implement mentality.

To appreciate Fodor's proposals it is important to keep in mind some of his previous work on perception and cognitive architecture, which I briefly review in the following section. Then I address how Fodor has dealt with the issue of drawing a line for mental representation in two stages of his work. First in his paper *Why paramecia don't have mental representations* where he draws the line on the capacity to respond selectively to non-nomic properties, and secondly in relation to his more developed *asymmetric dependence theory* put forward in Fodor (1987, 1990). The example of paramecia might strike the reader as unsuitable for present purposes given that, as I explained above, they can be ruled out from the scope of mentality due to their non-computational nature. However, Fodor's proposal can still be useful for us since it can be applied to the case of ruling out non-mental computer agents as well.

---

[16] To simplify the exposition I take Fodor's preferred term *mental representation* as equivalent to the term mental symbol I introduced in chapter 1. I believe this is innocuous insofar as both terms refer to the fundamental symbolic structures used in psychological explanations to describe thought and reasoning.

3.2 Brief Overview of Fodor's Account of Perception and Cognition

Fodor's view lies on the theoretical framework of CTM—of which he is one of the main contributors—, in particular in relation to perceptual systems (Fodor & Pylyshyn, 1981; Fodor, 1983). So it shall be useful to review some of this framework before going into the details of Fodor's proposals. I focus on the case of vision because it has extensively been studied from the viewpoint of CTM, however the main tenets of this account are supposed to be applicable to other perceptual systems as well. Broadly speaking, in the context of CTM perception is a cognitive mechanism that encodes and transforms information about the world and delivers it to central cognition in a format appropriate for thought and reasoning. Perceptual processes are supposed to comprise three stages that involve the following systems:

1- *Transducers*

2- *Input systems*

3- *Mechanisms for belief fixation*

Transducers constitute the first layer of perception and carry out the task of converting energy coming from the environmental stimulus into action potentials transmitted by neurones. In the case of vision, transducers are photoreceptors in the retina which transform luminous energy into electric energy appropriate for nerve conduction and computational processing. They do so by means of a purely physical process of encoding environmental information about distant objects conveyed by the light, a process normally described in terms nomic (or lawful) covariation (see 1.3). The information encoded by transducers is then processed by input systems, which can compute that information in order to infer properties about the distal objects, as Fodor (1983) summarises:

> The character of transducer outputs is determined, in some lawful way, by the character of impinging energy at the transducer surface; and the character of the energy at the transducer surface is itself lawfully determined by the character of the distal layout. Because there are regularities of this latter sort, it is possible to infer properties of the distal layout from corresponding properties of the transducer output. Input analyzers are devices which perform inferences of this sort. (p. 45)

Input systems (called "input analyzers" in the quote) are special-purpose computational mechanisms of early visual processing that analyse and interpret the information about light reaching the retina. This stage of visual perception is important since the information encoded by transducers is variable, relatively fragmentary and mathematically insufficient to recover data about distal properties. So perceptual systems must constraint the possible interpretations that could be drawn from this information in order to generate the right representation of distal environmental objects. For example, if we are looking at a house and start moving towards it, our retinal image of the house gets bigger and bigger, and if we do so at different times of the day the intensity of light that reaches our retina might also change. However, we still seem to see the same house, with the same size and colours. That is possible thanks to the computational processes carried out by input systems which have the function to extract information about environmental invariants.

According to Fodor (1983), input systems work in rather isolation from the rest of cognition, which is the reason why they are often regarded as *modular*. They can map from the information about a proximal stimulus that the retina provides "to a representation of the distal stimuli as an array of objects in space" before that representation is coded as a perceptual belief (p. 53). This implies that input systems constitute the stage in visual perception where elements of the distal environment are first individuated or categorised, what happens before those representations meet background information (i.e. stored in memory) involved in central processes of belief fixation. As Fodor himself notes, processes carried out by input systems are compatible with the computational stages put forward by Marr's theory of vision, who described the algorithms involved in the mapping of transducer information onto a representation of a three dimensional object in the distal layout (see 1.4.2).

Fodor calls the products of input systems *perceptual categories* or *percepts* since he regards them as representations of basic categories of objects and properties of the environment. These categories involve not just properties of objects such as colour, shape, size and motion, but also sometimes the representation of whole objects such as a dog or a tree. What is notable about percepts is that they can be generated just by means of modular computational mechanisms that process information encoded by transducers. Therefore their production does not demand the possession of other

representations or the integration of percepts with previous beliefs about the world. In this sense it seems clear that Fodor considers appropriate to characterise percepts as genuine representational states that denote basic categories, and to accept that beliefs (and propositional attitudes in general) are not the only form of representational state. More precisely, his idea suggested in 1983 is that representations are formed at the level of input systems and later on "corrected" under the light of background knowledge at the level of belief fixation (p. 102). But can percepts, or information coded by mechanisms like input systems, be properly called mental representations? Do percepts have a content that is somewhat equivalent to that of mental representations in general?

Considering Fodor's early writings (at least until 1983) the author would have responded negatively to this last question. That is because before that time he endorsed a view about content called *functional-role semantics* (Rives, 2010), and thus believed that the contents of mental representations are at least partially determined by the functional relations they bear with other representations. Given the view about perception and cognition summarised above, this implies that the content of representations only become fixed at the level of belief fixation, viz. where they are integrated with other representations and thus acquire their particular functional role. In sum, according to this view percepts are considered precursors of, but not already a kind of, mental representation.

In later writings, though, Fodor has become convinced that functional role semantics leads to many problems and has developed an informational approach to representational content. This allows that some percepts, at least those encoding certain environmental properties and bearing the right nomic relations to them, could count as mental representations. A clarification is in order, though. Fodor sees the nature of mental representations as distinct from the nature of the mind more generally. This is because he individuates mental representations by their contents, while he individuates the mind in terms of being a *system* of mental representations. To clarify this issue consider the following quote from Fodor (1987):

> I'm leaving open that a good reconstruction of intentionality might recognise things that have intentional states but no propositional attitudes; hence, things that have intentional states but are not intentional systems. For example, it doesn't seem to me to count against a theory of intentionality if it entails that the curvature of the bimetalic strip in a thermostat represents the temperature of the ambient air. By contrast, a theory that entails that

> thermostats are intentional systems—that they have beliefs and desires—would thereby refute itself. (p. 163, n. 1; see also Fodor, 1986, p. 21, n. 3)

So Fodor takes mental creatures to be intentional systems, viz. possessors of symbolic structures with causal roles that issue in behaviour, while when he talks about intentionality he is referring to content. Then since mental representations are individuated by their contents, and given that Fodor attempts to explain the content of mental representations by an informational theory that do without their causal roles, it turns out that mental representations can in principle be individuated in creatures that do not have minds. For example, Fodor (1998) says:

> [T]he content of propositional attitudes depends on the content of mental representations, and since the intended sense of 'depends of' is asymmetric, ... [my account] tolerates he metaphysical possibility of mental representation without thought. (p.9)

Then it appears that according Fodor's metaphysical picture it is possible to conceive a creature having a mental representation without having a mind. This is certainly hard to swallow, but for present purposes I propose to bracket this issue and evaluate how Fodor accounts for mental representations in the first place. In any case, according to the author this is a crucial step towards explaining why some creatures have minds and why others do not, since mental representations are a necessary—though perhaps not a sufficient—condition for having mentality. In the remainder of this chapter I discuss two stages of Fodor's informational approach to mental representations, where he directly deals with the issue of what are the minimum conditions an information-bearing state coded by perceptual systems has to meet to be regarded as a genuine mental representation.

3.3 Fodor's First Line: Selective Response to Non-nomic Properties

As mentioned above, in his paper *Why paramecia don't have mental representations*[17] Fodor proposes a criterion to draw a line between creatures with and without mental representations. More precisely, he attempts to identify a certain capacity or property that could make it plausible to ascribe mental representations to entities who posses it, and then he uses paramecia as an example of a species that

[17] Unless otherwise indicated, all section references are to this paper.

clearly lacks that capacity and therefore does not meet the criterion for having mental representations.

The paramecium is a single-celled animal that lives in freshwater environments. It is also a photosensitive organism, what means that it exhibits tropic behaviours in response to light intensity. Fodor starts conceding to paramecia the capacity to "see" light in the sense of instantiating a property that covaries with certain property of light, which causes a manifest behavioural response (e.g. movement in the direction of less intense illumination). Then Fodor makes an important distinction between nomic and non-nomic properties of objects. Nomic properties are those which enter into lawful relations with other properties of nature. For example, the frequency of electromagnetic radiation is a nomic property of light since there are physical laws that relate this property with others. Nomic properties are therefore also properties to which organisms can be sensitive in virtue of instantiating a nomological covariation between some property of their transducers and a property of an external object. The capacity of the paramecia to "see" light falls into this category.

On the other hand, non-nomic properties of things are those that cannot enter into any lawful relation. For example, the property of *being a left shoe* or the property of *being a house*[18]. Even though Fodor concedes that these might be real properties, he claims that there are no laws that hold for left shoes or houses in virtue of being that property. He then draws attention to the fact that organisms which are sensitive to external properties just by means of nomological covariation cannot be sensitive to non-nomic properties, since the latter cannot covary in a lawful way with anything else. Fodor argues that this is the case of the paramecium: it is only capable of "seeing" the nomic properties of things, given that its sensorium can only covary with properties that can enter into nomic relations with them. But non-nomic properties such as *being a left shoe* or *being a house* are out of the scope of the sensory systems of paramecia, because these properties cannot causally interact with their transducers.

Then Fodor suggests a criterion to differentiate entities which can have mental representations from those that cannot:

---

[18] In this thesis I follow Fodor's orthographic convention of using italics to denote properties and quoted formulas to express the contents of symbolic structures.

> [A]ny system that can respond selectively to non-nomic properties is, intuitively speaking, a plausible candidate for the ascription of mental representations; and any system that can't, isn't. (p. 11)

A key expression of this quote is *respond selectively*, for in some way a paramecium may be able to, for instance, "see" a house when confronted to it, but it will not be capable of having a behavioural response attributable to the property *being a house* since it cannot respond selectively to houses. All the paramecium can "see" are, say, levels of light intensity and frequency, which are just nomic properties of light. Fodor claims that the paramecium is a bad candidate for having mental representations because it can only react to nomic properties. On the other hand, human beings are paradigmatic examples of entities who can represent and respond selectively to instantiations of non-nomic properties. We can see houses or left shoes and act in ways that are a direct consequence of representing these particular properties. Fodor claims that this capacity to respond selectively to non-nomic properties is the most distinctive feature of cognition and that it permits us to distinguish paramecia from organisms that possess mental representations:

> [S]elective response to non-nomic properties is, on the present view, the great evolutionary problem that mental representation was intended to solve. And the solution to that problem was perhaps the crucial achievement in the phylogeny of cognition. (p. 14)

Where precisely in the phylogenetic continuum mental representations emerged is taken by Fodor as an empirical matter that depends on which properties are nomic and on the psychological capacities given organisms actually have. In his paper he is not interested in specifying where the line must be drawn, but in providing a criterion for the attribution of mental representations that intuitively leaves paramecia out of its scope.

At this point it can be useful to see Fodor's proposal in the light of the three levels of perception explained in the previous section. According to his view, paramecia are just capable of reaching the first of the three levels of perception (i.e. they just have transducer mechanisms) and thus can only respond to nomic properties of the environment. What they lack, though, is the computational machinery to map from the nomic properties detected by transducers onto percepts that represent non-nomic properties. In the text Fodor emphasises this point by associating transducers with the detection of nomic properties:

> ... in short, the point about transducers is that they respond selectively only to nomic properties. (p. 16)

> ... And conversely, if a property is nomic it is always possible to build a transducer to it. (p. 22, n. 10)

Then it follows that the relevant distinction between us and paramecia is that "only the former can respond selectively to properties that are *not* transducer detectable" (p. 15). Fodor clarifies that he does not want to draw the line on transducers but on nomicness, because the latter is a more fundamental notion than the former, however from the above quote it is clear that he takes properties that are not transducer detectable as non-nomic. Then the crucial step in the phylogeny of mental creatures appears to be the development of perceptual capacities (i.e. input systems) that can infer properties of objects that are not transducer detectable. This introduces a "semantic connection into the causal chain" that goes from information encoded by transducers to percepts (p. 14). Then the burden of the argument falls on having perceptual capacities, as Fodor makes clear in the following passage:

> What distinguishes intentional systems from the rest is that, whereas we've got perceptual categories, what they've got is, at most, sensory manifolds. (p. 20).

In regard to the last stage of perception (i.e. belief fixation), in the paper under discussion Fodor remains silent about what else beyond perceptual categories is required for having perceptual beliefs. For the purposes of his paper, however, it suffices for him to state that having representational states is a necessary, but not a sufficient, condition for having beliefs, which are just mental representations with characteristic functional roles. Since paramecia cannot respond selectively to non-nomic properties and therefore cannot have mental representations, they are *ipso facto* incapable of having beliefs. In any case, it remains clear that for Fodor mental representations are a more fundamental aspect of mentality than beliefs.

3.3.1 Objection 1: The Line is Drawn too Low

Certainly Fodor is successful in making the point that paramecia cannot represent non-nomic properties and that appears to be a good reason for not ascribing mental representations to them. However, I believe that he overestimates the capacity to

respond selectively to non-nomic properties as a criterion for the attribution of mental representations and ends up situating the line for the attribution of mental representation too low, thus permitting us to ascribe them to entities to which it would be clearly implausible to ascribe mentality. I shall use the example of a vending machine to illustrate how under Fodor's criterion, a mindless entity like this should be ascribed mental representations, making up an argument that takes the form of a reductio ad absurdum. Of course, the same argument could be extended to any other artefact or biological system equivalent to a vending machine from an information processing viewpoint.

First, a few words about the kind of vending machine I have in mind. It has an input system that consists in a coin acceptor, which can distinguish between coins of different value based on weight, size and magnetic content. It comprises an internal processing system that carries out computations over the amounts of money inserted into the machine, in order to validate the purchase of beverages and calculate the change to be dispensed. The machine also has an output mechanical system for dispensing beverages and another for dispensing change, and for the sake of the example let us imagine that this vending machine accepts only UK coins and that the only beverage it dispenses is one that costs one pound. A critical aspect of this vending machine is that its internal processing system is actually a computer in the sense defined in the previous sections. Even though for practical reasons vending machines are normally much simpler than this, I believe the example works since it is certainly plausible to conceive a vending machine equipped with a computer like the one just described.

My argument against Fodor is that a vending machine like the one described above can have an informational-bearing state that can be described as denoting the non-nomic property of *being one pound's worth*. To see the point it can be useful to compare the way the coin detection process works with the first two stages of a perceptual system described in 3.2. First, it has transducers that transform nomic properties of coins (viz. weight, size and magnetic content) into electric signals suitable for the internal processes of the machine. Then it carries out calculations over these signals such as adding the values of the different coins, and produces outputs to the beverage dispenser and the coin changer device in an analogue way as perceptual modules output to central cognition. Since from an information processing viewpoint

the vending machine implements a process comparable to the first two stages of perception, it seems that according to Fodor's criterion we would have to consider the machine's capacity to respond to the property *being one pound's worth* as one that yields a mental representation, a consequence I take to be unacceptable. Next, I address two possible replies Fodor could give to this argument, which are based on distinctions he makes in the paper under discussion.

First, Fodor could object that the vending machine is really detecting a non-nomic property (such as *being one pound's worth*). Before unfolding this reply let us recall that the reason why it is impossible for the paramecium to detect non-nomic properties is that those properties cannot covary lawfully with their transducers. In contrast, creatures capable of mental representation can detect non-nomic properties by means of their input systems. For example, to perceive a house a creature first has to detect nomic properties of the house reaching its retina (i.e. patterns of light intensity) to then infer the non-nomic property of *being a house*. As Fodor notes, this would not be achievable without the mediation of inferential processes because the set of nomic properties associated with non-nomic properties such as being a house is "vastly and *open-endedly* disjunctive" (p. 19). Houses can have an indefinite variety of nomic properties in relation to their size, colour, shape, etc., and therefore no transducer can be tuned to respond to houses by means of a lawful relation.

Similarly, the representation of the property of *being one pound's worth* can also be generated by means of quantifying over an indefinite number of possible representations of nomic properties; just consider all the possible combinations of weights, sizes and magnetic contents of coins that could make for one pound. Then Fodor's objection could be the following: when a vending machine responds to the insertion of coins that sum one pound, it is not really responding to the property of *being one pound's worth*, but just to a conjunction of transducer detectable properties such as certain weight, size and magnetic content. After all, contrary to what happens to genuine perceptual systems, the vending machine only has a limited capacity to detect particular nomic properties of coins, and is far from being able to infer the property *being one pound's worth* from an open-ended disjunction of possible nomic properties because most of them are out of its scope.

I believe this reply does not work because it overestimates the capacities of perceptual systems. As happens with the vending machine, the scope of nomic properties our perceptual systems can detect is limited, given their transducer and processing capacities (e.g. our visual system is insensitive to certain properties emitted by objects, such as ultrasonic sound waves and their magnetic field). Then to deny the vending machine's capacity to respond selectively to non-nomic properties on the grounds that their transducer detectable properties are limited is misleading, given that our perceptual systems are constrained in a similar way. Furthermore, it should be noted that possible combinations of coins that could make the vending machine be loaded with one pound is considerable[19], what shows that there is a significant variability in the way nomic properties of coins could mediate in the instantiation of an information-bearing state that corresponds to one pound's worth. So even the vending machine has a restricted scope of physical properties it can extract from proximal stimulation, and the combinations of those properties it can use to detect a non-nomic property is rather vast. It is not totally open-ended, but in this respect the machine does not differ from perceptual systems in general.

A second rejoinder Fodor could put forward to deal with the counterexample of the vending machine is to concede that it can have information-bearing states generated by inferential processes and that those states are not transducer detectable, but argue that those states still cannot count as mental representations because the actions carried out by the machine do not *respond selectively* to them. Imagine that you insert two 50 pence coins and the machine dispenses a beverage. There is a configuration of nomic properties detected by the coin detector, viz. two equal conjunctions of size, weight and magnetic field corresponding to a 50 pence coin. Let us call this configuration "2x50swm". This corresponds to a transducer detectable information-bearing state that carries information about nomic properties of the two coins. When the machine dispenses a beverage it is responding to *2x50swm*, which happens to be equivalent with the non-nomic property of *being one pound's worth*. Fodor could then contend that the behaviour of the machine is not sophisticated enough to give us sufficient grounds to determine whether it is responding to one property or the other. Dispensing the beverage would not suffice to determine whether the machine is discriminating between these

[19] Supposing that the machine accepts coins ranging from 1 penny to 1 pound, there are 4563 ways to make a pound.

two possible internal states, and we would not be justified in ascribing it a representation of being one pound's worth to explain this action. Then Fodor could call us to be conservative and refrain from attributing a mental representation to the vending machine unless we have enough behavioural evidence to discriminate whether it is responding selectively to the property *being one pound's worth* or just to the conjunction of nomic properties *2x50swm*.

In response to this rejoinder, I argue that the vending machine can respond to its input in a way sophisticated enough to attribute to it an information-bearing state of the property of *being one pound's worth*, something that according to Fodor's account would lead to the implausible conclusion that it has a mental representation. As Fodor himself acknowledges (p. 5) at the end it is a matter of inference to the best explanation whether we are justified or not in attributing the representation of properties that are non-nomic. And the case of the vending machine appears to be just a case where this attribution is justified, for at least the following two reasons.

First, the vending machine dispenses a beverage selectively when loaded with one pound worth, even when it involves adding different kinds of coins. The fact that the machine produces the same response under a wide range of possible inputs allows us to generalise from the varying configurations of proximal stimuli and ascribe it an info-bearing state that covaries with being one pound's worth. Instead, if we insist in explaining its actions in terms of responses to *2x50swm* or any other particular configuration of nomic properties we would miss out this important generalisation that is relevant for explaining its behaviour.

Secondly, the vending machine can dispense change that corresponds to the difference between the sum of the values of the inserted coins and one pound worth. To put it in algebraic terms, the machine carries out the following calculation:

$$x - 1 = y$$

Where x is the sum of the values of the coins inserted and y is the quantity to be given as change. There are many combinations of coins that can load the machine with more than one pound worth. For instance a two pound coin, 50 pence plus three 20 pence coins, etc. In all these cases the machine reliably gives change that is worth

precisely the value that *y* represents in the equation above. In order to explain how the machine can produce this behaviour we need again to generalise from the particular processes carried out when each combination of coins is inserted, and ascribe the machine the capacity to compute this calculation. And since the fixed value in the equation is representing the property of *being one pound's worth*, we would be justified to attribute that information-bearing state to the machine.

3.3.2 Objection 2: Nomic / Non-nomic Distinction is Irrelevant

Fodor's proposal can be read as starting from the assumption that the capacity to encode environmental information by means of transducers is not sufficient for a creature to yield mental representations. This assumption is quite uncontroversial and normally regarded as the basic problem that any informational approach to representation has to deal with, for many non-mental entities do bear environmental information by virtue of having properties that covary lawfully with other properties. For example, three rings bear information about the age of a tree and thermometers bear information about the temperature in a room. On pain of having to ascribe mentality to entities such as trees or thermometers, informational accounts to representation have to spell out the processes required for transforming environmental information into a full-fledged mental representation (see 1.3).

According to Fodor, the main characteristic of perception is that it is equipped with input systems capable of taking information encoded by transducers as premisses and deriving perceptual categories as conclusions. Indeed, in the paper under discussion Fodor points out that the reason why paramecia cannot have mental representations is precisely because that they lack those perceptual capacities, adding that the same argument can be made by appeal to the incapacity of paramecia to respond to non-nomic properties—or properties that are not transducer detectable. However, I believe this last claim is ill-founded and at odds with his own view of psychological explanation. It is simply a mistake to situate the capacity to yield perceptual categories as somewhat equivalent to the capacity to detect non-nomic properties, for most perceptual categories are, in fact, about nomic properties (i.e. about natural kinds in general). It is possible, for instance, to imagine a creature whose perceptual systems are

only attuned to represent nomic properties (e.g. animals, plants, etc.). In this case, the creature would have perceptual categories even though cannot represent non-nomic properties, something at odds with to Fodor's criterion.

One way to avoid this objection could be to reformulate Fodor's criterion by saying that it is not the capacity to respond to non-nomic properties what makes for mental representation, but the capacity to respond to distal environmental properties without at the same time entering into nomic relations with them. Then perceptual categories, due to their inferential nature, would be capable of denoting environmental objects without being nomically related with them. But this way of looking at Fodor's criterion cannot work if we take into consideration the viewpoint of CTM I presented the first chapter (which is compatible with Fodor's own view, see Loewer & Rey, 1991), according to which the way mental representations enter into a sound metaphysical picture of the mind is through playing a causal role in scientific explanations of behaviour. Those explanations are typically nomological, involving laws that quantify over environmental objects, mental states and behaviour. For example, a psychological explanation of spider-avoidance behaviour would involve laws linking spiders to mental representations of spiders, and linking those representations with avoidance behaviour. But then, it makes no sense to claim that what characterises representational creatures is their capacity to respond to properties with which they cannot have lawful relations. For if there cannot be natural laws relating mental representations with their referents, then there simply cannot be psychological explanations based on those representations. And what is worst for Fodor, his proposal presently addressed would be incompatible with his own more recent attempts to naturalise mental representation along informational lines, where he champions that "semantic facts are somehow constituted by nomic relations" (Fodor, 1998, p. 73). In the following sections I address his more recent informational approach.

A final, alternative way to make Fodor's proposal more plausible and to avoid the previous objections could be to follow his emphasis on non-nomicness and state that what matters for having mental representations is not to have perceptual categories in general, but to have perceptual categories about non-nomic properties particular (such as being a left shoe or a crumpled shirt). Then, the line for mental representation would

not be drawn at the level creatures with perception, but those capable to produce percepts denoting non-nomic properties of the environment.

This seems arbitrary, though, since there is no clear reason to draw the line for having mental representations at the capacity to respond to non-nomic properties instead of nomic properties in general. After all, it seems plausible to suppose that some non-human animals do have mental representations by virtue of perceiving natural kinds such as trees or horses, even if they lack the capacity to represent more abstract properties like the ones instantiated by left shoes or crumpled shirts. Then nomicness does not appear to be the main issue, but the inferential nature of perception and the capacity to go beyond mere responses to transducer detectable properties. In fact, later stages of Fodor's work go along these lines by trying to account for the informational relations percepts have to hold with the environment in order to become genuinely representational.

3.4 Fodor's Second Line: Asymmetric Dependence Relations

In later writings Fodor acknowledges that the relevant difference between paramecia and us is not the kind of property we can respond to (viz. nomic or non-nomic) but that we have the capacity to respond to not transducer detectable properties. And given the background on perceptual theory given above, that amounts to saying that we can process information beyond the outputs of transducers and infer percepts through our input systems. As Fodor (1991) recognises, "the polemically relevant point about transduction is not that it's nomic but that it's *non-inferential*" (p. 257). Then it turns out that whether an animal can have nomic relations with its referents is not what is at issue, but how those relations are grounded. As Antony and Levine (1991) put it when commenting on Fodor:

> [T]he fundamental difference between representational systems and non-representational systems is to be found in the kind of nomic relationships into which the systems can enter. Thus, the defence of intentional realism need not depend upon the distinction between transducible and non-transducible properties, even if the distinction can be made. (p. 11)

So the job for a causal or informational account of mental representation is to specify which are the right causal/nomic relations these representations have to bear

with environmental properties in order to have genuine semantic content. As Fodor (1990) says:

> If there's going to be a causal theory of content, there has to be some way of picking out *semantically relevant* causal relations from all the other kinds of causal relations that the tokens of a symbol can enter into. (p. 91)

In a series of writings, Fodor (1987, 1990) discusses the kind of nomic relationships that might ground the generation of genuine mental representations, however I now focus on his initial discussion in *Psychosemantics*[20]. He again starts from the basic assumption that the capacity to encode environmental information by means of transducers is not sufficient for a creature to yield mental representations. As commented in 3.3.2, environmental information can be borne by non-mental entities and thus cannot be the whole story about the nature of mental representations. Then Fodor puts forward two ways in which nomic relationships have to constrained in order to establish a genuinely representational relation, which together conform his alternative informational approach to representation. Those constraints are presented as a means to solve two main problems that any informational approach has to face. Below I briefly present those problems, to then explain how Fodor's approach attempts to deal with them.

- *The all-problem*: If there is a law connecting A with B, then it is nomologically necessary that if A is the case, then B. Therefore if 'cow' tokens carry information about cows, then every time a cow is instantiated in the world a corresponding tokening of 'cow' has to be instantiated. But this is not true of ordinary mental representations insofar as not every instantiation of a cow actually causes a 'cow' token. For instance, only a minimal fraction of the cows that exist in the world happen to cause tokenings of 'cow' in my mind, and moreover, cows that exist in isolated places might never be the cause of 'cow' tokenings at all. Therefore an informational approach has to be constrained in some way to explain how it is that just some cows cause 'cow'. And as Fodor remarks, to avoid being question-begging these constraints have to be specified without appeal to other mental representations.

---

20 Unless otherwise indicated, all section references are to this book.

- *The disjunction-problem*: if the mental representation 'A' refers to B by virtue of being nomologically connected with Bs, then it cannot also be nomologically connected with non-Bs, such as Cs, since in that case 'A' would be referring to the disjunct (B or C). For example, it is plausible to conceive that some 'cow' tokens can sometimes be caused by horses by virtue of some nomological relation holding between properties of horses and 'cow' tokens, but if that is the case then 'cow' would not refer just to cows but to (cows or horses). This is called the disjunction problem since mental representations are normally supposed to bear reference relations to some particular properties and not to a disjunct (and less to an open disjunct as it turns out) of properties as the informational approach appears to imply.

3.4.1 The All-problem

Fodor's way of dealing with the all-problem is to specify certain sufficient conditions for the instantiation of 'cow' tokens such that when those conditions are met, cows cause 'cow'. Since those conditions are supposed to be stated in non-representational terms, they would allow an informational approach to explain in a non-question begging way why not all cows actually cause 'cow' tokens. Fodor develops a twofold process to account for those conditions. The first corresponds to the encoding of information by transducers, which he describes as purely *psychophysical*. It starts from the physical process that happens every time our sensory systems get in touch with energy coming from objects in the environment, and ends with the encoding of environmental information. For example, there are certain conditions under which red objects cause the tokening of an inner state carrying information about the redness of the object. In Fodor's words:

> Psychophysics purports to specify what one might call an 'optimal' point of view with respect to red things; viz., a viewpoint with the peculiar property that any intact observer who occupies it *must*—nomologically must; must in point of psychophysical law—have 'red there' occur to him. (p. 115)

But of course, mental representations are not the output of transducers, and so this proposal needs to be accommodated for the case of representations that are generated by inferential processes. Mental representations of distal environmental

objects such as horses or trees cannot be generated just by means of transducers and so their representation cannot be explained by mere appeal to psychophysical laws. Psychophysical circumstances can tell you when someone will "see" a horse, but not when she will "see as" a horse. As Fodor claims, "there are no psychophysically specifiable circumstances in which it is nomologically necessary that one sees horses as such" (p. 117).

Here Fodor adds a second step to the process, which corresponds to the mediation of inferences. Mental representations such as 'horse' or 'tree' are mediated by inferential processes drawn from the perceiver's "background cognitive commitments" (p. 117). The idea is that after transducers encode environmental information, perceptual processes pick up that information and through computation and integration with stored information generate a mental representation. But as Fodor recognises, to appeal to inferences and background commitments is question-begging, since they presumably involve previous representations and theories from which we draw the inferences that allow us to token 'horse' or 'tree'. This is more clear of representations of more abstract kinds such as protons, which require certain scientific knowledge to be tokened. The author avoids this critique by saying that the representational capacities involved in this process are not determinants of the content of a mental representation. As he says, "for the purposes of semantic naturalisation, it's the existence of a reliable mind/world correlation that counts, not the mechanisms by which that correlation is effected" (p. 122). And since the mechanisms required to sustain the fixation of content are computational, they can, according to the author, be specified in causal-syntactic terms, without appeal to representational notions. I will return to this issue in 3.5 when putting forward a critique to Fodor's proposal.

3.4.2 The Disjunction-problem

Concerning the disjunction-problem, Fodor attempts to distinguish between the nature of the nomic relations a mental representation holds with its referents, and the nomic relations it might establish with anything else distinct from them. So for example, in order to avoid 'cow' tokens being about a disjunct such as (cows or horses), there must be some way to tell apart the nomic relation 'cow'-cows from 'cow'-horses.

Fodor's suggestion is to state that the causal route between the mental representation and its referent is special in the sense that it does not depend on any other relation to exist. Hence 'horse' tokens are supposed to be caused by cows only because 'cow' tokens are, and not vice versa. Put in Fodor's terminology, the point is that:

> the causal connection between cows and 'horse' tokenings is, as I shall say, *asymmetrically dependent* upon the causal connection between horses and 'horse' tokenings. (p. 108)

The same idea can be framed in terms of counterfactuals. The nomic relations holding between 'cow'-cows and 'cow'-horses are different in their counterfactual properties, because while 'cow'-cows can hold without there been 'cow'-horses relations, the reverse is not the case. If there were no 'cow'-cows relations, there could not be nomic relations between cow and any environmental property distinct from cows. In more simple words, any nomic relation between non-cows and 'cow', is parasitic on there being a 'cow'-cows relation.

Fodor claims that his solution to the disjunction-problem is non-question begging because it is based on dependencies between nomic relations of properties, relations which are compatible with a naturalistic viewpoint and do not need to be formulated in representational terms. In sum, according to Fodor what makes a token a mental representation is that it bears a nomic relation with a certain environmental property which constitutes its referent, insofar as any additional relation holding between the representation and properties of the environment depends asymmetrically on the relation with its referent.

Taking together Fodor's solutions to the all-problem and the disjunction-problem, we can now sum up. According to the author, what distinguishes us from paramecia and other entities that lack mental representations, is that we can bear the right kind of nomic relations with certain environmental properties. Those relations are mediated by sustaining mechanisms that go far beyond perception, since they involve background knowledge and theories we have about the world. But since for the purposes of fixing the mind/world relation those mechanisms can be specified in computational terms, no appeal to other representational contents is required. And importantly, not any nomic relation between symbols and referent will do for the purposes of fixing content. The relation has to be asymmetrical, in the sense that any

other relation between the mental representation and environmental properties must depend, or be parasitic, on the relation between the mental representation and its referent.

3.5 Problems with Fodor's Second Line

When dealing with the all-problem Fodor appeals to background commitments (i.e. stored theoretical knowledge) in order to account for what fixes the nomic relations mental representations bear with their referents. As commented in 3.4.1, the author anticipates the objection that this could be question begging by arguing that it is not the mental representations that make up the theory but its computational architecture what fixes the relation. In Fodor's words:

> the content of a theory does not determine the meanings of the terms whose connections to the world the theory mediates. What determines their meanings is which things in the world the theory connects them to. The unit of meaning is not the theory; it's the symbol/world correlation however mediated. (p. 125)

Then the structure of the theory that mediates the representation/world correlation is supposed to be somewhat separable from the contents of its representations, and to make this detachment plausible Fodor appeals to the distinction between the syntactic and semantic components of computer systems. He claims that the structure of theories responsible for establishing the representation/world correlations that fix representational contents can be specified in purely syntactic, computational terms. Thus Fodor:

> The picture is that there's, as it were, a computer between the sensorium and the belief box, and that the tokening of certain psychophysical concepts eventuates in the computer's running through certain calculations that in turn lead to tokenings of 'proton' (or of 'horse' or whatever). (p. 123)

These computations involve inferences drawn over true beliefs, but what fixes the belief's contents is their relation with the world, not with other beliefs involved in the inferences. A consequence of this view is that even false theories would be capable of delivering mental representations, insofar as they ensure a reliable representation/ world correlation. Fodor sees this as an advantage of this theory since it makes possible

that different people—who might have disparate theories about the world—could share the contents of their mental representations. This idea of distinguishing between mechanisms that enable the fixation of content and those that determine the contents themselves, is further developed on Fodor's later work (e.g. 1998). In sum, what fixes the content of a representation is the nomological relation it bears with the property of the environment it denotes, however mediated, even if that relation has never been instantiated in the actual world. Meanwhile, the mechanisms that sustain or fix that nomological relation are supposed to be required for having content, but not relevant for determining the nature of the contents themselves.

However, I believe this alleged irrelevance of sustaining mechanisms for determining content is problematic, at least for our present purposes, because they seem to be as crucial as the nomic correlations themselves for explaining what makes genuine mental representations possible. Recall that we are trying to find out, in a naturalist context, is what it is for a computational symbol to have the content characteristic of mental symbols by its own right, without the need of an external interpreter. Fodor's project of naturalising a theory of content pursues a similar objective, viz. to articulate, "in nonsemantic and nonintentional terms, sufficient conditions for one bit of the world to be *about* (to express, represent, or be true about) another bit" (p. 98). His proposal is, in short, that the nature of mental symbols can be explained by appeal to their semantic relations with their referents, and gives an account of those relations in nomic and counterfactual terms. But is that response satisfactory?

I believe it is not. Consider the following example[21]. Long before the time of Newton, mariners knew that there was a correlation between the rise and fall of the tides, and the position and phase of the moon. But a complete account of that correlation —even if the account involves nomic regularities and counterfactuals—would be insufficient to explain the tides. The mariners had no knowledge of the causal connection between the moon and tides, and to whatever extent they thought they had an explanation, it had probably to do with Gods' benevolence or some other sort of

[21] I borrow this example from Salmon (1989, p. 47) who presents it as a counterexample to the *deductive-nomological model* of explanation. Briefly, he attempts to show that scientific explanations based on nomic correlations are at best incomplete without an account of the causal mechanisms underlying these correlations.

supernatural mechanism. It was no until Newton elucidated the casual connection that we had a proper explanation of the tides.

The same moral can be extended to Fodor's account of the nature of mental representations. He offers an explanation based on (asymmetrically dependent) representation/world nomic correlations. But as it happens with the tides, this explanation is poorly illuminating about what grounds these correlations. The sustaining mechanisms appear to be required to furnish them, but since Fodor's account quantifies over all the possible sustaining mechanisms that could fix those correlations, his explanation is left incomplete. Instead, it would be much more illuminating about the nature of mental representations to know the computational principles or limits under which the sustaining mechanisms operate. Let me illustrate why the omission of these mechanisms is problematic with another example.

Imagine that in the future, scientists are able to build a robot that bears mental representations. Even though this is still science fiction, it can be taken as a working hypothesis of CTM that a robot endowed with a computational architecture and information processing capacities of the right complexity should be capable of thinking. Following Fodor's view, the scientists should at least have equipped the robot with perceptual systems and central computational mechanisms capable of reliably connecting its inner symbolic structures with their referents, in a way that fixes the appropriate nomological relations between them. So if the robot is able to think, say, about horses, it would need to possess computational mechanisms for sustaining a 'horse'-horse correlation. And in order to be genuinely semantic, that correlation would have to be nomic and constrained in such a way that any non-horse property causing 'horse' tokenings has to be asymmetrically dependent on horses causing 'horse' tokenings.

But as the example shows, what makes the robot capable of bearing mental representations is not just its capacity to relate symbolic structures with its referents in the appropriate nomic way, but also that it has the right computational architecture, viz. the appropriate sustaining mechanisms. Fodor could contend that when it comes to the metaphysics of mental representations, what matters are the semantic relations and not the sustaining mechanism, which are just an "engineering" fact about the robot with no

implications for a theory of mental representation (1998, p. 78). However, I believe that the engineering problem is quite relevant for the purposes of drawing a line for mental representation. For consider counterfactual situations: if the sustaining mechanisms had not been within certain computational limits, then the robot's symbolic structure could have had a different content or no content at all. There appears to be some minimum constraints in the computational architecture of a system that are crucial for its capacity to support semantic relations. And it is interesting to point out that the importance of those constraints for understanding what is distinctive of creatures seems to be a fundamental assumption of CTM. For if the mind is a sort of Turing machine, and given that not any implementation of a Turing machine is capable of instantiating a mind, it follows that what makes a mind possible is the implementation of a Turing machine of the appropriate complexity (see Kim, 2006, p.133).

It could be objected that sustaining mechanisms do not constitute a metaphysically necessary condition for content, given it is conceivable that, say, angels, could bear semantic relations with properties in the world without the mediation of any sustaining mechanism. But certainly they at least conform a nomologically necessary condition; creatures in the earth as we know it cannot bear semantic relations unless their computational architectures satisfy some minimum complexity constraints. Therefore, when faced with the question of what are the minimum conditions for having mental representations in nomologically possible creatures (which is the question we are addressing in this thesis), the sustaining mechanisms (represented by computational constraints) become an essential part of the response.

3.6. Conclusions

Fodor's proposals about what makes for mental representation are illuminating about the complex nomic relations computational symbols have to hold with their referents in order to count as mental representations. Indeed, the author might be correct when claiming that semantic content is determined by the sort of nomic relations he describes. However, his last and more plausible account says little about the computational and information processing means that are required for instantiating symbolic structures with the relevant nomic relations, and for this reason I believe

Fodor's view is not very useful for the present purposes, viz. what are the minimum conditions for possessing mental symbols. More has to be said regarding the "engineering problem", viz. which is the computational architecture required for processing information and coding computer symbols in way that makes possible the emergence of mental representations. In the following section I will explore Dretske's view which goes much further in this respect.

Chapter Four

Informational Approaches: Dretske on Drawing the Line

4.0 Introduction

In the previous chapter I started discussing informational approaches to representation, by focusing on Fodor's proposals. The upshot was even though his last proposal is on the right track, it is not very revealing of the computational and informational processing facts that make possible the possession of mental symbols.

In this chapter I discuss the view of Fred Dretske, who has also developed an informational approach that provides much more detail about the computational and information processing mechanisms that generate mental symbols. I critically present Dretske's proposals in two moments of his work. Basically, the author claims that in order to become mental symbols informational structures have to pass through a process of digitalisation and be coded as a cognitive structure. I also discuss his latter work, where he takes learning as the crucial aspect of coding that makes possible to yield mental symbols.

I conclude that even though Dretske makes significant progress towards understanding how informational structures could become mental symbols, he draws the line for mental symbols too low and cannot successfully deal with some counterexamples. Again, more has to be said about the computational architecture of central cognition in order to state what is special about computational agent that possess mentality.

4.1 The Flow of Information: From Analog to Digital Form

Dretske is responsible for one of the earlier and most complete informational approaches to representation. Even though his view has evolved thought the last three decades, the main tenets of what he put forward in his book *Knowledge and the Flow of*

Information[22] remain current in his work. As most proponents of informational approaches, Dretske takes environmental information as the main precursor of mental representations, but at the same time recognises that it is not enough. Cognitive systems have to code or transform this "raw" information into symbolic structures in order to become genuinely representational.

The first stage in the flow of information from the environment to the mind is the sensory system. At this stage transducers encode information about environmental properties and do so by means of covarying with environmental properties in a nomological way (see 1.3). Recalling what we said in the previous chapter, what happens at the surface of sensory systems is not particularly distinctive of cognitive systems. The capacity of the retina to covary with light, or ear bones to covary with sound waves, is equivalent in terms of information encoding to what goes on in tree rings or fuel gauges. But even though those sensory states can be considered as the most basic manifestation of information-bearing state, Dretske claims that they are already instantiating a property that has traditionally been considered characteristic of mental representations: *intentionality*.

More precisely, the idea is that all informational structures are intentional because they can carry information about certain environmental objects without carrying all the information that can be extensionally attributed to them. For example, imagine that for unknown reasons all dogs happen to be infested with certain parasite; then every object that instantiates the property *is a dog* will happen to instantiate the property *has a parasite*. However, as Dretske points out a structure carrying the information '*x* is a dog' does not necessarily contain the information '*x* is a parasite', since the lawful correlation that grounds the flow of information is between the dog and the informational structure, and not between this structure and the parasite. Parasites are just contingently correlated with dogs, and so there is no nomic connection between them. To put it in a more philosophical fashion, the intentional character of information-bearing states is expressed in their being insensitive to extensionally equivalent information, viz. extensional principles such as the intersubstitutivity of co-referring expressions are not always satisfied. Dretske claims that at this sensory level we have

[22] Unless otherwise indicated, all references from sections 4.1 to 4.3 are to this book.

what he calls *first-order intentionality*, since the information coded by them appears to be directed to environmental objects in a way that is not fully extensional.

This is a controversial aspect of Dretske's account, mainly because it draws the line for intentionality quite below the level of mentality. However, at the same time he acknowledges that to instantiate intentional properties is not sufficient for instantiating semantic properties and therefore mental representations. Entities that bear just environmental information have intentionality only in its most primary, first-order, manifestation, while genuine mental symbols have what Dretske calls *higher-order intentionality* (p. 173; see also Dretske, 1980). Therefore, for present purposes let us explore where he draws the line between mere information-bearing states and internal states with that qualify as mental symbols.[23]

According to the Dretske, the crucial difference is given by the capacity to process and transform information in the right way. The main part in this process happens through the digitalisation of raw information carried out by the sensory system. In Dretske's words:

> It is the successful conversion of information into (appropriate) digital form that constitutes the essence of cognitive activity. If the information that *s* is *F* is never converted from a sensory (analog) to a cognitive (digital) form, the system in question has, perhaps, seen, heard, or smelled an *s* which is *F*, but it has not seen that it is *F*—does not know that it is *F*. (p. 142)

To unpack this quote is it useful to explain the notions of analog and digital form. At the sensory level, a continuous and massive amount of information is registered and only a fraction of it ends up encoded in a symbolic structure. The richness of sensory information is a result of the direct impact of energy coming from environmental properties, which according to Dretske is coded in analog form. He compares analog information at this level with our phenomenal experience of the world, which is "informationally rich and profuse in a way that our cognitive utilization of it is not" (p. 150). He gives the example of visual experience. Imagine we are in front of a scene of a crowd of youngsters at play. We see, say, 27 children, and many details such

---

[23] At this point it is appropriate to make a terminological remark. As mentioned in the paragraph, according to Dretske's usage the term intentionality does not imply mentality. In fact, the author regards intentionality, meaning and representation as properties that even simple artefacts can instantiate by virtue of their information-processing capacities. Instead, and as shall become apparent in the course of this chapter, Dretske draws the line for mentality on the possession of beliefs.

as their colours, relative location, size, etc. However, at first sight we probably are unaware of many of those details and cannot tell exactly how many children are there (over a dozen? around 30?). Even though the information that there are 27 children has already been registered, it is still in analog form and needs some filtering in order to be available for thought. And the point is not just about phenomenologically conscious thought, but about cognitive processing in general. Unless some filtering is carried out and particular information extracted and encoded in the appropriate format, we simply cannot process it beyond our sensory systems. According to Dretske, this filtering amounts to digitalisation:

> Until information has been extracted from this sensory structure (digitalisation), nothing corresponding to recognition, classification, identification, or judgment has occurred—nothing, that is, of any conceptual or cognitive significance. (p. 153)

The general idea is, then, that a minimum requisite for having mental symbols is to possess a "digital converter" that transforms analog information into a format suitable for cognitive processing. This is not to say that sensory experience should not be considered as part of cognition, though. The point is that systems which lack the resources to code information in the appropriate digital form lack cognition altogether. In Dretske's words:

> [I]n order to qualify as a perceptual state (seeing s) a structure must be coupled to a cognitive mechanism capable of exploiting the information held in the sensory representation. (p. 258, n. 29)

This idea should be considered as fairly straightforward insofar as many simple artefacts such as a mercury thermometer have the capacity to code analog information from the environment, while they clearly lack mentality. But at the same time, it must be noted that not any digital conversion will do for generating mental symbols, since digitalisation is also widespread in simple artefacts. Take a mercury thermometer. The position of the mercury column is a continuous variable that registers temperature in analog form. But imagine that electrodes are inserted in its tube in such a way that when the mercury reaches 5ºC their contacts are closed and an electric impulse turns on a light. Then the light would be carrying a signal with the information that 'the room is at 5ºC', which would be in digital form because it has only two informationally relevant states (on and off). This example illustrates why it is implausible to attribute mentality just by virtue of being capable of digitalising information. There are subtleties in the

way information is digitalised, which constitute the core of Dretske's proposal, and will be matter for the following section.

4.2 The Digitalisation Process

From a more orthodox computational viewpoint, digitalisation can be described as a computational mechanism that starts with environmental information coded by sensory systems in analog form, that after successive inferential processes is transformed into an informational structure. Dretske characterises informational structures in general as containing multiple layers of information, one nested under another. Each layer conveys a piece of information and more peripheral layers carry more specific information about the source, while the inner layers carry information that is more general. For example, suppose someone sees a dog that happens to be a brown poodle. Among the multiple information coded by her sensory systems is that '*x* is a dog', but also more specific information about the source such that '*x* is a poodle' and '*x* is a brown poodle' (fig 1a). During the digitalisation process, informational structures are coded in such a way that a specific piece of the information it contains is singled out. On Dretske's view, once this information has been "completely digitalised" it qualifies as the semantic content of the structure. To reach this level, the informational structure passes through processes of coding that involve *filtering* and *selective sensitivity*. Below I explain them in turn.

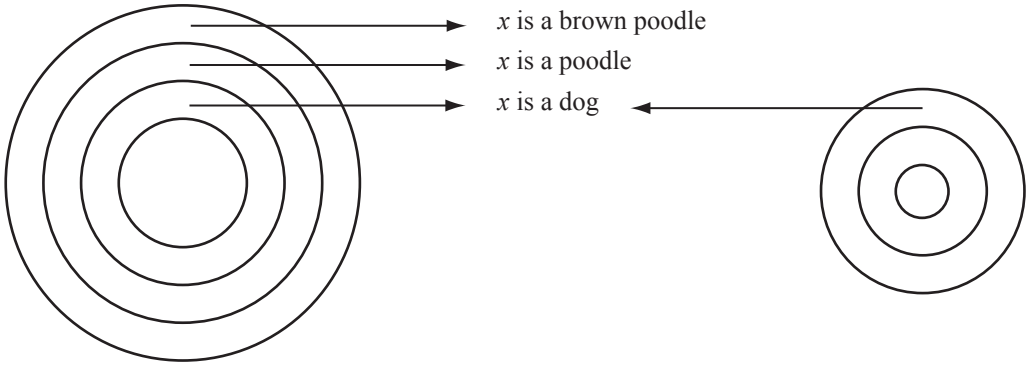

Figure 1a Figure 1b

Figure 1a shows a multi-layered informational structure that carries the information that '*x* is a dog' but also more general information about the source. Figure 1b illustrates an informational structure that has '*x* is a dog' at its more peripheral layer and therefore carries no more general information than this.

Filtering: when an informational structure undergoes digitalisation it loses information that is irrelevant for the purposes of isolating its semantic content. More precisely, information conveyed in the more peripheral layers of the structure is filtered, leaving the information corresponding to the semantic content at the outermost layer. Therefore, no more specific information about the source is nested in the informational structure. Returning to the previous example, imagine that the person that sees the brown poodle generates a symbolic structure that has '*x* is a dog' as its semantic content. During the conversion from analog to digital, its sensory information underwent a process where information more specific than '*x* is a dog' was filtered by the digital converter, in particular the information that '*x* is a poodle' and '*x* is a brown poodle' (fig 1b). Therefore, '*x* is a dog' ended up being at the outermost informational layer, viz. as the piece of information in which all other information is nested. According to Dretske, that means that this is the information that has been completely digitalised, and the one identified with the semantic content of the structure.

Selective sensitivity: as mentioned above, informational structures have many pieces of more general information embedded within them. This information can be nomically or analytically nested. Nomically nested information corresponds to pieces of information that are entailed by a natural law by the information carried by a signal. For example, if a signal carries the information that '*x* is a dog' then it must also carry the information that '*x* is a mammal' and that '*x* is an animal' (fig. 2a). On the other hand, analytically nested information is logically entailed. For instance, if a signal encodes the information '*y* is a square', it must also carry the information that '*y* is a rectangle' and '*y* is a parallelogram' (fig. 2b).

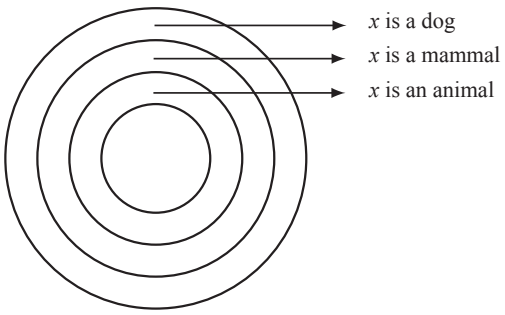

Figure 2a

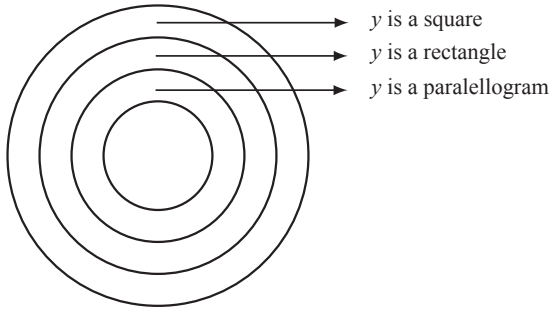

Figure 2b

Figure 2a shows a multi-layered informational structure that carries the information that '*x* is a dog' and also nomically nested information. Figure 2b illustrates an informational structure that carries the information that '*x* is a square' and analytically nested information.

While informational structures contain multiple layers of information, as explained above after the digitalisation process just one piece of information stands out from the rest, which corresponds to the one that has been completely digitalised and is situated at the outermost layer. Dretske claims that this is the only piece of information that is selectively sensitive (i.e. responsive) to that component of the incoming signal that defines the semantic content of the informational structure. In contrast, nomically or analytically nested information is not selectively sensitive to the information that is causally responsible for the production of the informational structure, at least not in the same way than the one that corresponds to its semantic content. Take again the example of a structure that has '*x* is a dog' as its semantic content. Even though that structure also carries the information about the property *x is a mammal*, only the property *x is a dog* is the one causally responsible for the production of a structure with that semantic content. For if *x is a mammal* had been the responsible for the generation of the structure, then its informational structure would had been different, one that has the information '*x* is a mammal' as its outermost layer.

The notion of selective sensitivity is also useful to tell apart the semantic content of a structure from information about the proximal events to which the delivery of this information depends. Any perceptual process that carries information about a signal also carries information about the means by which that information was produced. For example the information '*x* is a dog' covaries nomically with patterns of light in the retinal surface and with electric signals of neurones, and therefore carries information

about those properties nested in its structure. But again, it is only '*x* is a dog' that is selectively sensitive to the property of the environment that is accountable for causing an informational structure with that semantic content. As Dretske claims, all other information associated with its means of transmission depends on the causal link between the dog and the information that '*x* is a dog', while that link does not depend on any particular means of transmission, since multiple inferential routes and sensory modalities could had been effective to transmit the information that '*x* is a dog'.

At this point it is interesting to note some similarities with Fodor's view. First, the isolation of a specific piece of information that categorises a distal environmental property and abstracts from the proximal means of its production is equivalent to the process that gives rise to percepts. As explained in 3.2, percepts correspond to basic representations of distal properties that remain constant however the information that originates them is variable and incomplete. What makes those percepts capable of being about a distal property is the capacity of perceptual systems to encode information coming from that property and abstract from irrelevant information carried thought the coding process (i.e. generate perceptual constancies).

A second similarity is between the notion of selective sensitivity and Fodor's asymmetric dependence. According to Dretske's view, even though the piece of information a perceptual state has as its semantic content (i.e. the outermost piece) contains more information nomically and analytically nested in its structure, only the information carried by the semantic content is the one that has been completely digitalised. This corresponds to the piece of information that is causally responsible for the production of the perceptual state, and the one the perceptual system responds selectively to. On the contrary, nomically or analytically nested information is not selectively sensitive to the information that is causally responsible for the production of the perceptual state, and in this sense the nested information can be said to be asymmetrically causally dependent upon information the structure carries as its semantic content (cf. Adams, 2003).

4.3 Dretske's First line: Digitalisation and Cognitive Structure

Now I will focus on where Dretske situates the point in which the manufacture of informational structures gives rise to genuine mental symbols. An important part of that process corresponds to the encoding of environmental information picked up by sensory systems. As described in the section above, that information goes through digitalisation until one particular piece of information about the distal environment is completely digitalised. Dretske claims that that piece of information can be regarded as having semantic properties. One reason why it can be qualified as genuinely semantic is that it has what the he calls higher-order intentionality. Let me explain.

According to Dretske, first-order intentionality is reached by any informational structure since they do not carry all the information that is extensionally equivalent to the source (see 4.2). But as noted in the previous section, informational structures cannot avoid carrying additional information that is nomically or analytically nested in them. Any informational content that carries the information that '*x* is a dog' will carry the nomically related information that '*x* is a mammal', as well as the analytically entailed information of, say, '*x* is a canine'. However, by delivering a semantic structure the digitalisation process "*features* or *highlights* one of these components [of the incoming information] at the expense of others" (p. 181). As I explained when introducing the notion of selective sensitivity, Dretske claims that the piece of information that corresponds to the semantic content of a structure is primarily related with the source, in the sense that all the other (nomically or analytically) nested information depends on its relation to the source in order to generate a state with that structure. Then for example, a structure having the semantic content that '*x* is a dog' is sensitive to the property of being a dog in a way that grounds any other relation between the structure and other properties of the dog.

Then the informational structure reaches what Dretske calls higher-order intentionality, insofar as the principle of intersubstitutivity of co-referring expressions fails to apply to that structure. This happens not just for the case of contingently associated information—as it occurs with first-order intentionality—but also for nomically and analytically embedded information. Dretske sees having higher-order intentionality as a crucial step informational structures have to take towards the

acquisition of genuine mental properties, and indeed claims that "to qualify for *cognitive* attributes a system must be capable of occupying *higher-order intentional states*" (p. 172). And this is a consequence of the process of digitalisation, which yields structures with completely digitalised information that corresponds to their semantic content. Dretske summarises this idea by contrasting how a television codes information with how humans do:

> The crucial difference between the human viewer and the instrument is that the instrument is incapable of digitalizing this piece of information in a way a human viewer is. The television receiver slavishly transforms the information available in the electromagnetic signal into a picture on a screen without ever imposing a cognitive, higher-level intentional structure on any of it. (p. 183)

But even if a television is endowed with the capacity to completely digitalise some piece of information, this will not be enough for delivering a mental symbol. To completely digitalise a piece of information is a necessary, though not a sufficient condition for giving it a cognitive structure. Dretske forcibly makes this point in the following passage:

> I believe this is a mistake—a mistake fostered by a confusion of information-carrying structures on the one hand and genuine cognitive structures on the other. Even if we grant that the output of these preliminary neural processes has a semantic content, this does not, by itself, qualify them for *cognitive* status. For unless these preliminary semantic structures have a hand on the steering wheel, unless *their* semantic content is a determinant of system output ... they do not themselves have cognitive content (p. 200)

Then what Dretske understands as a genuinely cognitive informational structure is one that has semantic content, and at the same time has this content with a functional role within the system that determines behaviour. If an informational structure exercises no control over the output, then it does not qualify as a symbolic structure with any cognitive significance, and with that Dretske refers to structures like concepts, beliefs, and cognitive states that characterise thought and knowledge in general. Then I take it, by switching to my own terminology, that according to Dretske the lack of functional roles also disqualifies a structure to count as mental symbol.

An important motivation Dretske has for stressing the need of functional roles for having genuine cognitive states is that mere informational structures cannot, even if completely digitalised, carry false semantic contents. This can be understood in terms of

the disjunction-problem discussed in the context of Fodor in 3.4.2. The problem emerges because a genuine mental symbol is supposed to be possibly tokened by information coming from environmental properties distinct from the one that constitutes its referent. This gives rise to the problem of either defining its content as a disjunct denoting all the properties that could possibly token that symbol, or to deny that it can be tokened by information other than the one coming from its referent (i.e. cannot misrepresent). Both alternatives are implausible, given that mental symbols are standardly supposed to denote particular properties instead of a disjunct, and to be capable of misrepresenting their referents.

In order to account for misrepresentation, Dretske resorts to the functional roles of cognitive structures. Besides their semantic contents, cognitive structures also have what the author calls their *information-carrying role* (p. 192), which corresponds to a general type of cognitive structure acquired during a period of development or learning (L). During that period, certain information is completely digitalised and the system thus becomes selectively sensitive to that piece of information. Besides, the structure is attached to an information-carrying role in the system such as that of discriminating and identifying the property of the environment that is the source of that information. The outcome of this process is the fixing of a type of cognitive structure with the appropriate semantic content, which can then be tokened in new structures that inherit the same structure-type. In Dretske's words:

> Once this structure is developed, it acquires a life of its own, so to speak, and is capable of conferring on its subsequent tokens (particular instances of that structure type) *its* semantic content (content it acquired during *L*) *whether or not these subsequent tokens actually have this as their informational content* ... In short, the structure type acquires its meaning from the sort of information that led to its development as a cognitive structure. (p. 193)

Since those subsequent tokens need not carry the information that generated the structure type, those tokens can misrepresent. This is because they still count as having the same content, in the sense that they are tokens of a structure type with that semantic content, even though they might not actually be carrying information that corresponds to that content. Dretske sometimes uses the term *meaning* to refer to semantic content, as a way to highlight its capacity to be tokened by a structure carrying the wrong information or even no information at all (in which case can be said to be carrying "putative information"; p. 262, n. 8). For present purposes, though, the important point

is that we cannot have genuinely meaningful symbols without cognitive structures. The most that can be delivered from the digitalisation process is a completely digitalised informational structure, which might have the appropriate level of intentionality required for being semantic content, however it will lack genuine meaning insofar as informational structures as such cannot misrepresent. Their contents only become semantically significant when they are tokened in a cognitive structure that derives from the right structure-type.

4.3.1 Problems with Dretske's First Line

To evaluate how plausible is Dretske's criterion for drawing a line between mental and non-mental computational entities I shall examine whether his criterion would safely rule out entities that clearly lack mentality. Let us start by considering a simple artefact such as a refrigerator thermostat. This artefact consists on a mercury thermometer with electrodes inserted in its tube, so that when the mercury reaches 5ºC their contacts are closed and an electric impulse is transmitted to a cooling system, which is turned on until the contacts are opened again. In this way, the thermostat keeps the temperature inside the refrigerator constantly below 5ºC.

Thanks to the capacity of mercury to nomically covary with the thermal properties of its immediate environment, the thermometer instantiates analog informational structures that carry information about the temperature in the fridge and that at least have first-order intentionality. But the artefact is also capable of some degree of digital encoding. Imagine that the mercury column of the internal thermometer suddenly reaches the level marked as 6ºC and so an electric impulse is generated by its contact with the electrodes that is then transmitted to the cooling system. This electric signal carries the information that '*r* is over 5ºC', which has been digitalised from the analog information coded by the thermometer. For even though the mercury column of the thermometer already bears the information that '*r* is over 5ºC', it has it coded in analog form because is nested on the more specific information that '*r* is at 6ºC'. What happens is that when this analog signal passes through the electrical circuit, the information that '*r* is over 5ºC' is digitalised since more specific information (such as that '*r* is at 6ºC') is lost and so at this point the circuit does not convey more

specific information about how much over 5ºC the room's temperature actually is. (cf. p. 140)

So the thermostat could be regarded as a candidate for having semantic content since, at least to some degree, it is capable of digitalising a precise piece of information about the environment and delivering it to the cooling device. However, Dretske claims this is not the case since this piece of information has not been completely digitalised. There is digitalisation, since some process of filtering is going on, but the digitalisation is not complete insofar as the piece of information picked up by the electrodes and transmitted by the circuit is not selectively responsive to the information that should correspond to its semantic content, viz. that '*r* is over 5ºC'. Let me explain.

First, when the information that '*r* is 5ºC' reaches the cooling device it has not been completely digitalised because it is nested in more specific information related with the more proximal structures involved in the generation and transmission the information-bearing state, such as that the electrodes are in contact, the amount of current generated by them, the magnetic field that reaches the cooling system, etc. The cooling device is turned on in virtue of receiving information about temperature, but also in virtue of receiving information about these more proximal events (cf. p. 187). So, the argument goes, the information-bearing state never encodes the information that '*r* is over 5ºC' in way specific enough to constitute semantic content. In Dretske's terms, there are larger information layers in which this information is nested, and therefore the informational structure never carries the information that '*r* is over 5ºC' as its outermost informational layer.

A related reason is that the thermostat is a device that has been built to always respond to the same kind of information-bearing state, which, as previously noted, conveys information about the intermediate events by means of which the production and delivery of this information depends. So the information-bearing state cannot abstract from this more proximal processes to encode a more specific and distal piece of information about the world. Entities which produce states with semantic content, on the other hand, should be plastic enough to extract the same piece of information about the world from a variety of different physical vehicles that may deliver this information, in the same sense as cognitive systems can encode the same semantic content from

information conveyed by different sensory modalities (p. 188). Again, the informational structure is not selectively responsive to the information that '*r* is over 5ºC', because it carries that information in virtue of carrying information about the proximal means by which that information was produced and transmitted. That additional information is not asymmetrically dependent on the piece of information that has been digitalised, and therefore the latter lacks the higher-level intentionality required for qualifying as semantic content.

Having shown that the thermostat can be safely ruled out from the scope of mentality, let us now examine a more complex artefact, a vending machine. I refer to a vending machine of the same type as the one described in 3.3.1, when discussing Fodor's proposal. Even though a vending machine is more complex than a thermostat, it is also certainly incapable of instantiating any property that would go beyond mere information-bearing states, such as semantic or mental properties. To attribute mental symbols to vending machines would be to extend the psychological domain far beyond what seems plausible.

As with the thermostat, the vending machine is surely capable of digitalising information. It converts information about different features of the coins (i.e. weight, size and magnetic content) to an electric signal that carries information about which type of coin was inserted. This last information is coded in digital form because more specific information about features of the coin was filtered during the process, leaving an informational structure carrying information such as that '*x* is a one pound coin'. Now the question is whether the vending machine can carry a piece of information in completely digitalised form, and therefore have an internal state capable of bearing semantic content. Can the information that '*x* is a one pound coin' be selectively sensitive to that property of the coin inserted in the machine? Can this information be carried by the informational structure in a way that is not dependent on the proximal means that mediate its production and transmission?

If the internal state that bears the information that '*x* is a one pound coin' is completely digitalised then it should be possible for the machine to distinguish it from other states that may carry information that matches some of the information embedded in the former digital state. For example, the information that '*x* is a coin' would be

embedded in any state which also bears the information that '*x* is a one pound coin', since the former is (let us say) analytically nested in the latter. If the machine shows a behaviour that is caused selectively by the information '*x* is a one pound coin' and not by '*x* is a coin', this would provide us with evidence that at least the machine has two ways of encoding the information about a coin: one that digitalises the information that '*x* is a coin' and other that digitalises the information that '*x* is a one pound coin' (and of course also has the information that '*x* is a coin' nested in its informational structure).

I believe it is easy to show that the machine can make this discrimination. First, its coin detector can distinguish genuine coins from fake ones and from other objects that may fit into the coin slot, through its capacity to measure the weight, size and magnetic content of the object. Secondly, every time it is loaded with a valid coin it accepts it, thus responding to the property *x is a (UK) coin*. But also, it can identify different kinds of coins, running from one penny to two pound coins. So besides identifying coins, it can sort one pound coins from that set. This shows that the machine can digitalise the information that '*x* is a one pound coin', since it can instantiate two states that share the information '*x* is a coin' and sort one of them in virtue of carrying a piece of more specific information about the coin.

However, this still does not show that a piece of information has been completely digitalised, since it could be nested in more specific information about more proximal events involved in the generation and transmission of the information-bearing state. As I explained with the example of the thermostat and the way it encodes the information that '*r* is 5Cº', some devices have a fixed architecture that picks up information always in the same way. A thermostat cannot abstract certain piece of information from information related with its means of production, and so is not plastic enough to extract the same type of information from different kinds of incoming signals, something that according to Dretske is one of the main characteristics of systems that can encode states with semantic content. I believe that this argument works for the case of the thermostat, but I will show that it cannot be applied to the vending machine.

Let us examine how the vending machine encodes the information that '*x* is one pound's worth'. The machine dispenses a beverage selectively when loaded with one

pound's worth, and can produce the same output in response to different possible inputs, which could be any of the 4563 possible combinations of coins ranging from 1 penny to 1 pound that could make for a pound. Moreover, in order to dispense the right change the machine can carry out an algebraic operation where one of the quantities stands for *being one pound's worth* (see 3.3.1). These observations not only suggest that the machine is really instantiating an information-bearing state which carries the information that '*x* is one pound's worth', but also that this information can be sorted out from a variety of different inputs (and in case you are not satisfied with all the combinations of coins, just imagine a vending machine that can also scan banknotes). Therefore it seems plausible to grant the machine with the plasticity to extract the information that '*x* is one pound's worth' from a variety of signals, so isolating it from information about more proximal processes related with the input. All this suggests that the vending machine carries the information that '*x* is one pound's worth' in completely digitalised form, and therefore would have an information-bearing state with that semantic content. I believe this works as a counterexample to Dretske's proposal, as mentioned above.

4.4 Dretske after 1981

From the viewpoint of scientific realism, the metaphysical status of mental states is vindicated by their causal role in scientific theories that explain behaviour. In other words, it is in the context of psychological theories where psychological notions acquire ontological status. Otherwise, if notions such as mental symbols and reasoning processes play no role in behavioural explanation, their metaphysical condition appears to be close to epiphenomenalism.

In his work following 1981, Dretske[24] (1988, 1999) adverts to some of these ideas and puts behaviour at the centre of his theory of mental symbols. He thus shifts his attention from the informational origins of mental symbols, to the causal or explanatory role information plays in behaviour. As Dretske (1994) says:

[24] In his 1986, Dretske put forward a somewhat different (teleological) approach, however it was later abandoned for his 1988 account.

> [I]f information has to do with the nomic dependencies between events, then for information to do any causal or explanatory work in the world these dependencies have to do some causal or explanatory work in the world. (p. 262)

This view can be regarded as continuous with his previous work described above. On the one hand, Dretske regards mental symbols as originated by the effect of digitalised information on the crystallisation of cognitive structures with certain information-carrying role. On the other, he still believes that to count as mental symbols it is crucial for internal states to yield some functional role in behaviour. However, Dretske now focuses on how to link this informational approach to the origins of mental symbols with the role they have in explanations of behaviour. On his present view, what characterises the behaviour of mental creatures is that it is the expression of intelligent thought and purpose (or is the product of genuine *agency*, as he says in 1999). Therefore, if information is to be relevant for explaining behaviour, it must be causally linked with the mechanisms that give rise to purposeful behaviour. Let me explain this idea through examples.

Take again the case of the thermostat. According to Dretske's account, when a thermostat switches on a cooling device in response to a rise in the refrigerator's temperature its behaviour is caused by a state with the semantic content of '*r* is over 5ºC'. Or to use a different example, when I delete some letters that appear in the screen of my laptop by pressing the key DELETE, what causes my laptop to do so is an internal state it has that means 'delete a letter'. But even though the states responsible for these behaviours have a content, these are not purposeful actions in Dretske's sense. This is because, he argues, what is relevant to explain the production of these behaviours is not the content of the states, but some intrinsic (physical or functional) properties of the artefact.

This is clear from the fact that even if we change the informational contents of these states the behaviour of the artefacts would remain the same. For example, we could change the thermostat's state meaning to '*r* is too hot' or 'turn on the cooling device!' and it would continue switching the device on in the same way, because what really causes this behaviour is that (say) its mercury column reaches the electrodes and closes their contacts. And the same applies to the laptop's key; irrespective of its contents, what really explains the deleting behaviour is an algorithm that runs patterns

of zeros and ones in the laptop's processor. Then the question turns to what makes it possible for some entities to have states whose meanings are the real causes of their behaviours. Dretske summarises the whole idea in the following passage, suggesting that we might find purposeful action in animals:

> Machines don't think, and so nothing they do is governed by what they think, but their behaviour is sometimes controlled by internal states with a meaning remarkably like that which controls the behaviour of intelligent agents. Seeing exactly what is missing in the case of machines—why meaning doesn't actually control their behaviour—will give us a better understanding of how meaning gets its hand on the steering wheel in animals. (p. 23)

4.5 The Structuring Causes of Behaviour

Before going into the details of Dretske's present account about what makes for mental symbols, it is important to introduce a distinction the author formulates between two alternative causal explanations of behaviour (Dretske, 1988). When discussing the role alternative causal events (C) have in our explanations of certain behaviour (M), he writes:

> In looking for the cause of a process, we are sometimes looking for the triggering event: what caused the *C which* caused the *M*. At other times we are looking for the event or events that *shaped* or *structured* the process: what caused *C* to cause *M* rather than something else. The first type of cause, the triggering cause, causes the process to occur *now*. The second type of cause, the structuring cause, is responsive for its being *this process*, one having *M* as its product, that occurs now. (p. 42)

A typical example of triggering cause can be found in the behaviour of artefacts described in the previous section, which is properly explained in terms of the proximal physical or functional events that take place inside them. By contrast, the structuring cause focuses on the historical events that configured the internal events of the machine to have their current structure and behave the way they do. For example, the thermostat switches a cooling device on in response to temperature because it was designed, by its creator, to perform that way.

Dretske elaborates this distinction in his analysis of the role of semantic content in the causation of behaviour. As previously noted, the thermostat's state that controls the cooling device has a content, but it is not this content which explains the behaviour.

This is because the triggering causes of the thermostat's behaviour are physical events which are themselves not sensitive to content. But if we look at the structuring cause then we find out that meaning is really playing a role, because the creator of the thermostat was an agent that had the purpose of building the artefact to perform its function. Meaningful events inside the head of the creator shaped the structure of the thermostat, and thus are responsible, at least in this historical sense, of its current behaviour. But the point is that purposeful action is coming from "outside" since there is nothing inside the thermostat, no intrinsic properties of it, that have a meaning with causal powers over its behaviour.

An intriguing point is that structural causes are not always due to human creators. This is the case with biological organisms, whose intrinsic properties were designed by (presumably) processes of natural selection where no thinking or agency took place. For example, the hypothalamus functions as a thermostat to keep body temperature constant at about 37ºC. The triggering causes that explain its behaviour involve physical states that register body temperature which, following Dretske, have content. However, in contrast to the artificial thermostat the structuring cause of this biological thermostat does not involve purposeful agents but a gradual process of natural selection. The mechanisms that govern the behaviour of the hypothalamus are not dependent on the purposes of any thinking being. Does this confer on the hypothalamus genuine agency and therefore mental symbols?

Dretske's response would be no, because the structuring cause of the hypothalamus behaviour is still coming from "outside", this time not from some human mind but from a history of selection over the behaviour of previous organisms that evolved the hypothalamus. The meanings that may be found in the internal states of this biological thermostat have been fixed by processes that happened long before the existence of the organism that actually possesses them. As happened with the artificial thermostat, nothing internal to the hypothalamus explains why some of its states have meaning. These meanings have been fixed through structuring causes that are out of the control of the organism who has the hypothalamus, and therefore cannot be governed by it to produce purposeful behaviour. As the Dretske (1988) says, even though the cause of its behaviour "has meaning of the relevant kind, this is not a meaning that has *to* or

for the animal in which it occurs. That, basically, is why genetically determined behaviors are not explicable in terms of the actor's reasons." (p. 95)

4.6 Dretske's Second Line: Learning

Dretske claims that it is in simple cases of animal learning (viz. conditioning) where meaning starts to play a genuine explanatory role in behaviour. In the following I will focus on his paper *Machines, plants and animals: the origins of agency* where he directly addresses the issue of what makes a creature capable of thought and purposeful behaviour. There he gives the example of a foraging bird that learns to avoid a poisonous butterfly:

> A foraging bird tries to eat a Monarch butterfly. This butterfly has been reared on a toxic form of milkweed. Such butterflies are poisonous and cause birds to vomit. After one nasty encounter, the bird avoids butterflies that look like the one that made it sick. A day later our bird sees a tasty Viceroy, a butterfly with an appearance remarkably like that of the noxious Monarch. The Viceroy, though, is not poisonous. It has developed this coloration as a defence from predatory birds. It mimics the appearance of the Monarch so that birds will "think" that it, too, tastes nasty and avoid it. Our bird sees the Viceroy and flies away. (Dretske, 1999, p. 27)

Dretske argues that in this case it seems natural to say that the bird avoids viceroys because it appears to believe that the bug it sees tastes bad, and thus to regard the bird as having agency. It has an internal (perceptual) state that means (say) 'M-looking bug', and which explains why the bird flies away when it encounters one of these butterflies. But, what distinguishes this example from the case of the thermostats? After all, in both the artifactual and the hypothalamic thermostats there is an internal state that means that the temperature is too high and that is responsible for activating some cooling mechanism.

The key difference is that, as previously noted, in the case of thermostats the meaning of their internal state is not responsible for the behaviour since this meaning comes from "outside" the entity/organisms that possesses the thermostat, this "outside" been understood as not within the scope of the actual engagements the organism has with the environment. So, the argument goes, the structuring cause of the thermostat's behaviour relies on a human creator or natural selection, and nothing inside it has

determined the meanings of its relevant internal states. This contrasts with the case of the bird. Since this animal is capable of learning, an internal event that has occurred to the organism itself (i.e. perception and memory registration) is now implicated in behaviour. What makes the difference is that the meaning of the event has been structured from "inside" the organism, in contrast with the case of thermostats where meaning had been fixed from antecedent events that were not caused by the entity/ organism itself.

Of course, the meaning and behavioural role of the internal states of a thermostat can be modified at the present time, for example by calibrating its thermometer or by adjusting its connections with cooling or heating devices. But these modifications, Dretske argues, are due to the designer's (our) purposes, and the events that reconfigured the artefact and thus explain its new behaviour (i.e. the structuring cause) were never an achievement of the thermostat. The internal states may be meaningful for us, but not for the thermostat itself. On the other hand, in the case of the bird the fact that the structuring cause of its behaviour is the product of its internal responses to past experience, makes its internal states meaningful for the bird, and so directly implicated in its behaviour.

4.6.1 Problems with Dretske's Second Line

In the paper under discussion as well as in his 1988 book, Dretske draws the line for mental representation at the capacity to learn. His conception of learning is basically behavioursitic, in particular operant conditioning. A critique that could be made against Dretske is that this kind of learning requires representation (Gallistel, 1990), and therefore it begs the question about where representation begins. If the internal states that give rise to learning are already representational and thus have meaning, then learning cannot be the instance where mental representations emerge (Burge, 2010).

I believe this critique is misguided and probably leads to a futile terminological dispute about what is meant by *representation*. From his early writings (e.g. 1980) Dretske has been quite liberal concerning the attribution of representations and, indeed, the same can be said with the attribution of meaning in his 1999 paper. He regards

internal states that bear *natural meaning* as meaningful, however he is clear in his purpose of distinguishing between these representational states and genuine thoughts, the latter understood as (higher-level) internal states with a content of the same kind as beliefs. It is thus a mistake to say that Dretske cannot appeal to representational states involved in learning processes, since these processes are supposed to give rise to higher-level representational states which do develop mental properties not present in their predecessors. Perhaps it would be convenient to reserve the term representation for *mental* representations, and call other informational or computational states just informational structures (as I have been doing so far), but this is just a terminological issue that does not undermine Dretske's attempt to draw the line for mental representation at the level of learning.

Having said this, I believe Dretske's proposal is too liberal. If the threshold for purposeful action and thus to mental representation is situated at the capacity to learn by operant conditioning, then we would have to attribute mental representations to some artefacts and animals which intuitively lack mentality. Starting with artefacts, it is certainly possible to program some robots with learning algorithms that make them able to develop behavioural effects similar to classical and operant conditioning. For example, the robot Amelia (Touretzky & Saksida, 1997) was designed to "learn" to sort objects into bins based on colour. The robot had to be trained by receiving a reward signal when its desired response occurred, and after short period it was able to discriminate the objects based on colour and to drop them in bins at certain location, in accordance with the desires of the trainers. The experimenters concluded that Amelia exhibited most of the hallmarks of operant conditioning. In addition, robots SAIL and Dav, based on a connectionist architecture which had not been previously programmed for any particular task, have shown to be capable of learning a variety of skills such as autonomous navigation and speech recognition (Weng, 2004).

If we take experiments like this seriously, then we should acknowledge that some computational artefacts can develop forms of learning of the sort Dretske describes as constitutive of purposeful action. However, this appears to be implausible, as Dretske (1999) himself recognises when he states that "machines don't think, and so nothing they do is governed by what they think, but their behaviour is sometimes controlled by internal states remarkably like that which controls the behaviour of

intelligent agents" (p.22). We could then interpret Dretske as claiming that machines such as Amelia are just mirroring genuine learning, and so are a mere simulation of purposeful behaviour. But it is wrong to interpret him that way, since he intends to draw the line for genuine thought on the capacity to learn and not in some additional distinction between machines and animals that would make the former a simulation of the latter.

As mentioned, a similar counterargument can be put forward by appeal to animals. Recall the habituation response of *Aplysia* presented in section 1.4.3, which is a basic form of learning. But most notably for present purposes, *Aplysia* is also capable of associative learning such as classical conditioning (Hawkins & Kandel, 1984) and therefore is comparable to the case of learning by the foraging bird presented by Dretske (and discussed above). But as shown in 1.4.3, these sorts of learning behaviour can be satisfactorily explained in terms of the physical domain, and there is no justification to deploy symbolic and computational processing to account for them. Even if we take into account biological organisms such as *Aplysia*, learning as such does not appear to be a safe place to draw the line that marks the origins of mentality.

4.7 Conclusions

In the context of informational approaches to representation, Dretske provides one of the most detailed proposals about how information could be regarded as the basic ingredient for making a mind. He goes deep into the engineering problem of understanding how the gap between informational structures and mental symbols could be bridged, certainly deeper than Fodor does. One distinctive aspect of his view (in particular the one presented in 4.6) is, however, that he includes the aetiology of symbolic structures as relevant for determining which of them count as mental symbols. But irrespective of the particular problems of etiological approaches to mental symbols in general (see 2.3.2 and the next chapter for discussion), I have argued that Dretske's proposal draws the line for what makes for mentality too low, and is thus susceptible to counterexamples. If his view implies that vending machines and simple animals such as *Aplysia* have mental symbols, then it ends up being too liberal and needs further refinements. I believe that those refinements come in the way of adding more

complexity to the computational architecture of central cognition, and explaining how in that context symbolic structures could play the functional roles that characterise psychological explanations. But this shall be explored in the remainder of this thesis.

Chapter Five

Teleological Approach: Burge on Drawing the Line

5.0 Introduction

Tyler Burge has put forward an alternative way to draw the line for the origins of mental representations. His proposal shares the basic tenets of computational and informational approaches presented in the previous chapters, and like Fodor he claims that percepts can already be considered genuine mental representations. However, he adds teleology to his account, by arguing that the capacity of perceptual systems to yield mental representations is grounded on what he calls *representational functions*.

In the present chapter I review Burge's view, beginning by showing how it departs from previous informational and teleological approaches, to then present his own proposal and raise two objections to it. Overall, I conclude that even though some notions he introduces—such that of *agency*—can be useful for the purposes of this thesis, Burge's notion of representational function is problematic and ends up drawing the line for mental representation too low.

5.1 Burge's Project in the Context of CTM[25]

> Perceptual representation is where genuine representation begins. In studying perception, representational psychology begins. With perception, one might even say, mind begins. (Burge, 2010, p. 367)

In his recent book *Origins of Objectivity*, Tyler Burge (2010)[26] develops an account of the minimum conditions for having mental representations. He claims that the most elementary forms of mental representation are already present in perceptual systems. He calls these forms *objective empirical representations*, viz. the

[25] Parts of this chapter are adapted from my article *Is perception representational? Tyler Burge on perceptual functions.*

[26] Unless otherwise indicated, all references to pages correspond to this book.

representation of basic environmental kinds, properties or relations[27]. Burge presents empirical evidence that suggests that perceptual systems appear to be widespread in the animal kingdom, even in phylogenetically primitive animals such as arthropods. Therefore, the author contends that the line for the origins of mental representation should be drawn at a low stage in the evolutionary tree of life.

Burge grounds his account in scientific work on perceptual psychology in accordance with the cognitive tradition. Accordingly, he agrees with the basic tenets of the computational approach to perceptual psychology that I described in 3.2 in the context of Fodor's account. Here is a brief description of it:

> The current Establishment theory (sometimes referred to as the "information processing" view) is that perception depends, in several respects presently to be discussed, upon inferences ... And since, finally, the Establishment theory holds that the psychological mechanism of inference is the transformation of mental representations, it follows that perception is in relevant respects a computational process. (Fodor & Pylyshyn, 1981, pp. 141-142).

According to this theory (which I shall call *computational approach to perception*) perception is a threefold process composed by transducers, input systems, and mechanisms of belief fixation. Information coming from the environment is encoded through transducers and then transformed by input systems into perceptual categories or percepts, which as mentioned in 3.2 are taken as basic categories of properties of the environment. A characteristic feature of percepts is that they can represent environmental invariants. This is possible thanks to the mediation of the inferential mechanisms of input systems, which constrain the possible interpretations of the sensory input in order to yield a constant perception of distal environmental properties. As it was for Fodor, according to Burge this is the stage of information processing where genuine mental representations are formed, or in his own terms, where objectification occurs. In Burge's words:

> A perceptual system achieves objectification by—and I am inclined to believe *only* by—exercising *perceptual constancies*—given, of course, the background of relations to the environment through individual functions just sketched. (p. 408)

For an explanation of the "individual functions" mentioned in the quote we will have to wait until section 5.5. For present purposes, what I want to highlight is that

[27] To simplify the exposition, in the rest of this chapter I shall just refer to environmental *properties*.

percepts, viz. the stage of perception where perceptual invariants are detected, are considered by Burge as the primary form of mental representation.

But even though there are evident similarities, Burge's view differs from Fodor's in some important respects. Burge contends that informational approaches to representation are deflationary in the sense of being too permissive at attributing mental representations. I shall critically present his arguments on this matter in 5.2. Besides, Burge goes on to say that to explain the emergence of genuine representations some teleological notions have to be added. However, he departs from mainstream teleological theories by not relying on a biological notion of function. In 5.4 I discuss Burge's arguments on this respect. When it comes to his positive view, the Burge develops a teleological approach to perceptual functions which he calls *representational function*s. I present his proposal in section 5.5 to then raise some objections to it in the final sections of this chapter.

Before passing to the next section, it is convenient to make a terminological note regarding Burge's use of the term "representation". Contrary to authors such as Fodor or Dretske that have no problem with saying that, for example, thermometers can represent the temperature in a room, Burge proposes to restrict the use of the term "representation" to (scientific) psychological explanations. He grounds his view in a form of scientific realism like the one I presented in 1.1.2, thus limiting the use of the psychological term "representation" to account for events that are better explained by the psychological level. So to facilitate the exposition of Burge's view, in the remainder of the present chapter this term shall be considered as equivalent to mental representation or mental symbol.

5.2 Against Informational Approaches

As noted above, Burge's proposal builds upon some principles of the computational approach. He also endorses some ideas from informational approaches to representation, at least in the sense that information-bearing states are precursors of perceptual representations. He accuses, however, informational approaches such as Fodor's and Dretske's of being "deflationary" in the sense that they draw the line for

representation too low by describing in representational terms the behaviour of animals and artefacts that clearly do not demand a psychological explanation. Following the path of scientific realism, Burge contends that in those informational approaches "representation is to be assimilated to notions that have no *distinctive* theoretical relation to psychology as it is ordinarily understood" (p. 293).

Recalling the discussion of previous chapters and along general lines, informational approaches are formulated by following the distinction between sensory information and percepts. While sensory information is directly correlated with proximal environmental properties, percepts are inferentially mediated. Some authors have proposed that the inferential route that runs from sensory information to percepts can be regarded as setting a normative standard for what is an accurate percept, and interestingly open the possibility of error. As Fodor and Pylyshyn (1981) put it:

> The standard approach to this problem within Establishment theories is to connect misperception with failed inference ... These inferences depend upon generalisations gleaned from past experience, and the generalisations are themselves nondemonstrative, and hence fallible. (p. 153)

So the idea is that when a percept is the result of the right inferential process then it is accurate, and when this process fails but at the same type the percept is instantiated then we have case of misrepresentation (cf. p. 92). One problem with this idea is to determine which is the right inferential process without begging the question by presupposing what is normatively correct or incorrect. A common way to develop this view is to link perceptual accuracy with some sort of regularity of statistically constant inferential route, thus explaining misrepresentation in terms of statistical atypicality. But Burge (p. 299) replies that causal-inferential routes cannot be considered right or wrong by themselves, and infrequent or abnormal percepts need not be mistaken. It is perfectly possible, for instance, for a perceptual system to perceive accurately certain environmental object even though it is highly infrequent or even if it has never appeared before. According to Burge, the only way to assign normativity to perceptual processes is to supplement them with teleology. But before considering this idea let us examine how Burge targets the specific informational approaches of Dretske and Fodor.

As explained in the previous chapter, Dretske considers learning as a crucial step towards the conversion of information-bearing states to mental representations. Dretske (1981) describes learning as the (main) way information can be crystallised into a cognitive structure capable of misrepresenting, while in later works he highlights that only through learning a creature can be regarded as self-determining some of its inner symbols and in consequence be ascribed as possessing genuine mental representations.

Burge raises several critiques to Dretske's view. The main one is similar to the critique I put forward in 4.6.1, namely that basic forms of learning such as associative conditioning are already present in artefacts and animals that clearly lack mentality. For example, Burge writes:

> Flatworms and snails exhibit habituation and trial-and-error association that straightforwardly meet the requirements of this conception of learning. [However,] I think that anyone who hoped to draw an interesting distinction between biologically functional information-carrying and some more psychologically distinctive kind of representation would not draw it just below snails and flatworms. (p. 307)

Another worry presented by Burge is that, in a way, learning also draws the line for representation too high. This is because nothing appears to rule out, as a matter of principle, that a creature who cannot learn but is equipped with innate perceptual mechanisms could yield representational states. This argument works for Burge since he grounds his view of perceptual representations in their role in (actual) psychological explanations and not in the aetiology of the respective psychological capacities. Then, according to him, representational states derive from perceptual mechanisms that fulfil certain functions irrespective of whether they are innate or acquired. I shall return to these issues in the next section and in 5.5 when presenting Burge's own positive view. Overall, I believe that in general Burge's critique to Dretske's proposal is compelling, and compatible with my own discussion of this proposal in the previous chapter. Now, let us examine how Burge deals with Fodor's version of informational approaches.

As described in 3.4, Fodor puts forward a theory of mental representations that grounds their semantic properties on nomological relations they bear with their respective referents. In order to be genuinely semantic, those relations have to be primary, in the sense that any other nomic relation between the mental representation and other environmental properties has to be asymmetrically dependent upon the one it

bears with its referent. Burge in a footnote (p. 307) takes issue with Fodor's view and presents two worries.

The first is that Burge regards as implausible to claim that there are laws connecting "higher" representational states with their referents, and that the only way to do so implies formulating those laws or law-like patterns in representational terms. To evaluate this critique let us consider Fodor's own discussion of higher (i.e. abstract) representational states such 'virtuous' (Fodor, 1990, p. 111). According to Fodor, what happens in this case is essentially the same as what happens with the instantiation of any other mental representation; the token 'virtuous' is caused by the property of *being virtuous*. But, is it plausible that *being virtuous* is a real property of the environment, to which we can bear nomic relations? Fodor would respond affirmatively, by appealing to the (inferential) mechanisms that sustain the relation between mental representations and their referents (Fodor, 1998). He contends that properties such as being virtuous have a real, however mind-dependent and relational, metaphysical status. They are real because there seems to be something like *being virtuous*, in the same sense than there is something like *being a dog* or a *being doorknob*. Indeed, Fodor considers it preposterous to deny that people can normally tell whether someone or something has one of those properties or not. And they are mind-dependent because to be virtuous "*just* is to have that property that minds like ours (do or would) lock [i.e. get fixed] to in virtue of experiences of typical instances of" *being virtuous* (Fodor, 1998, p.137).

Burge does not appear to be satisfied with a response of this kind, and in his second worry he points out to the case of uninstantiated properties (e.g. unicorn) to show how implausible he finds the appeal to laws connecting those properties with representational states. Without getting into the details, the way Fodor (1990) has replied to attacks of this sort by formulating his theory of content in purely nomological, and not causal, terms. This allows representations to be about uninstantiated properties even if there cannot be causal relations between them, insofar as they are nomologically linked. All he needs is that the property in question is nomologically possible, viz. that there are possible worlds where unicorns would cause tokens of 'unicorn' in our heads. Even though Burge does not address this reply directly, he considers Fodor's proposals unsuitable for a scientific account about mind and indeed "very remote from any actual theorising about representational phenomena" (p. 307).

I believe that instead of constituting a refutation of Fodor's proposal, Burge's considerations reflect a fundamental methodological difference between his view and Fodor's. Let me explain. Burge endorses a form of scientific realism according to which representational contents are individuated by way of being part of relevant explanatory distinctions made by perceptual psychology (p. 293). And since scientific explanations typically describe causal mechanisms, Burge assumes that the individuation of content has to involve causal relations with the environment. When it comes to representations of uninstantiated properties, he claims that those relations are indirect, mediated by their associations with other representations that do engage in causal relations with the world (p. 86). For example, 'unicorn' would be explained by appeal to its relation to representations of horses, horns, etc.

Fodor, on the other hand, makes no such commitment to scientific realism, at least for the purposes of individuating representational contents. Consider the following passage from Fodor (1994):

> [C]oncepts aren't individuated by the roles that they play in inferences, or, indeed, by their roles in any other mental processes. If, by stipulation, semantics is about what constitutes concepts and psychology is about the nature of mental processes, then the view I'm recommending is that semantics isn't part of psychology. (p. 122)

Fodor individuates mental representations (i.e. concepts) by specifying their contents, and since contents are not determined by their roles in scientific theories, they are not individuated by appeal to psychology. Of course, Fodor does not deny that mental representations are part of psychological explanations, since he indeed claims that psychological laws quantify over representational contents. But in contrast to Burge, he does not circumscribe his account of representational content to (causal) psychological theories. Fodor's theory of content is therefore open to include relations between representational states and uninstantiated properties, relations that however are metaphysically possible (viz. in possible worlds), they cannot figure in causal explanations of psychology. This is a debatable issue where I shall not enter, however it seems that Fodor's approach is at odds with my own commitment to scientific realism presented in the first chapter. In this sense, Burge's critique seems plausible, and compatible with the critique to Fodor I advanced in section 3.3.2.

5.3 Teleology Enters the Scene

Before addressing Burge's particular version of perceptual functions and as a way of preparing the following discussion, I shall introduce teleological approaches to perception in general. Teleological theories attempt to apply the notion of function to explain how perceptual systems work. To ascribe a system with a function means to understand its mechanisms as aimed at a certain end or goal (telos). A characteristic of functional explanations is that they are normative, in the sense that a functional system ought to perform its functions, and failure to perform them is a kind of error.

One advantage of ascribing functions to perceptual systems is that they could be used to explain the normative character of perceptual states. By ascribing perceptual mechanisms with the function of detecting properties of the environment, teleological theories attempt to characterise them as having the purpose of instantiating an accurate representation of these properties. And in the case that those mechanisms end up detecting a different property than the one corresponding to their function, teleological theories describe them as a case of malfunctioning and misrepresentation.

Teleological theories typically make use of a biological approach to functions, which analyses them in terms of their aetiology, viz. by identifying the function of a system with reference to the reasons why the system has the function it does (Wright, 1973). Since the mainstream view among philosophers of biology is that the best explanation of why systems have functions is natural selection, teleological theories normally define functions by appealing to how they evolved by natural selection (Allen, 2009). For example, they consider that the function of the heart is to pump blood because this function played some role in enhancing the survival and reproduction of the organism, and hence the species.

This version of teleological approaches to mental representation is what I shall call *teleo-biological theories*. In addition to giving a naturalistic and mind-independent account of function, teleo-biological theories also can distinguish the system's functions from other accidental effects they may eventually have. For example, the heart makes a noise that might be useful to know the mood of a person. However, (arguably) to make that noise is not a function of the heart insofar as this played no role enhancing the survival of its possessor during its phylogeny. In the same way teleo-biological theories

can be used to distinguish between genuine and accidental functions of perceptual systems. So it is often claimed that perceptual systems evolved the function of detecting certain specific environmental properties, even if sometimes they accidentally represent properties they were not designed to represent. This last case would correspond to cases of misrepresentation.

5.4 Against Teleo-biological Theories

In *Origins of Objectivity* Burge develops a teleological approach to perceptual functions, which he calls representational functions. However, he explicitly puts forward his approach in opposition to teleo-biological theories. One reason he gives to support this claim is that teleo-biological theories are "deflationary" (as he did with informational approaches, see 5.2) in the sense that they lead to the attribution of mental representations to simple organisms that do not even have perceptual systems. For example, proponents of teleo-biological theories such as Millikan (1989) and Dretske (1986) claim that some bacteria can have representations by virtue of having the biological function to detect and respond accordingly to certain environmental conditions. But as Burge argues, nothing in the way bacteria process information suggests that they go beyond mere sensory registration of information, or that they reach some level of constant detection of distal environmental properties. Moreover, an explanation based on purely biological and informational notions (i.e. physical- and computational- level explanations according to my terminology) can offer a comprehensive explanation of the behaviour of the bacteria, while no descriptive or explanatory advantage is gained by the use of representational notions.

Although Burge's critique seems essentially right, it should be noted that proponents of teleo-biological theories often take sensory states such as those of the bacteria as having just rudimentary forms of representational content, not comparable to those of belief (see Dretske, 1986; Papineau, 1987). But given how Burge understands the term "representation"—viz. as it is applied in psychological theories—to use it to explain phenomena that just involve sensory information is certainly misleading. In cases like this, a straight biological explanation would suffice, and the notion of representation does not seem to contribute anything relevant. For Burge, psychological

and biological explanations constitute different explanatory levels and he views teleo-biological theories as futile attempts to reduce the former to the latter. He claims that psychological explanations involve distinctive psychological notions such as representation and veridicality conditions, that cannot be reduced to biological terminology without losing important explanatory virtues.

A second attack Burge makes against teleo-biological theories is directed against their supposed pairing between representational accuracy and evolutionary success:

> The key deflationist [teleo-biological] idea in explaining error is to associate veridicality and error with success and failure, respectively, in fulfilling biological function ... Explanations that appeal to biological function are explanations of the practical (fitness) value of a trait or system. But accuracy is not *in itself* a practical value. Explanations that appeal to accuracy and inaccuracy—such as those in perceptual psychology—are not explanations of practical value, or of contributions to some practical end. (p. 301)

Burge identifies perceptual accuracy with veridical representation and argues that pairing veridicality with biological success is problematic. A common case used to illustrate this problem concerns predator-detection systems. For instance, several species of birds have evolved systems that respond to aerial predators by eliciting a fleeing response (Marler & Hamilton, 1966). Since the main predators during their evolution were hawks, a teleo-biological explanation would say that this system has the representational function of detecting hawks. But note that under certain ecological conditions it would have been perfectly possible for these birds to evolve predator-detectors that were highly inaccurate. Imagine, for example, that the energy consumed by the fleeing response is very low, whilst the real occurrence of a predator almost always results in being caught. Then even if the predator-detector is highly inaccurate and triggers many false alarms (e.g. by responding to any winged-silhouette), it could still have been recruited by evolution to perform a hawk-detection function. Burge offers a variant of this example by pointing out that fleeing responses to false alarms could also have improved fitness by means of increasing strength and agility, and in this way favoured the selection of hawk-detectors even if they misrepresent most of the time (p. 302).

Defenders of teleo-biological views have responded to cases like this by accepting that inaccurate perceptual systems could have evolved by natural selection (Godfrey-Smith, 1992; Millikan, 1989). According to teleo-biological theories, all that

matters is that hawks were the relevant environmental condition that explains the selection of the predator-detector during the evolutionary history of the birds. Even if the system was highly inaccurate and gave rise to many false alarms, the reason it was selected is that the few times it was successful in detecting hawks it had a significant effect in enhancing the survival of the species. Therefore the system has the biological function of detecting hawks, and in cases when it responds to any other winged-silhouette it is just misrepresenting hawks.

But Burge replies that views like this are counterintuitive and at odds with perceptual psychology, for nothing in the bird's perceptual computational machinery appears to have the capacity to discriminate between hawks and other aerial objects with winged-silhouettes. When an aerial predator approaches all the perceptual system can probably do is to infer from sensory information the perceptual invariant of a winged-silhouette. Then the system would be successful when detecting winged-silhouettes, even if what explains its evolutionary origin was the detection of hawks. Burge takes cases like this to support this claim that "the function fulfilled by representational success, by perceptual veridicality, is not a biological function" (p. 308).

However, I believe his move is too fast, since he misses one possible reply from teleo-biological theories. For it could be the case that the bird's predator-detection system has the biological function to detect hawks *in virtue of* detecting winged-silhouettes. As Neander (1995) suggests, both functions need not be mutually exclusive if we take them to be complementary functions at different levels of description. The function to detect winged-silhouettes can be regarded as the underlying mechanism that enables the bird to carry out its biological function of detecting predators. Even though at the level of early vision this mechanism cannot detect hawks, the fact that it is a crucial part of a larger system that evolved with that function of detecting predators suffices to ascribe it a biological function of detecting hawks (provided that it was the main predator during the evolutionary history of the bird).

This view seems compatible with both Burge's account of representational functions and teleo-biological theories of perception, and so it is perhaps surprising that he gives no attention to it in his book. I shall return to Burge's critique to teleo-

biological theories in the final section of this chapter. For now, let us focus on Burge's positive account of perceptual functions.

5.5 Burge's own Proposal: Drawing the Line on Perceptual Functions

According to Burge, an essential aspect of representations is that they bear reference relations to a subject matter, such as objects or properties of the environment. These reference relations are "established by a person or animal ... by the way of some thought, cognition, perception, or other psychological state or event" (p. 31). Paradigmatic forms of representation are propositional thought and concepts, however Burge believes that the "the most primitive form of representation is perception" (p. 9), in particular the detection of distal environmental properties carried out by percepts.

As it is widely acknowledged, Burge recognises that any account of representation must explain how they could fail to refer to what they are supposed to be about. To use a common terminology, they must explain how misrepresentation is possible (cf. Warfield & Stich, 1994). Thus a central issue for Burge's representational account of perception is to explain how percepts could have what he calls *veridicality conditions*, viz. the perceptual analogs to truth-conditions of belief. This normative character of percepts that has been troublesome for informational approaches (see chapters 3 & 4) and is something teleological approaches have attempted to figure out, as mentioned above.

Burge puts forward a particular teleological approach to perceptual functions, which he characterises as "representational functions" to emphasise the alleged representational nature of perception. As I explained in the previous section, he also departs from standard teleo-biological theories. His main motivation is that he believes standards of veridicality do not need to mesh with any practical value and therefore that representational functions are essentially independent from biological success. In Burge's words:

> Biological functions and biological norms are not the only sorts of function and norm that are relevant to explaining the capacities and behaviour of some animals. Given that veridicality and non-veridicality cannot be reduced to success and failure (respectively) in

> fulfilling biological function, we must recognise a type of function that is not biological function, a representational function. (p. 339)

As Burge acknowledges, *biological functions*[28] "are functions that have ultimately to do with contributing to fitness for evolutionary success" (p. 301) and "their existence is explained by their contribution to the individuals' survival for mating, or perhaps in some cases the species' survival" (p. 326). This corresponds to a standard teleo-biological approach that analyses functions in terms of their aetiology, often by reference to the process of natural selection[29]. In contrast, Burge's notion of representational function is consistent with a non-etiological, and often called *dispositional*[30], construal of functions, viz. one that does not define their nature it terms of aetiology but in terms of their current roles in carrying out some capacities of the organism. More precisely, Burge's representational functions have their metaphysical grounds on scientific realism, viz. the idea that we can adopt a positive epistemic attitude towards the theoretical components of our best scientific explanations. As mentioned in 5.1, this allows Burge to take a realistic stance towards representational functions given the assumption that the most successful explanations in perceptual psychology constitutively make use of them:

> The conclusion that perception has a representational function... derives from reflecting on the nature of explanatory kinds in perceptual psychology ... There is extensive empirical support for explanations in which the representational aspects of perceptual states are explanatorily central ... Such explanations evince the existence of perceptual states. So they support the claim that there are representational states that have representational functions. (p. 310)

It is important not to read Burge as arguing that representational functions did not evolve by natural selection. On his account he can just remain neutral about aetiology and instead focus on what functions are settled by our best current explanations of how perception works. It is also interesting to note that a similar analysis of functions is commonly adopted by computational approaches to psychology. These characterise psychological capacities such as perception, memory or decision-

---

[28] Biological functions are not always characterised in teleo-biological terms, but for expository purposes I shall follow Burge in doing so.

[29] Natural selection need not be the only source of aetiology. Some teleo-biological approaches also claim that functions can result from learning or conditioning. See e.g. Papineau (1987).

[30] For a good exposition of both opposing theories of functions in the context of psychological explanation see Price (2001). Thereafter, I stick to the term "dispositional" to refer to non-etiological functions.

making by looking at how they are actually structured in terms of their input-output relations, regardless of their historical origins (e.g. Cummins, 1983; Crane, 1995; for discussion on dispositional theories of function see Koons, 1998, and Fodor, 2000).

Accordingly, Burge believes that several cognitive capacities have non-biological functions. In addition to perception, he alludes to functions for belief-formation, deductive reasoning and primitive agency. One peculiar aspect of Burge's proposal is that biological and representational functions actually coexist in the same organism, in a way that gives rise to a complex array of different functions and normative constraints. I find this functional picture puzzling, but I shall reserve my arguments for the next section and conclude this exposition by trying to explain how Burge suggests these functions could be arranged in a whole organism.

A basic idea behind most teleological approaches is that functions are identified in the context of a functional analyses of the organism, where it is decomposed into systems (e.g. circulatory system) which are themselves decomposed into their parts (e.g. heart, arteries, veins, etc.). All these subsystems are at least partly explained in terms of their causal contribution to the functioning of the whole organism (sometimes called by Burge "individual functions", as mentioned in 5.1). In the author's words:

> Whole animal function is exemplified by the basic biological activities—eating, navigating, mating, parenting, and so on. These activities are functional in the most commonly cited sense of biological function ... They are distinctive in being functions of the whole individual—not the individual subsystems, organs, or other parts. (p. 326)

In his book under discussion Burge describes biological functions as coordinated sub-systems organised to maximise fitness. But on the other hand, representational and other non-biological functions are also compositionally described. For example, Burge points out that perceptual systems deliver accurate representations to belief-formation systems which have the function of generating true propositional representations, which then interplay with systems of deductive inference, and so forth.

But how could both biological and non-biological functions be integrated? At this point it is pertinent to introduce Burge's notion of agency. He characterises agency as the capacity to generate "functioning, coordinated behaviour by the whole organism, issuing from the individual's central behavioural capacities, not purely from

subsystems" (p. 331). The author sees agency as a property of organisms whose actions issue from central capacities that coordinate its subsystems, towards the fulfilment of whole-animal functions. He sets this sort of centrally-driven actions in contrast with typically peripheral movements such as reflexes, or certain processes carried out within a cell, which are not imputable to an individual as a whole.

Interestingly, Burge finds agency in very primitive organisms, even some which lack a central nervous systems. For example, he claims the paramecia's eating and swimming behaviour count as agency. The key point is that there is a coordination of different anatomical structures stemming from within the organism, that result in activities that at least contribute to the satisfaction of basic biological functions. Given that organisms such as paramecia and amoebas have agency but lack perceptual functions, a consequence of Burge's account is that agency is more primitive than perception. More precisely, he considers agency as a precondition for the emergence of perception and representation. When animals evolved perceptual systems, some of their actions started to be guided by representational states, actions attributable to the whole individual given that agency was already in place. Burge calls this *psychological agency*, and claims that it marks the point in evolution when the first properly psychological act was performed.

We can now return to Burge's account of representational function. He claims that agency is what makes possible the integration of biological and representational functions insofar as they operate in coordination towards the fulfilment of functions of the whole individual. To put it roughly, once agency is present, it is the individual who perceives and not just its subsystems. The notion of agency also helps Burge to explain why perceptual systems are not just peripheral, automatic computational subsystems such as reflexes, that do not feature in representational explanations. This is because cognitive psychology considers perceptual processes within explanations of behaviour imputable to the whole-organism and not merely to its computational subsystems. Hence Burge believes that what makes percepts genuinely representational is the conjunction of having the computational machinery for generating percepts and the possession of whole-individual agency, viz. having perceptual systems integrated with central cognitive capacities that result in behaviour.

So far I hope to have given a comprehensive presentation of Burge's account. Even though in his critique of informational and teleo-biological approaches the author makes some insightful points and appears to be compatible with the basic framework of this thesis, I believe his final account of perceptual functions is not convincing. In the following sections I shall discuss two problems concerning his account.

5.5.1 Objection 1: Burge's Mixed Account of Functions is Problematic

In general terms, the idea that perception and cognition have a functional organisation is widely accepted. Disagreements often hinge on whether cognitive functions should be characterised in etiological or dispositional terms, and on other issues that arise from this. This seems natural insofar as both teleological theories constitute different epistemological and metaphysical approaches towards the ascription of functions.

However, I believe problems begin when Burge combines representational and biological functions, since each comes from different teleological approaches that need not always agree about how to characterise the same function. For example, suppose that our best physiological theories explain how the heart works by ascribing it the function of pumping blood. Then from a dispositional approach it would be a fact that the heart has precisely that function. But imagine that research on the evolutionary origins of the heart finds out that the heart was not selected because it pumped blood, or that it simply did not evolve by natural selection (e.g. as a result of genetic drift). Then from a teleo-biological viewpoint the heart would have a biological function that is different from its current one, or worse, no function at all. As proponents of teleo-biological approaches have pointed out, biological and dispositional attributions of functions can be divergent and typically pursue different explanatory aims, and therefore it is recommendable to keep them separate (e.g. Godfrey-Smith, 1993).

Alternatively, some authors have proposed a pluralist view where both teleological approaches coexist. For example, Preston (1998) claims that etiological and dispositional functions can be complementary, and are required to cover the full range of functions—from artefacts to natural entities—as well as to account for how functions

change over time. This is not the place for a full discussion of pluralist theories of this kind, though. My present purpose is to argue that even though Burge's view could be interpreted along these lines, this brings up certain problems with his characterisation of perceptual functions. More precisely, combining etiological and dispositional functions can generate conflicting ascriptions of representational contents. I elaborate this idea below.

A well known problem associated with teleological theories of representation is how to avoid the indeterminacy of content (cf. Fodor, 1990). Recall the example of the predator-detection system of some birds and imagine that it responds to flying boomerangs in exactly the same way as with hawks. An information-processing explanation of how its perceptual system manages to detect environmental properties might fit equally well with the bird's perceptual states having as representational content 'winged-silhouette', 'boomerang', and perhaps some other similar objects. But this leads to the problem of determining which of those things the birds actually represent. At this point teleo-biological approaches are often called to disambiguate; they can argue that the function of the system is to represent winged-silhouettes, because winged-silhouettes and not boomerangs (or other objects) were selectively responsible for the evolution of that system (cf. Sterelny, 1990).

But, of course, this is not precisely Burge's strategy given that he rejects teleo-biological theories of perception. However, he adopts a mixed approach where dispositional functions of perception are somewhat "constrained" or "framed to fit" with biological functions of the organism. In Burge's words:

> the framework for perceptual reference and perceptual representational content is set by organism's responses to the environment in fulfilling individual biological functions, in the evolutionary prehistory of the perceptual system. (p. 321)

To see why this mixed approach to the individuation of content is problematic, consider the following scenario. In areas populated by birds, throwing and catching boomerangs becomes an extremely popular game, however associated with an unhappy consequence: an important number of birds die because of boomerangs falling over them. Eventually, some species of birds manage to scape safely from boomerangs thanks to their possession of predator-detector systems such as the one described above. The system is recruited, so to speak, to respond to boomerangs and elicit a flight

response[31]. Now, the predator-detector system would be performing the function of detecting boomerangs, at least from a dispositional viewpoint. We can even imagine that in this particular environment it becomes normal for the birds to avoid boomerangs, and therefore our customary explanations of their behaviour would have to incorporate their capacity/disposition to respond to boomerangs. Then it turns out that even though the system did not evolve for that reason, it happens to be perfectly fit for detecting boomerangs, and from a dispositional viewpoint the perceptual system of the bird would have the function of detecting boomerangs (to make the case more dramatic, we let us imagine that even though the system evolved as an adaptation for detecting hawks, those animals became extinct in the area and the only actual "predators" are boomerangs).

But recall that Burge's mixed view of functions also contains an etiological factor, where "antecedent interactions between moving bodies and operations of perceptual mechanisms are central to the explanation of the kinds (primarily the representational content) of perceptual states" (p. 71). These antecedent interactions bring us back to evolutionary history and would lead us to the conclusion that what the birds represent are hawks. Then we end up having at least three candidates for what is the representational content of the perceptual system: 'hawks', 'boomerangs' and 'winged-silhouettes'. I do not see how Burge's account would help to solve this indeterminacy. Depending on whether we give more relevance to a dispositional or an etiological approach, the contents we attribute will oscillate between this space of alternatives. And given that contents are supposed to be psychological kinds relevant for explaining behaviour, it is hard to see how such a degree of indeterminacy could be tolerated.

5.5.2 Objection 2: Passage from Accuracy to Veridicality is not Clear

Given that Burge rejects purely teleo-biological accounts of perceptual functions and that, as I argued above, his mixed account of functions is problematic, perhaps his proposal could be improved if framed in straight non-etiological, dispositional terms.

---

[31] For the sake of the argument, let us suppose that this is just an ontogenic recruitment, and no adaptive modifications to the system have occurred yet.

That is, by assuming that all functions are determined by their current roles at work in a system, as revealed by our best scientific explanations. In this section I argue that even if framed in this way, Burge's reasons for drawing the line at the level of perception are not compelling because they rest on unjustified assumptions, in particular, on the premise that perceptual veridicality is entailed by the detector-accuracy of its computational mechanisms.

Let us start by recapitulating Burge's account. The author submits to the basic tenets of computational and informational approaches to the mind, and draws the line for the origins of mental representations at the level of percepts. In order to justify his account, and in particular to explain the normative character of percepts, he adopts a teleological approach. Perceptual systems are supposed to have representational functions, which fulfil the role of yielding percepts with veridicality conditions. As the following quote shows, Burge parallels veridicality with accuracy in perceiving something:

> A *veridical* perception is a *correct* or *accurate* perception. A *veridical* thought is a *true* thought. Truth and accuracy (correctness) are subclasses of veridicality. (p. 39)

And also takes veridicality to be an outcome of the representational function of perception:

> Perceiving is a type of veridical representation. The representational function of a perceptual system is to represent veridically. Veridical perception is necessarily and constitutively a kind of success for a perceptual state or perceptual system. It is fulfilment of a kind of function. (p. 309)

The idea that a perceptual system, that has the function of being accurate in detecting a certain environmental property, can generate percepts with veridicality conditions appears to be uncontroversial. But to see what is misleading about Burge's account let us step back to the computational domain (without assuming the psychological domain) and ask what would make a computational system, that has the function of being accurate in detecting certain environmental property, capable of generating percepts with veridicality conditions. Note that since in this case we are not assuming that the computational system is perceptual (i.e. psychological), the capacity of being accurate does not entail the capacity of being veridical. What is missing then?

Burge's response would probably be that not any detector-system actually makes for veridicality, since this is only the outcome of systems endowed with genuine representational functions. But I believe this response sounds poorly revealing about the nature of representational states, since it just leads to replacing one question (what makes certain computational states representational?) with another (what makes certain computational functions representational?) without answering either of them. We pass from one puzzling notion (mental representation) to another (representational function). Thereby, we might ask Burge where representational functions come from.

A response to this question has already been advanced in 5.5. A computational detector-system fulfils representational functions if its behaviour is best described by appeal to psychological-level explanations, which means that its computational symbols have to be described as percepts (on pain of trivialising psychological explanations, see p. 342). As Burge notes, "such explanations evince the existence of perceptual states" (p. 310). According to him, the leading exemplar of psychological theorising in the context of perception occurs in visual psychology, and the crucial visual process Burge regards as characteristically psychological is one we are already familiar with: the generation of perceptual constancies (see 3.2 and 4.2). What this perceptual process does is to overcome the problem of underdetermination of the retinal input, viz. to infer information about distal environmental properties from proximal sensory information that is mathematically insufficient to determine it. Burge also acknowledges that the inferential processes that mediate perceptual constancies "are computational ... [and] describe quasi-algorithmic, quasi-automatic transitions in the perceptual system in ways that enable one to model perceptual systems on a computer." (p. 356)

Burge goes on to offer several examples of visual constancy capacities (pp. 342-366). For instance, one is convergence, which yields constant perception of distance, and other is lightness constancy, that delivers constant perception of surface lightness. He argues that these capacities are present in a wide range of animals, from arthropods to mammals, and concludes that his examples illustrate

> the role of [constancy] formation principles in explaining formations of perceptions. Each exemplifies the explanatorily non-trivial invocation of states with representational content (and veridicality conditions) that distinguishes psychology from biology. (p. 347).

So far, Burge proposes that a characteristically psychological aspect of representational functions is their computational capacity of detecting environmental invariants. I believe this is unconvincing, since the same capacity can be ascribed to the detector-systems of non-mental machines. Recall the example of the coin acceptor of the vending machine presented in 3.3.1. Its capacity to sort out the property of *being one pound's worth* and deliver a signal that is selectively sensitive to it, appears to be preserving the constancy of the detection of that property from a variety of inputs (i.e. a variety of possible coins and values). To consider a more sophisticated kind of artefact, take a digital camera. As it is well known, the way cameras capture light is similar to how the human eye works, for example in terms of image focusing and light adjustment. In both cases, light coming from objects in the environment is reflected onto a surface that transforms patterns of light into electric signals. In addition, some modern digital cameras have face-detection technology, which basically consists in algorithms that scan the image and detect the shape of human faces[32]. This process appears to be comparable to what happens in the input-systems of the visual system and thus to be able to deliver a constant detection of faces that overcomes the underdetermination problem of extracting that property from the information registered by the lens.

I take it for granted that to ascribe mental representations to the vending machine and the digital camera is clearly implausible. Can they count as counterexamples to Burge's proposal then? It seems to me that they do since they are rather equivalent to the examples of visual perception offered by Burge. In both cases there is a flow of information from the environment that goes through computational coding that delivers an informational structure that singles out some particular distal environmental property. And importantly, in the case of artefacts the mechanisms that mediate detection of environmental properties are fallible, and thus allow the possibility of misrepresentation, by the same means used by Burge to account for misrepresentation in genuine perceptual systems—he claims that misrepresentations can be explained as "malfunctions of or interferences with the [computational] system", due to their fallibility under possible (adverse) conditions (p. 346).

---

[32] However immensely more complex, human perception also appears to be equipped with hard-wired algorithms for detecting the shape of faces.

At his point Burge might resort to the notion of agency, which as explained in 5.5 is supposed to constitute a prerequisite for the emergence of perceptual systems. However, I believe this notion is not of much help. For as Burge understands it, agency is a very elementary capacity of organisms, to some extent equivalent to the definition of autonomous agent I presented in 2.3. It precedes perception and psychological capacities in general, and as such, does not involve any characteristically psychological properties. I do not see why all autonomous agents capable of detecting environmental invariants would have to qualify as possessing mentality, any more than robots such as the Mars rovers described in chapter 2 would do. I believe, though, that the basic idea behind the notion of agency can be useful in this case, but it must be one that captures distinctively mental categories. In the final chapter of this thesis I develop a proposal along these lines.

So I conclude that, overall, Burge is wrong in assuming that there is a clean passage from accuracy to veridicality in the case of the computational systems that mediate the detection of distal environmental properties in some animals. In these cases accuracy in detection does not entail veridicality. There is certainly not an a priori entailment, neither an intuitive connection as can be gathered from the case of the artefacts showed above.

5.6 Conclusions

In this chapter I have examined Burge's account of perceptual functions and argued that it has many problems that weaken his case that mental representations are originated at the level of percepts. However, several aspects of his view can be useful for, and compatible with, the line of argument I am developing in the present thesis. One is his emphasis on scientific realism and the appeal to psychological-level explanations to distinguish mental symbols from (merely) computational symbols or other informational notions. Another is the notion of agency, in particular the idea that typically psychological explanations presume that representational states are guiding the behaviour of whole-agents, instead of their (subpersonal or computational) parts. In the final chapter of this thesis, I return to those ideas when putting forward my own proposal for a criterion for drawing the line between agents with and without mentality.

Chapter Six

Bermúdez and Carruthers on Drawing the Line

6.0 Introduction

In this chapter I address two more philosophers who have explored the minimum conditions for having mentality: José Luis Bermúdez and Peter Carruthers. They propose forms of symbolic processing and cognitive architecture that, they claim, deserve to be described in psychological terms, and put forward some criteria for attributing mentality to animals.

After critically presenting the views of Bermúdez and Carruthers, I conclude that they do not offer a satisfactory criterion for distinguishing computational from mental symbols. Bermúdez's proposal attempts to formulate a framework for psychological explanation that does without a standard—inferential—model of psychological explanation. However I argue that it has many problems.

Concerning Carruthers's view, he follows a more traditional version of psychological explanation and proposes a cognitive architecture that captures what he takes to be the core of mentality. I contend, however, that his account does not satisfactorily distinguishes mental from non-mental computational architectures. I conclude by introducing an alternative framework based on a personal-level approach, which aims to do better in capturing what is paradigmatic of psychological explanations, and which I shall explain and develop in the next—and final—chapter of this thesis.

6.1 Bermúdez on Thinking Without Words

In his book *Thinking Without Words*[33] José Luis Bermúdez explores how psychological explanations could be formulated to account for the behaviour of

[33] Unless otherwise noted, from this section until 6.3 all page references are to this book.

creatures who lack language, whether animals or human infants[34]. Even though his primary concern is to develop a framework of psychological explanation applicable to nonlinguistic creatures, he also deals with what are the minimum conditions under which psychological explanations apply. In this sense, his proposal is of particular interest for our purposes since it is only when a creature's behaviour can be explained in psychological terms that we are justified to regard it as cognitive, instead of nonpsychological or merely mechanistic. As Bermúdez (1995) has claimed elsewhere:

> Explanations of behavior, particularly when dealing with the cognitive abilities of non-linguistic creatures, quite rightly operate with a principle of parsimony. Appeals to representational states should be made only where it is theoretically unavoidable, where there is no simpler mechanistic explanation of the behavior. (p. 346)

In general, Bermúdez's proposal is compatible with the naturalistic and computational background put forward in the first chapters of this thesis, in particular with respect to the symbolic, information-bearing, nature of thoughts. The author believes that creatures can be perceptually sensitive to their environments by means of picking up and coding information in a way that singles out particular distal properties of their environments. A crucial aspect of this process is that it manages to represent environmental invariants, so that "the most primitive form of categorisation is grounded in perceived similarity" (p. 94). Therefore, Bermúdez submits to the standard idea that basic forms of perceptual categories are delivered through the coding of perceptual constancies (see 3.2). They determine which properties of the environment the creature is sensitive to, and also shape under which *mode of presentation* those properties will be represented by each species:

> [D]ifferent types of nonlinguistic creature will carve their environment up in different ways as a function of being perceptually sensitive to different object-properties. The essence of perception under a particular mode of presentation comes because different nonlinguistic creatures will perceive different types of similarity between these objects*. (p. 95)

The expression "object*" is intended to denote environmental objects as regarded from the particular mode of presentation under which creatures apprehend them, instead of according to our own (human) perceptual and linguistic categories. The author believes that in this way it is possible to determine the "ontology" each creature has, viz. the contents by which their symbolic structures carve up their environments.

---

[34] Henceforth, just "nonlinguistic creatures".

There is, however, one important aspect in which Bermúdez's approach departs from the background presented in the first chapters of this thesis. It has to do with the nature and scope of psychological-level explanations. On the one hand, the author accepts that there is a traditional view of psychological explanation which he calls *standard belief-desire explanation* (compatible with the one presented in 1.4.1). This account, endorsed by authors as diverse as Davidson and Fodor, characteristically involves the postulation of beliefs and desires, and reasoning processes carried out over them. But even though Bermúdez agrees in that this account is suitable for describing human behaviour, he contends that it is inappropriate for explaining the behaviour of nonlinguistic creatures. Bermúdez then works out alternative types of psychological explanation that according to him are required if we want to ascribe psychological states beyond the domain of human beings.

His main motivation for departing from the standard belief-desire model is that he regards nonlinguistic creatures as incapable of engaging in genuinely inferential symbolic processing. I discuss his arguments for this in section 6.3. For the moment, let us focus on his overall view of psychological explanation, to then explore his alternative account of psychological explanation that does without the standard belief-desire model. To start with, consider two characteristics the author presents as essential to psychological explanations (p. 10):

(1) They serve to explain behaviour in situations where the connections between sensory input and behavioural output cannot be plotted in a law-like manner.

(2) They rely on the cognitive integration of different psychological states.

Each of these characteristics[35] serves to rule out from the scope of psychology certain alternative forms of behavioural explanation. For example, fixed action patterns and most types of associative conditioning fail to meet (1) since their explanations basically reduce to some sort of input-output link, while in contrast, psychological explanations typically appeal to inner states that function as intermediaries between

[35] To simplify the exposition I have omitted a third characteristic of psychological explanations that Bermúdez presents in his book, which is that they appeal to psychological states that admit of misrepresentation. The omission is innocuous for present purposes since the author endorses a standard approach to deal with misrepresentation along information-processing lines (see chapters 3 and 4), and so the capacity to misrepresent is assumed to be present in creatures endowed with perceptual systems. Indeed, the author gives little attention to this issue in his book.

sensory input and behavioural output. A related point is that in fixed action patterns or associative conditioning no significant interactions between inner states such as beliefs and desires are supposed to take place, and so (2) is also not satisfied. It is noteworthy that this is consistent with the idea, put forward in 1.4.3, that both fixed action patterns and associative conditioning correspond to a nonpsychological, but physical, level of explanation.

Bermúdez then considers another (alleged) alternative to the standard belief-desire model of behavioural explanation, which he calls *minimalistic*. Since its most paradigmatic version is Gibson's theory of *affordances*, and to simplify the exposition, I shall focus on it (Gibson, 1979). This theory explains behaviour in terms of affordances, which are basically perceptual states in which the environment somehow "offers" the creature potential actions to carry out. A characteristic of affordances is that, according to Gibson, they are directly picked up by perceptual systems in the sense that the contents of perception themselves present the creature possible courses of action. He then explains behaviour in terms of the perception of affordances that directly manifest possible courses of action the creature might follow in accordance to its desires and needs. Even though in Gibson's theory there is a sort of cognitive integration between perceptual states and desires—and so (2) could be met, Bermúdez considers it insufficient to qualify as psychological explanation because it fails to satisfy (1). The reason why is that in Gibson's theory behaviour is directly attached to perceptions of the immediate environment, and so the creature cannot go beyond the aforementioned input-output link. As Bermúdez remarks, an "action requires psychological explanation just if its occurrence could not have been predicted solely from knowledge of the environmental parameters and sensory input" (p. 129); and since minimalistic explanations always operate over immediate perceptual states, he concludes that they cannot provide a framework for psychological explanations. But what alternatives are left? As Bermúdez notes, the obvious one is to appeal to inferential capacities:

> The natural way to understand what I am calling nonimmediate perception is in inferential terms—and this is certainly how it has been understood by many philosophers who have considered the matter. (p. 52)

According to Bermúdez the resort to inferential capacities amounts to return to the standard belief-desire model which, as noted, he regards as not applicable to

nonlinguistic creatures. Therefore the goal becomes to find further alternative types of psychological explanation that do without inferences but at the same time are capable of fulfilling the requirements for psychological explanation presented above. In the following section I present Bermúdez's proposal on this matter.

6.2 Bermúdez's Line for Mentality: Proto-inferences

As noted above, Bermúdez claims that two essential characteristics of psychological explanations are that they are not restricted to the "here and now" of what is immediately perceptible and their reliance on cognitive integration. He then attempts to develop a type of psychological explanation that, without resorting to inferential capacities, is able to satisfy these characteristics. His strategy has two steps. The first is to account for symbolic structures that can be viewed as beliefs and desires and applied to nonlinguistic creatures, and also be capable of denoting states of affairs that are not immediately perceptible. The second is to explain how those structures could be integrated and engage in decision-making processes comparable to the inferential operations performed by linguistic creatures. I explain those two steps below.

Let us start with Bermúdez's account of symbolic structures suitable for explaining the behaviour of non-linguistic creatures. For this purposes, he develops a version of *success semantics*, where the content of a belief is defined as its utility condition and the content of a desire as its satisfaction condition. As the author acknowledges, this can be understood as a form of functionalism, where mental states are individuated in terms of their functional relations to one another and with behaviour. In the following passage he explains this idea in more detail:

> Beliefs, according to success semantics, are causal functions from desires to actions. The content of a belief is its utility condition, where a utility condition is the state of affairs whose holding would result in the satisfaction of desires with which that belief is associated. True beliefs are such as to cause actions that satisfy desires ... A particular desire has the content that it does in virtue of its satisfaction-condition, where the satisfaction-condition of a desire is the state of affairs whose holding leads to the cessation of the behavior to which the desire gives rise. (p. 105)

The fact that Bermúdez's approach can be understood in functionalist terms might suggest that it is compatible with CTM (see 1.2), but this is not so. Even though

he individuates mental states by their functional roles and explains their content by information-processing means, his view is not properly computational because it does not structure mental states at the level of their vehicles (pp. 111-116). That is, they lack a logical structure from which syntactic operations could be performed over them. As I shall explain in section 6.3 this is the reason why, according to Bermúdez, animals cannot carry out inferential thought.

Now let us see how Bermúdez's version of success semantics could satisfy the two characteristics of psychological explanations presented in the previous section. First, mental states understood in these terms can project from what is directly perceivable and so go beyond the "here and now". This is possible because belief contents are defined by appeal to their disposition to satisfy the desire(s) with which the belief is conjoined, and so there is no need of either spatial or temporal contiguity between the belief and the state of affairs defined by its utility condition. Indeed something similar occurs with desires, which satisfaction-condition can be a state of affairs that is not immediately perceivable. Secondly, in this view there is cognitive integration insofar as behaviour stems from the combination between (at least) an instrumental belief and the desire to be satisfied by means of the course of action specified by the belief.

Someone might wonder, however, how that integration between mental states could be possible without them having vehicles and inferential structure. As Bermúdez notes, psychological explanations operate by integrating mental states in processes such as decision-making that rationalise the behaviour being explained. But if there are no inferences going on, how could mental states combine and interact in a way complex enough to account for rational processes? This leads us to the second step in Bermúdez's strategy mentioned in the opening paragraph of this section, which is to account for how his proposed symbolic structures could be integrated and engage in decision-making processes comparable to those performed by linguistic creatures. For this purposes he develops a framework for symbolic processing that can satisfy minimal rational constraints without appealing to inferences. It is grounded on what he calls *proto-inferences*. For ease of presentation, I shall follow Bermúdez and focus on the case of decision-making. In this respect, the author observes:

> What then is involved in genuine decision-making? The minimal requirement is that the selection of a particular course of action should be made on consequence-sensitive grounds ... Deciding is not simply selecting. It is selecting for a reason. (p. 124)

Then a creature engaged in a process of decision-making has to be capable of deciding between two (or more) courses of action by assessing their consequences. This involves the explicit representation of alternative contingencies, an evaluation of the possible outcome-situations that might result from the actions that could be performed. In terms of Bermúdez's version of success semantics, those representations correspond to instrumental beliefs which utility-condition corresponds to the alternative outcome-situations that would satisfy a desire.

Let us consider one of Bermúdez's examples by way of explaining his proposal. A thirsty animal approaches a watering hole, in an environment where there are gazelles and lions. In order to drink water safely, the animal has to discriminate whether lions (viz. predators) are present or not. According to Bermúdez framework, it is plausible to assume that the animal can have beliefs such as that 'the gazelle is not at the watering hole' and 'the lion is not at the watering hole'. Let A stand for the former and B for the latter sentence. The author argues that after perceiving that 'the gazelle *is* at the watering hole' (not-A) it will be possible for the animal to conclude that 'the lion is not at the watering hole' (B) by reasoning from an excluded alternative in terms of the disjunctive syllogism "A or B, not-A, therefore B". Note that the basic logical connectives involved in this reasoning are negation and the material conditional. Bermúdez claims, however, that these logical connectives are out of the reach of nonlinguistic creatures, since to master them it is required to apply them to complete propositions, something that crucially depends on the capacity to carry out second-order reflection (which in turn depends on language, see 6.3). To sort this problem, Bermúdez formulates nonlinguistic, "protological", analogues of them, which are summarised below:

- *Protonegation*: In contrast with the negation operator, which applies to whole propositions, protonegation consists just of a proposition with a negative predicate. So for instance the protonegation of 'the lion is *not* at the watering hole' would be 'the lion is at the watering hole'. According to Bermúdez it is plausible to grant this primitive form of negation to nonlinguistic creatures since it just presupposes the

ability grasp pairs of symbols that are contraries, e.g. of presence and absence, or of safely and danger, which clearly appears to be more fundamental than the mastery of the negation operator.

- *Protocausation*: This is supposed to be the precursor of conditional reasoning and a primitive form of causal reasoning available to nonlinguistic creatures. Bermúdez considers it to be widespread in the animal kingdom since the ability to detect causal regularities and to distinguish genuine causal relations from accidental regularities has an obvious survival value. Returning to the previous example, protocausation would be what makes it possible for the animal to track a causal relationship between the fact that a gazelle is at the watering hole and the fact that lions are not present.

These two protological capacities are supposed to ground primitive forms of reasoning Bermúdez takes to be analogous to certain fundamental inference forms. For example the disjunctive syllogism "A or B, not-A, therefore B" from the example above can be understood in terms of the modus ponens "if not-A then B, not-A, therefore B". Overall, the main goal of Bermúdez's proposal of protological operations is to demonstrate that the beliefs and desires of nonlinguistic animals can engage in primitive forms of practical reasoning, even though they lack syntactic structure and cannot take part on second-order reflection. In his words:

> protoinferences at the nonlinguistic level are not made in virtue of their form. Creatures who engage in, for example, proto-*modus tollens* need not have any grasp of the form of an inferential transition as truth-preserving. In fact, they cannot have any such grasp, since that would involve second-order reflection on the evidential relations between propositions ... (pp. 148-149)

After this brief presentation of Bermúdez's framework for nonlinguistic thought and reasoning, let us sum up the minimum conditions under which psychological explanations apply according to this view. It seems that the psychological domain is appropriate to describe the behaviour of creatures at least capable of:

- Generating symbolic structures that fulfil the functional roles characteristic of beliefs and desires, and can stand for states of affairs that are not immediately perceivable.

- Grasping the distinction between two pairs of contrary concepts (protonegation).

- Distinguishing causal regularities between events from accidental ones (protocausation).

- Integrating their symbolic structures to make them engage in processes of protoinference that result in adaptive behaviour.

In sum, Bermúdez's version of psychological explanation appeals to mechanisms that seem basic enough to be present in most nonlinguistic creatures, and thus to apply widely at least from insects upwards in the animal kingdom. Many animal behaviours formerly explained in terms of fixed action patterns or associative conditioning, would then be describable as the result of thought and reasoning. I believe, however, that Bermúdez's version of success semantics has several problems, and so is not a plausible alternative to standard psychological explanations. I present my objections below.

6.2.1 Objections

A general strategy adopted by Bermúdez in his book is trying to understand in operational terms how the ascription of psychological states and rationality to nonlinguistic creatures could be justified. As explained above, Bermúdez's version of success semantics identifies the content of a belief with its utility condition, which is the state of affairs that would have to obtain for the various desires with which it is associated to be satisfied, and the content of a desire with its satisfaction condition, which is the state of affairs whose holding leads to the cessation of the behaviour to which the desire gives rise. Note that this characterisation of mental states ultimately rests on observable behaviour: the utility condition of a belief is understood in terms of its disposition to satisfy a desire, which is in turn defined in terms of its disposition to cease the behaviour it normally causes.

Despite the behaviourist flavour that emanates from operational definitions of behaviour, Bermúdez is certainly not a radical behaviourist because he describes beliefs and desires as inner states that cause behaviour. However, his operational approach to mental states still resembles some forms of behaviourism, since he ends up analysing beliefs and desires in terms of observable behaviour, just as analytical behaviourists

attempted to do (Rey, 1997). As Fodor (2003) points out in his review of Bermúdez's book, this resemblance also makes this approach subject to the same worries raised against behaviourism. To illustrate why the use of operational criteria for defining a desire does not work, Fodor gives the following example:

> Getting food terminates your hunger. Whether it also stops your hunger behaviour depends on the circumstances; notably on what you have in mind. It won't stop your scrounging for food if you have in mind not just to do some eating, but also to do some hoarding. (p. 17)

The point of the quote is that behaviour often underdetermines the mental life of the cognitive agent. What is satisfied by getting food is the desire to eat, not the (eating) behaviour the desire gives rise to, since for instance the eating behaviour might carry on even if the desire has already been satisfied. As cognitivists approaches often insist, what is important for psychological explanation is not the behaviour, but the inner processes that may or may not produce the behaviour. It should be noted, though, that Fodor's example is somewhat misguided for two reasons. One is that if it is the desire for hoarding what motivates the behaviour, then this is not precisely a case of hunger behaviour, but of hoarding behaviour. The satisfaction condition of the desire would then simply be to store up sufficient peanuts—and not getting fed, as the examples wrongly assumes. So, there would be nothing misleading with the fact that scrounging for food continues after having eat enough.

Secondly, if what Fodor means with the desire for hoarding is a complex mental state —such as a second-order one that suppresses, so to speak, the desire to eat in order to get food stored—then Fodor is missing the point. For Bermúdez is aware of some of the limitations of his version of success semantics, and explicitly points out that it is not purported to apply to second-order desires—which are only available to linguistic creatures on his framework. Indeed, he claims that his "model only works for relatively simple desires ... [where] There is a very clear sense in which we can identify when the behavior associated with the desire for food actually ceases and hence works backward to its satisfaction-condition." (p. 68)

But having said this, I believe the general critique still works against Bermúdez's view, since its operational character makes it also incapable of accounting for some complex forms of animal behaviour. For example, consider the case of navigation by path integration present in insects such as wasps and honeybees. As

explained in 2.4, path integration involves the integration of information about the vectors travelled, and by this means wasps can remember the location of food sources and also find their way back to their nests. Imagine that a wasp has a belief that codes information about the vector flight that leads to a food source, and that its content is something like 'food is at vector flight V32'. From the viewpoint of Bermúdez's version of success semantics, that belief content would correspond to "the state of affairs whose holding will bring about the satisfaction of desires with which that belief is conjoined", e.g. the desire for food.

But this way of defining a belief is insufficient to account for the complexity of some of the wasp's beliefs. Note that the insect's mechanisms of path integration are rather flexible, and allow the wasp to recalculate its location with respect to its destination or visible landmarks, and eventually take a new route towards its destination, e.g. fly to its nest in a straight line even though its initial route was done in zigzag. For that to be possible, along with the belief that 'food is at vector flight V32' the wasp would have to possess more specific beliefs about sections of the vector travelled, which could be deployed to recalculate the vector flight. One of these more specific beliefs could be, for instance, the belief that 'from the hive to this point the vector flight is V21'. Let us see how this belief could be individuated according to Bermúdez's approach. We would have to start looking for the desire that is supposed to be satisfied by that belief. But since the belief could be used in different calculations, it might eventually lead to distinct satisfaction-conditions such as reaching food or returning to the nest. Then, that specific belief does not seem to be associated with any particular desire, such the desire for food or the desire for homing, but to be better understood as a belief that takes part in reasoning about navigation.

This example shows a case where some beliefs of a nonlinguistic animal do not seem to be directly associated with the satisfaction of any particular desire or satisfaction-condition as Bermúdez's view requires. Rather, they appear to be better understood in traditional cognitivist terms, viz. as inner processes with the potential to generate a variety of behaviours. It should be noted, though, that in his book Bermúdez makes efforts to give a more cognitivist character to his account. He admits that since success semantics is basically an extensional way of individuating mental states, it has

the problem of not saying much about the nature of mental content and its explanatory role in generating behaviour. Then Bermúdez acknowledges that his

> version of success semantics needs to be further supplemented by showing how it can capture the mode of presentation under which its utility and satisfaction conditions are apprehended. (p. 92) ... [The content-constituents of thoughts] must be capable of performing genuine explanatory work when embedded in different thoughts in different contexts. (p. 97)

With the notion of "content-constituent" the author attempts to capture the particular mode of presentation the content of beliefs or desires have, and as the quote suggests, he also sees as a requirement for his account that those content-constituents have cognitive significance, viz. that they translate into some difference in behaviour. In this way, Bermúdez works towards making the individuation of content more fine-grained than standard versions of success semantics, allowing, for example, beliefs with the same utility condition to have different content-constituents. To explain this idea let us return to the example of wasps. In addition to their capacity to navigate using path integration, wasps can also guide themselves by relying on landmark cues. In terms of success semantics, we might ascribe the wasp with the belief that food is in a certain location, a belief whose utility condition would be the state of affairs (i.e. getting food) whose holding would satisfy the desire to feed. Now, considering the two navigational systems of the wasp, we could add that it has two ways of coding the location of food: one in terms of the vector flight required to reach it, and other in terms of its spatial relation with landmarks. These would correspond to alternative content-constituents for the same belief described above, and interestingly for Bermúdez's purposes, can be shown to have cognitive significance. For even though they have the same utility condition, depending on which navigational system the wasp is using its frame of reference will change and this will have consequences in its behaviour. And this could be empirically tested, for example, by manipulating the relevant parameters such as the location of landmark cues, of after capturing a wasp and releasing it in a place where its previous vector flight is no longer useful.

But even though according to Bermúdez's refined version of success semantics mental contents are described as playing a significant role in determining behaviour, the explanatory framework the author presents to account for them is, I contend, still unsatisfactory. Note that if a belief, with certain content-constituent, is to interact with

desires in order to fulfil their satisfaction-conditions, there must be some mechanism that explains how behaviour comes up from that interaction. Bermúdez's proposal in this respect, is to appeal to Gibson's theory. Let me explain why it does not work.

As mentioned in the previous section, Bermúdez claims that genuine decision-making involves acting on consequence-sensitive grounds, that is, by being sensitive to the information available about the likely outcomes of each possible course of action. According to the standard belief-desire model, this decision-making process takes the form of an expected utility calculation where the payoffs of (some of) the different outcomes associated with each course of action are explicitly represented and compared. But of course, according to Bermúdez this sort of decision-making is not available to nonlinguistic creatures, and at this point is where he resorts to Gibson's theory. The general idea is that "the instrumentality of a particular course of action is manifest in the content of perception—and of course a single perception can reveal different potential courses of action" (p. 135). He adds that even though the contents of perception contain information about the likely outcomes, that "is not necessarily registered as information about likely outcomes. The animal just sees what to do by comparing [the relevant perceptual contents]" (pp. 137-138).

But how precisely is it that the information about the relevant outcomes is "manifest" in the contents of perception in a way that the animal can "just see what to do" by looking at them? Sometimes Bermúdez appeals to Gibsonian terminology as a way of making sense of these ideas:

> Gibson's theory is that perception is not neutral. It is not just a matter of seeing various objects that stand in spatial relations to each other. It involves seeing our own possibilities of action—seeing the possibilities we are "afforded" by the environment. If this is right then we can see how a given behavior might be selected from a range of alternatives in a way that does not involve a process of [standard] decision-making. (p. 121)

The problem is that it is not clear at all that the theory is right. Moreover, and as many authors from the cognitivist tradition have pointed out, the appeal to Gibson's theory does not help much to clarify the mechanisms that allegedly make possible to link the contents of perception with desires and action schemas, without any inferences, memories or calculations involved in that process (e.g. Fodor & Pylyshyn, 1981). For example, it is hard to imagine how a wasp released in a remote site could opt to fly

straight to its nest without relying on a stored representation of the vector of the route between nest and feeder, their relation to local landmarks, and on the capacity to carry out calculations over these information.

In sum, it seems that Bermúdez's attempt to develop a type of psychological explanation that does without structured content vehicles and inferences is unsatisfactory, and incapable to account for complex but however widespread forms of animal behaviour. He draws elements from the behaviourist and Gibsonian traditions in order to develop a mentalistic version of them, however their help is limited since his view inherits well-known problems associated with those traditions.

The upshot at this point is that Bermúdez's framework of psychological explanation for non-linguistic creatures does not provide an adequacy criterion for drawing the line between agents with and without mentality. Besides, it is interesting to note that when it comes to the standard-belief desire model of explanation—which roughly matches what I have called the psychological level—, Bermúdez contends that it can only be applied to language-using creatures. So if we assume the failure of his mentioned alternative framework of psychological explanation, Bermúdez's view could be interpreted as drawing the line for mentality at the level of language-using creatures. In the next section I shall argue that his arguments for this are also unsatisfactory, however.

6.3 Language and Second-order Thoughts

As mentioned above, Bermúdez assumes that the standard belief-desire model of psychological explanation cannot apply to the behaviour of nonlinguistic creatures. The reason for that is that he regards language as necessary for developing thoughts with syntactic structure and capable to engage in inferences. He emphatically writes:

> [W]e have no theory at all of formal inferential transitions between thoughts that do not have linguistic vehicles. Our models of formal inference are based squarely on transitions between language sentences (as codified in a suitable formal language). (p. 111)

The central idea is that without the provision of natural language it is not possible to develop inferential forms of thinking. This idea rests on two assumptions:

(1) inferential thinking requires the mastering of logical operators, such as tense operators and modal operators, which in turn demand *metarepresentational* capacities (i.e. thinking about thoughts), and (2) metarepresentational capacities are only possible when thoughts are linguistically vehicled. In the reminder of this section I argue that both assumptions are flawed, and therefore that Bermúdez's main claim that nonlinguistic creatures cannot perform inferences does not follow.

Let us start addressing (1). Sometimes in the course of thinking we apply a rule over some thoughts we previously had. This certainly involves the metarepresentational capacity to use one (higher-order) thought to target some other (first-order) thoughts. A straightforward example is the use of a rule of conditional inference, such as modus ponens, to manipulate some of our thoughts according to that rule. Bermúdez claims that to be able to carry out logical reasoning of that sort we need to master logical connectives, such the material conditional, which involves picking up, so to speak, two thoughts and manipulating them according to the rule of modus ponens so as that the second thought will be recognised as true if the first is recognised as true. But even though the process of picking up and manipulating symbol structures certainly involves the metarepresentational capacity of recognising and targeting them from a higher-order level of processing, it is not clear that this is required for a process to be recognisable as modus ponens (or any other rule of inference). For a first-order sequence of symbolic processes might have the structure of modus ponens and, as Carruthers (2004b) points out, there is no reason to believe that to run such a process requires engaging in metarepresentational thinking at the same time.

Moreover, as I explained in chapters 1 and 2, the computational theory of mind precisely provides a framework for understanding how inferential processes could be carried out by nonlinguistic creatures and even machines. All that is required for a system to be regarded as implementing an inference is for it to possess symbolic structures with representational properties and a syntactic component capable of mechanising algorithmic processes in an autonomous way. The debate about whether those inferences could be taken as genuine often focuses on whether the representational component of computers has intrinsic content or not, and on whether the computational level of explanation maps onto an autonomous natural domain. But the debate hardly relates with the capacity of computers to develop higher-order levels

of processing, such as having symbols that target other symbolic structures. Instead, that capacity appears to presuppose, but not to be necessary for, the ability to carry out inferential operations in general.

The second assumption of Bermúdez's argument mentioned above is that metarepresentational capacities are only possible when thoughts are linguistically vehicled, that is, that public language sentences are the only possible kind of vehicle for thoughts that can be the objects of higher-order thinking. The main argument Bermúdez puts forward to defend this claim is grounded on introspective evidence. He uses reflective thinking, viz. evaluating evidential and inferential relations between thoughts, as an example of "thinking that will paradigmatically involve a direct and conscious cognitive access to the target thoughts" (p. 159). He claims that every time we engage in conscious reflection over our propositional thoughts they have the form of sentences of public language, and that therefore language appears to be the only possible vehicle for metarepresentational thinking.

There are several problems with this argument. One is that reflective thinking does not appear to be the only form of metarepresentational thought, and so even if reflective thinking always involve public language, this does not deny that there could be other forms of thinking about thoughts that do not depend on language (Carruthers, 2004b). This could be the case, for example, in certain metacognitive forms of thinking such as hypothesis testing or belief revision that occur at a subpersonal, unconscious level of symbolic processing. Another worry with Bermúdez's argument relates with his use of evidence from introspection. As Fodor (2003) contends, introspective reports are, at best, partial accounts about what goes on inside our heads. It could be perfectly possible that some forms of reflection over thoughts which are not vehicled by sentences in public language take place well below the level of conscious awareness.

Finally, Bermúdez does not convincingly rule out alternatives to public language as vehicles for thought that could allow for metarepresentation. He discusses two alternatives: pictorial models of representation and the *language of thought hypothesis* (LOT). Briefly, pictorial models claim that mental states represent environmental properties by virtue of resembling or having a structural isomorphism with them. Following most authors within the cognitive tradition, Bermúdez points out that even

though pictorial models can have a structure (e.g. in the case of mental maps), they lack the expressive power to represent propositional contents, or the relations they might bear with one another. For instance, consider how we could possibly represent the difference between relations we express using the words "and", "or", "if", in terms of pictures (see Crane, 1995, and Cummins, 1989, for standard critiques to pictorial models). Bermúdez concludes that since pictorial models cannot provide the vehicles for expressing propositional contents, their transitions and relations, pictorial representations lack structured elements capable of becoming the objects of higher-order mental processes.

Now let us turn to LOT[36], the alternative to which Bermúdez gives more attention, and not surprisingly given that it proposes that thoughts can have the structure of sentences, however they do not correspond to public-language sentences. Firstly, Bermúdez takes issue with this view by employing the introspective argument sketched above, stating that conscious thinking characteristic of reflexive thought is based on public-language sentences instead of LOT. But as mentioned, this argument is feeble both because of the weaknesses of introspective arguments themselves and for the fact that most of thinking according to LOT is supposed to occur at a subpersonal level, inaccessible to introspection. And this last point is relevant to Bermúdez account given that his view of perception appears to assume that some sort of symbolic processing is carried out by perceptual systems situated at a subpersonal level (e.g. early vision, see 1.4.2). And as the following passage shows, he acknowledges that unconscious symbolic processing might happen elsewhere:

> It might well be the case that certain types of hypothesis testing and refinement do take place at the subpersonal level. Something like this happens, according to Fodor, when we learn a language. Nothing that I say is incompatible with that proposal, since my claim is simply that such processes would not count as instances of second-order dynamics [i.e. metarepresentations]. (p. 159)

But why cannot processes carried out in LOT be instances of metarepresentations? Recall that LOT is basically a computational theory of mind that is explicit on its claim that the syntactic structure of thought has an expressive power equivalent to that of any other language, and as I explained in the first chapter, this is supposed to be achievable by means of symbolic structures and computational

[36] Following Bermúdez, here I focus on Fodor's version of this view. See Fodor (1975, 1987).

transitions. Why cannot there be computational systems with hierarchical structures, that take first-order symbolic structures as objects for higher-order algorithms? Nothing seems to rule out that possibility, and Bermúdez does not provide reasons why this is to be ruled out as a theoretical alternative to his account.

So far, I have given a review of Bermúdez's proposal regarding whether and when psychological explanations can be applied to explain the behaviour of non-linguistic creatures. I conclude that the framework of psychological explanation he advances for non-linguistic creatures is too problematic to provide the basis for making a distinction between agents with and without mentality, and that even his criterion for applying the standard belief-desire model of psychological explanation is misleading. In what remains of this chapter, I shall address the view of another philosopher who explored the grounds for ascribing mentality to non-human agents.

6.4 Carruthers's Line for Mentality: Core Cognitive Architecture

In a series of writings, Peter Carruthers (2004a, 2006) has suggested that the line for mentality should be drawn at an unexpectedly low level in the evolutionary tree of life, since what he takes to be the essence of the mind is supposed to be present in a wide range of animal species, starting from arthropods. Contrary to Bermúdez, though, he grounds his approach on what I have been calling the standard belief-desire model of psychological explanation. As a first approximation to his view, consider the following quote from Carruthers (2004a):

> What does it take to be a minded organism, then? We should say instead: you need to possess a certain core cognitive architecture. Having a mind means being a subject of perceptual states, where those states are used to inform a set of belief states which guide behaviour, and where the belief states in turn interact with a set of desire states in ways that depend upon their contents, to select from amongst an array of action schemata so as to determine the form of behaviour. (p. 207)

It should be noted that Carruthers is strongly committed to the computational theory of mind (see 1.2), and therefore the "core cognitive architecture" he describes has to be understood in this context, viz. as involving symbolic structures that process and transform information about the environment and interact causally according to

algorithmic processes to generate behaviour. This core architecture is schematised below.

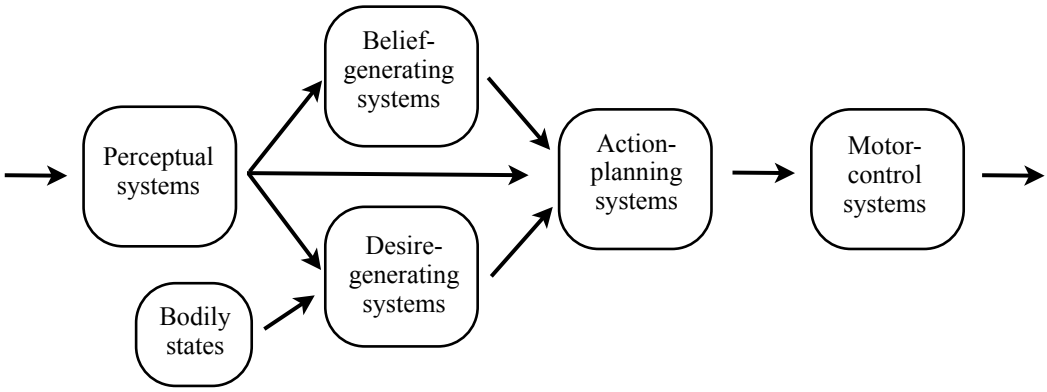

Figure 1: Carruthers' core architecture of the mind (adapted from Carruthers, 2004a).

So according to Carruthers' view, minded creatures must have perceptual systems capable of transmitting information about the environment to belief-generating systems and to desire-generating systems that also receive information from bodily states. Beliefs and desires are supposed to be symbolic structures capable of interacting causally one with another following algorithmic procedures commanded by action-planning systems. These systems are informed by beliefs and motivated by desires to generate action schemata, which are finally carried out by motor-control systems. Actually, Carruthers believes that animals typically possess several of these (belief/ desire-generating, action-planning, etc.) systems, which have been designed by natural selection to deal with different circumstances relevant for the animal. However, I shall not be concerned with these complexities here and instead focus on his fundamental claim that the core cognitive architecture depicted above is sufficient to make for a mind. Let us now focus on the arguments he puts forward to support his proposal.

First, Carruthers appeals to common sense to rule out what he regards as too demanding conditions for having mentality. For example, he describes Davidson's proposal that mental states must have contents specifiable in terms of public language, and that the only way to possess those contents is to be an active participant in a linguistic community (Davidson, 1984). Carruthers also mentions McDowell's claim

that to have a mind we need to be capable of evaluating the normative force of our reasons for believing and acting (McDowell, 1994a), and concludes that "common sense has little difficulty with the idea that there can be beliefs and desires that fail to meet these demanding conditions. This suggests, at least, that those conditions are not conceptually necessary ones" (Carruthers, 2004a, p. 205).

Then Carruthers turns to the opposite end of the spectrum and claims that it is also implausible to ascribe a mind to creatures whose actions are just driven by innate motor schemes or acquired habits learnt through some form of conditioning. In this respect the author is in line with the idea I put forward in chapter one that behavioural explanations just based on fixed action patterns or associative conditioning can be described from the physical level without there being any justification of using psychological vocabulary in them. And clearly, Carruthers's core cognitive architecture is more demanding than this. According to his view, symbolic structures informed by perception have to be capable of interacting in a flexible manner, and causing behaviour in way that reflects their representational properties.

A second line of argumentation takes the form of inference to the best explanation. Carruthers contends that many cases of animal behaviour that had previously been described in terms of innate or conditioned action patterns, actually demand a psychological explanation. He cites many cases of animal behaviour in support of his claim. For example, the capacities for navigation of bees and Tunisian desert ants, memory and learning in food-catching birds, and planning in jumping spiders and rats. Carruthers convincingly argues that these capacities cannot be explained by fixed action patterns or by associative mechanisms "unless those mechanisms are organised into an architecture that is then tantamount of algorithmic symbol processing" (2004a, p. 209), i.e. the core architecture put forward by him. Most of Carruthers' examples come from arthropods, and not without a reason. He believes that if he manages to demonstrate that various arthropods possess the mentioned core computational architecture, then it would be indeed very plausible to propose that this architecture is rife in the animal kingdom.

Let me briefly present the case of honeybees—Carruthers' leading example—to illustrate his proposal. As I have mentioned elsewhere in this thesis, honeybees have

notable navigational capacities that make them able to fly from their hives to sources of food and return. These capacities have been studied in detail by several scientists using techniques such as harmonic radar, which permits tracking the fly-paths of individual bees (Menzel et al., 2012). They have revealed that honeybees rely on landmarks and the position of the sun (as expected at a given time of the day) to orientate, and that they can use this information for dead reckoning (calculating their position by estimating the direction and distance travelled). This allows the insects to reach their destination by using a route they have never flown before, something that cannot be explained by appeal to fixed action patterns or associative conditioning.

In addition, honeybees can communicate their findings to other honeybees by performing a sort of dance inside the hive. Some features of the dance such as the angle of movement as measured from the vertical, and the number of "waggles" they make at some point, convey information about the expected angle relative to the direction of the sun for the time of the day and the distance to the food source. The bees in the hive are not just able to integrate this information and fly to the food, but also to evaluate it along a number of dimensions. For example, they are less likely to fly to distant sources of food, and show preference for rich sources of food. These findings suggest that honeybees are able to encode, store and perform calculations over different kinds of information, to then use it flexibly to reach their goals. Thus it seems plausible to conclude that the best explanation for these complex behaviours is that honeybees can carry out computational processes over causally efficacious and structured representations (see also Gallistel, 2009; Tetzlaff & Rey, 2009).

6.4.1 Objections

On first consideration, Carruthers's proposal appears to be intuitively plausible. It would certainly be chauvinistic to claim that to have mentality it is necessary to have the same complex psychological capacities we have, such as language or the ability to reflect about our own decisions. As Carruthers says, it is possible to conceive minds that are simpler than ours. But when he appeals to how limited are explanations based on innate or conditioned action patterns to account for complex forms of animal behaviour, as a means to justify his claim that mentality is widespread in the animal kingdom, I

believe he falls into a false dilemma in the sense explained in 2.2. That is, Carruthers fails to appreciate that psychological explanations are not the only alternative to innate or conditioned action patterns, for there is an alternative explanatory framework for animal behaviour that corresponds to what I have been calling the computational level.

To illustrate the point, it can be useful to look at some passages where Carruthers admits that some non-psychological forms of behaviour can be rather complex and allow some degree of variation. For example, he sets forth the mating behaviour of the male cricket. Crickets typically sing to attract mates, however sometimes this can also attract predators and parasite flies. So some crickets adopt the alternative strategy of staying silent near a singing cricket, and try to mate with the females attracted by the song. Interestingly, this strategy is flexible since the same cricket that stayed silent, when not nearby a singing cricket, can switch to the strategy of singing. Then, Carruthers (2004a) comments:

> Admittedly, such examples suggests that something *like* a decision process must be built into the structure of the behavioural program. There must be some mechanism that takes information about, for example, the cricket's own size and condition, the ratio of singing to non-singing males in the vicinity, and the loudness and vigor of their songs, and then triggers into action one behavioural strategy or the other. But computational complexity of this sort, in the mechanism that triggers an innate behaviour, isn't the same as saying that the insect acts from its beliefs and desires. The latter is what mindness requires, we are assuming. (p. 212)

Indeed, the cricket's computational capacities are limited. Although they can run algorithms that ramify onto alternative courses of action, their operations are otherwise rigid in the sense that they cannot be altered, for example, by further evidence the animal might obtain through perception. But Carruthers is wrong in assuming that what comes next, in terms of computational complexity, is the mental domain. The next step is, instead, just to be a computer. The reason is that what makes the cricket's mating mechanism so modest is that it cannot instantiate a universal machine, viz. it lacks the resources to perform the fundamental set of operations that characterise a computer (see 2.3). The computational capacities of the cricket are on this respect equivalent to those of the electronic calculator mentioned in 2.3; it has in-built mechanisms that can perform certain basic algorithmic procedures, but it cannot, even in principle, be programmed to run any other algorithm. Then, to repeat, Carruthers appears to be omitting a level of explanation that is between the physical (i.e. fixed innate patterns

and associative conditioning) and the psychological level of explanation: the computational level. And since as I argued in chapter 2 there is no reason for assuming that all computers have mentality, Carruthers seems to be committing the mistake of overlooking the possibility of animals being non-mental computers.

Someone might object to this critique by claiming that it wrongly assumes that any computational architecture that could in principle be realised in a machine, since it would then be explainable from the computational level, cannot make for mentality. Following this assumption, for example, it could be argued that Carruthers's proposal is mistaken simply on the grounds that it is possible to imagine his proposed core computational architecture being implemented in a robot. But this assumption is certainly misleading since it would lead us to reject that even humans have a mind, provided that we are committed to CTM and therefore accepting that we have a mind precisely because our brain is running the right computer program.

But I am not making this assumption, though. My point is not that Carruthers's proposal is unconvincing simply because it is possible to realise his core computational architecture in a machine, but because it is not at all clear that this core architecture is the right criterion to draw the line for what qualifies as a mind. Sometimes Carruthers regards his account as highly commonsensical, however this is open to question. Recall, for instance, the examples of the Mars rovers given in previous chapters. They are programmed with symbolic structures capable of representing aspects of the environment, and interacting with other structures that can motivate and guide action. Those machines appear to be running something like Carruthers' core computational architecture, however commonsense is certainly not aligned with the idea that they have mentality.

Perhaps what the Mars rovers lack is the capacity to carry out genuine reasoning. Indeed, Carruthers defends his view by pointing out that on his proposed cognitive architecture there are practical rules at work—such as a practical syllogism—and claiming that they correspond to exercises of practical reasoning. But note that if we are intended to differentiate between computational processes that count or not as reasoning, then to look at the logical form of their inferences is not of much help. For we can perfectly imagine a merely computational agent being commanded by an

inferential mechanism with the form of a practical syllogism, involving the manipulation of representational and motivational states, however without being compelled to describe it as a genuine reasoner.

I believe the same idea applies, for instance, to the navigational module inside a honeybee. The fact that they compute inferential processes that mediate between input and output mechanisms does not seem to be sufficient to describe them as genuine reasoning. It could be replied that human beings presumably also possess computational modules inside their heads—often described from a subpersonal level of explanation—and that there is nothing wrong in characterising them as instances of reasoning, and so we should do the same with the modules inside a honeybee. But I believe this reply does not work because Carruthers's core computational architecture in which the honeybee module is supposed to be embedded is significantly different from the cognitive architecture of human beings. As I shall explain with more detail in the final chapter, human computational modules can be described as subpersonal processes, that is, as part of a broader system that is properly described from a personal level of explanation. But this is something that does not happen with the honeybee module. It cannot be considered as subpersonal processing, because the overall computational architecture of the honeybee is not up to the mark for being captured by personal-level explanations (see the next chapter). Then, the core computational architecture put forward by Carruthers appears, however, at least as an incomplete attempt to account for what is paradigmatic of mentality. As I shall argue below, personal-level explanations provide a better framework for the purposes of capturing what is essential of psychological explanation.

In some passages Carruthers defends his criteria for mentality by appeal to the usual practice of cognitive scientists and comparative psychologists of describing animal behaviour in psychological terms. Carruthers (2004a) notes that the scientific literature on animal cognition "is replete with talk of information-bearing conceptualised states that guide planning and action-selection (beliefs), as well as states that set the ends planned for and that motivate action (desires)" (p. 206), and argues that this constitutes a good reason for taking the ascription of a belief/desire psychology to animals seriously, even if it clashes with some philosophical views. As Carruthers (2006) points out:

> The main point is that when science and philosophy come into conflict, it is generally the philosophers who should give way. For all that such non-empirically minded philosophers have to guide them is their 'intuitions'. And why should those count for much when set against the scientists' data and careful theorising? (p. 67).

However, Carruthers is misleading when he resorts to scientific practice to defend his view. On the one hand, it is possible that many scientists are also falling into a false dilemma when justifying the ascription of a belief/desire psychology to animals by appeal to the insufficiency of fixed action patterns or associative conditioning in explaining their behaviour. And on the other, some comparative psychologists actually recommend abstaining from attributing mental representations to animals. For instance, Chater and Hayes (1994) contend that the evidence obtainable from animal behaviour often underdetermines the nature of their inner structures, and that to describe them by using our own linguistic categories can barely be justified. Indeed, the authors conclude

> that advances can be made in the study of animal cognition by making pragmatic, flexible, and as far as possible, minimal assumptions about the content of animals' representational states. (p. 239; see also Shettleworth, 2010)

Carruthers might partly agree with this conclusion in the sense that we should remain neutral about the precise contents of animal representations, since those contents could, in fact, not be specifiable by means of our linguistic categories. He claims that, instead, we should better characterise them "from the outside, by means of an indirect description" (2004a, p. 206), thus making few commitments about their precise contents. But the problem with this external characterisation of the animal's inner symbolic structures is that it is, again, assuming that those structures have mental properties. The capacity to discriminate and categorise environmental properties and behave in an intelligent way that reflects those discriminations, can sometimes be explained from the computational level without the help of psychological notions. Some creatures might process information through complex computational architectures such as Carruthers', but do so by means of computational symbols, not mental symbols.

6.5 Conclusions

In this section I have discussed the proposals of of Bermúdez and Carruthers concerning what are the minimum conditions for having mentality. Both authors start

from a rather conventional cognitivist viewpoint, however Bermúdez develops an alternative model of psychological explanation that attempts to do without inferential processes. I have argued, though, that several tenets of Bermúdez's account are problematic and that his model of psychological explanation ends up being insufficient to explain some complex forms of animal behaviour.

Concerning Carruthers, he argues that the standard model of psychological explanation actually applies to most animal species, since they implement a computational architecture that can be described as a belief-desire system that causes behaviour. I contended that his reasons in defending this criterion for mentality are unsatisfactory, and sometimes rest on dubious appeals to commonsense and scientific practice. His proposed core cognitive architecture does not seem to capture what is paradigmatic of having a mind, or at least does it in a way that is incomplete in comparison with what in the next chapter I shall characterise as a personal-level approach to psychological explanation. In the following—and final—chapter of this thesis, I shall develop this idea and adapt it for the purposes of explaining the behaviour of non-human entities.

Chapter Seven

The Agent Level: A Proposal Towards Drawing the Line

7.0 Introduction

In this last chapter I will conclude this thesis by putting forward my own hypothesis on the correct way to draw the line that separates computational agents with and without mentality. My strategy will consist in identifying the contrast between psychological- and computational-level explanations with what is known as the personal-subpersonal distinction, by arguing that it provides an especially satisfying way to distinguish what are the main aspects of a properly psychological explanation.

After introducing the main tenets of the personal-subpersonal distinction, I attempt to vindicate the plausibility of adopting a realistic approach towards personal-level explanations by expounding how they can satisfy the requirements of explanation and supervention. Finally, I develop a non-human analogue to the personal level called "agent-level" and explore what constraints it imposes to the possibility of finding mentality in non-human computational agents.

7.1 The Personal-subpersonal Distinction

The *personal-subpersonal distinction* was first proposed by Daniel Dennett in 1969 and has been widely used and debated by philosophers of mind and psychology. In this section, I introduce the personal-subpersonal distinction focusing on Dennett's characterisation of it, to then in the following sections advance towards a more general formulation of this distinction. According to Dennett (1969), when studying the human mind and behaviour our explanations normally take place at a *personal level* of analysis, that is, in terms of activities and states that belong to a person. However, the same phenomena can also be approached from a *subpersonal level*, which instead of focusing on persons goes deep into the underlying cognitive or neural mechanisms that enable a person to have the mental or behavioural properties under scrutiny. He suggests that

once we understand the contrast between personal- and subpersonal-level explanations, we can appreciate how they can become complementary levels of inquiry about the mind.

Dennett gives the example of pain. This is a paradigmatic mental state that corresponds to a personal level of analysis, since explanations related to pain are typically formulated in terms of how they occur to people. By contrast, Dennett observes, knowing the subpersonal processes associated with pain will not enhance our understanding of how people experience or discriminate their pains. Moreover, he goes on saying that if we want to study what underlies a personal level explanation of pain, then

> we must abandon the explanatory level of people and their sensations and activities and turn to the *subpersonal* level of brains and events in the nervous system. But when we abandon the personal level in a very real sense we abandon the subject matter of pains as well. (1969, p.95)

This is not to say that a subpersonal account of pain cannot shed light on the nature of pain and its characteristic behaviour, though. Dennett's point is that when we shift to the subpersonal level we no longer have *pain* within the scope of our theoretical vocabulary, however we get further theoretical tools to explain the operations and mechanisms that make possible the phenomenon of pain. Thus understood, it is natural to associate the personal-subpersonal distinction with the idea of multiple levels of analysis depicted in chapter 1. Consider the example of an enzyme catalytic reaction. We can study this biochemical process from the viewpoint of basic physics, and in this way reveal the atomic interactions that happen during the enzymatic process. This explanation can undoubtedly enhance our understanding of the catalytic reaction itself, even though the term *enzyme* might not be present at this physical level of analysis. So the point is that it is possible to study certain phenomenon from the theoretical level (downwards) where it supervenes, even if that implies abandoning the vocabulary of the level where the phenomenon was initially described.

One controversial aspect of Dennett's example of pain, however, is that it alludes to a conscious mental state and thus may lead to the conclusion that personal level phenomena are always picked out in terms of consciousness. But Dennett understands the personal level in a more fundamental way, grounded not on whether mental states

are conscious but on their capacity to provide intentional explanations of people's behaviour (which indeed match psychological-level explanations, see 7.2.1). So from a more general viewpoint, personal-level explanations consist on those that can be properly ascribed to intentional systems, which Dennett (1979) defines as follows:

> An intentional system is a system whose behaviour can be (at least sometimes) explained and predicted by relying on ascriptions to the system of *beliefs* and *desires* (and other intentionally characterised features—what I will call *intentions* here, meaning to include hopes, fears, intentions, perceptions, expectations, etc.). (p. 271)

Dennett claims that when an agent exhibits behavioural patterns that can be satisfactorily explained and predicted by adopting an *intentional stance*, then we can qualify the agent as an intentional system and ascribe mentality to it. And even though alternative explanations can in principle account for the same behaviour (e.g. subpersonal-level explanations), the intentional stance arguably offers an advantageous explanatory approach when dealing with agents with minds (I shall say more about these explanatory advantages in 7.3.1). Intentional explanations can also be conceived as properly situated at a personal level; they make intelligible how a person could perceive and generate beliefs about the world, think and act on reasons.[37]

As is suggested by the term, the subpersonal level is commonly defined as concerned with the parts, or sub-systems, of people. In Dennett's words, "subpersonal theories proceed by analysing a person into an organization of subsystems ... and attempting to explain the behavior of the whole person as the outcome of the interaction of these subsystems" (1979, p. 153). The idea that personal-level explanations are properly attributed to a person as a whole, while subpersonal-level explanations describe the functioning of its parts, has become the most common way to draw the personal-subpersonal distinction.

To complete this initial depiction of the personal-subpersonal distinction, it must be noted that personal-level explanations characteristically have a normative dimension concerned with rational norms. Broadly speaking, behaviour is supposed to be governed by normative principles of (instrumental) rationality, which constrain how the person's

---

[37] A note of caution is in order here. In accordance with what I have said throughout this thesis, I adopt a realistic reading of the personal-subpersonal distinction, understanding it as mapping onto real natural domains of inquiry. But as it is well known, it is not clear whether Dennett would agree with that. His view sometimes oscillates from interpretativism to mild forms of realism (cf. Dennett, 1979, and 1991, for both extremes).

beliefs and desires come together to bring about actions, in order to satisfy its desires and goals. According to Dennett's particular approach, as a result of being constrained by rational norms, personal-level explanations proceed by ascribing beliefs and desires a person *ought* to have in order to satisfy its goals, according to rational standards such as consistency and coherency. In 7.2.4 I shall argue that this idealised conception of rationality is misguided, however, and present an account of rationality more suitable for the present thesis.

This sort of normativity is supposed to constrain the behaviour of a whole rational agent, and thus not be operative at the subpersonal level. Even though subpersonal-level explanations can have a normative dimension—in particular when computational capacities are understood teleologically and therefore aimed at certain ends or goals—this normativity is concerned with the functioning of particular subsystems and not with explanations that rationalise how a (whole) person attains its desires and goals. I will say more about why ascriptions of rationality demand a personal-level approach in 7.4.3.

By way of summary, the main ideas behind this preliminary characterisation of the personal-subpersonal distinction are contrasted in the table below. In the following sections of this chapter I delve more deeply onto these ideas, to then conclude with a revised version of personal-level explanations attempted to map neatly on what I have called the psychological domain.

	Personal level	**Subpersonal level**
Subject matter	Behaviour and cognitive capacities attributable to a whole person	Mechanisms and subsystems that make personal level capacities possible
Theoretical vocabulary	Couched in terms of commonsense psychology	Couched in terms of computation and informational theories
Normative dimension	Standards of rationality apply	No place for rational standards

7.2 Specifying the Distinction

7.2.1 The Personal-subpersonal Distinction and Hierarchical Levels of Explanation

In the first chapter of this thesis I delineated the common picture of distinguishing three hierarchical levels of explanation for the mind, viz. the psychological, the computational and the physical level. A natural way of extrapolating this analysis into the personal-subpersonal distinction presented above, is to identify the psychological level with the personal level, and the computational and physical with the subpersonal level. Let me explain with more detail.

The psychological level as presented in 1.4.1 deals with the subject matter of personal-level explanations and deploys the same theoretical vocabulary. Both levels attempt to make understandable the behaviour of a cognitive agent taken as a whole, and proceed by ascribing it propositional attitudes governed by norms of reason. Their explanations roughly coincide with the categories of commonsense psychology, and generally strive to legitimise their vocabularies and generalisations under a the light of science. Subpersonal-level explanations, on the other hand, can be identified with both the computational and the physical levels, given that the two are situated "below" the level of whole persons.

Take again the case of pain. We have two ways of elucidating this personal level phenomenon. One could be properly computational, and proceed by analysing how information coming from nociceptors is coded and categorised though different computational stages of processing. On the other hand, an alternative explanation could consist on a straightforward physical account describing, for instance, the neural pathways and regions of the brain that are activated during painful experiences. Both computational and physical level explanations can be regarded as complementary, but it is worth pointing out that in the context of cognitive science it is customary to situate the computational level as the one immediately "below" the personal level. Then following common usage, and for the purposes of mapping the personal-subpersonal distinction onto the standard threefold analysis of the mind already mentioned, I will identify the subpersonal level with the computational level.

7.2.2 Subpersonal-level Explanations are not Purely Syntactic

In section 2.3.1 I argued against the view that computational-level explanations just correspond to a formal or syntactic description of the computational operations performed by mental creatures. A similar view is sometimes adopted by defenders of the personal-subpersonal distinction, and can be tracked back to Dennett (1982) and his assertion that while the mind is a semantic engine, when we look at the brain all we find is a syntactic engine. John McDowell (1994b[38])—another prominent defender of the personal-subpersonal distinction—pushes this idea forward when he writes:

> we have inside us something that is not intelligent at all (it knows nothing and understands nothing); even so, we can be enormously helped in finding it comprehensible how we can be intelligent, [by means of subpersonal-level explanations] .... That makes it possible to understand how this mindless internal control system enables us to do what it takes to display genuine mindedness, namely to live competently in an environment. (p. 200)

McDowell concedes that the subpersonal level is an invaluable approach to make sense of the mental phenomena that figure at personal-level descriptions, since it can show us the computational mechanisms that "enable" us to possess mentality. However, he suggests that we should distinguish this enabling explanation from a "constitutive" one, that is, from explanations about what literally grounds or fully explains the phenomenon under study. According to McDowell, the enabling conditions described at the subpersonal levels are concerned with the syntactic component of the mind, leaving the representational component completely outside the scope of subpersonal-level explanations (indeed, according to McDowell any ascription of representational contents at the subpersonal level must follow an "as if" fashion).

I believe this way of framing the personal-subpersonal distinction is misleading, though, for the following reasons. First, the way it relates both levels is unclear and inconsistent with a scientific picture of hierarchical levels of organisation, where all the processes at a higher level are supposed to be implemented by processes situated at the next level down. And McDowell's notion of "enabling explanation" does little to clarify how the personal and subpersonal levels are related, and ends up being too weak to account for inter-level relations. Something can be the enabling condition for certain phenomenon, in the sense of explaining how it can be possible, but be just a partial

[38] All page quotations in this section correspond to this paper.

account of what underlies that phenomenon. This could be the case, for instance, of the relation between a power supply and a the sounds emitted by a radio. The power supply enables the radio to emit sounds, but much more than the power supply is needed to account for what grounds that phenomenon. In contrast, the subpersonal level is supposed to offer a complete account of what underlies personal-level explanations, and therefore is not just an enabling but also a constitutive condition for the mind. In one part (p. 203) McDowell tries to shed light on the notion of enabling condition by associating it with a causal explanation, but this only further distorts the picture. Higher levels do not stand in a causal relation with lower ones, but in a relation of supervenience. And if supervenient levels do not constitute the metaphysical ground for the levels situated "above", then those higher levels remain as a floating mystery.

Another way of showing why McDowell's view is unsatisfactory, can be to consider his claim that "a brain knows nothing and understands nothing" (p. 201). The rationale behind this assertion is that the brain is normally studied from a subpersonal perspective, one of computation and information-processing where terms such as *knowing* and *understanding* do not exist. But even though in some sense this is true, it is just a consequence of the division of labour proper of the scientific approach of multiple levels of analysis, not a fact about the nature of the brain. Take the analogue case of the assertion "human bodies cannot feel the ambient temperature". Even if it would be right to say the the theoretical vocabulary that properly describes bodies (say, physiology) has no place for the term *feel*, this does not mean to say that bodies cannot feel. Bodies do feel, but how we describe this phenomenon will depend in which level of analysis we adopt. Therefore, brains do understand, however this assertion has to be understood as formulated from one particular level of description.

7.2.3 The Personal and Subpersonal Levels do not Collapse into a Single Level

Some philosophers, in particular some who endorse psychological functionalism, tend to assimilate explanations couched at the personal level with subpersonal-level explanations. A common motivation behind this has to do with the attempt to legitimise commonsense psychological explanations by analysing them at a

functional level that is somewhat continuous with a rigorous and scientifically grounded functional description of the workings of the nervous system.

One consequence of a straightforward personal-subpersonal assimilation can be the elimination of personal-level explanations altogether. That would happen if the personal level ends up playing no distinctive explanatory role and the phenomena it was purported to explain is totally redescribed in subpersonal terms. Even though this is not the place for assessing arguments for this sort of eliminativism, it is worth mentioning some reasons for resisting it. One is that the personal level does have its explanatory merits, which I shall defend in section 7.3.1. A further reason for preserving this level can be deduced by looking at the term *subpersonal*; it is conceptually tied to the notion of personal level, which it is supposed to elucidate. But the point is more than just terminological. Personal-level descriptions are required for isolating the relevant subpersonal processes we need to explain human behaviour, and without an explanatory vocabulary that includes personal-level categories such as perception, decision-making and action, we would be left without a way of telling apart computational agents with and without mentality. I expand this idea in section 7.3.2.

Some philosophers who follow psychological functionalism have attempted to preserve something like a personal-subpersonal distinction, however I believe they have done so in a misleading way (see below). They concede that there is a relevant distinction to be made between a description of a whole person's behaviour and a functional analysis of its capacities in terms of subsystems, but according to their view personal-level explanations just correspond to a subset of subpersonally characterised computational processes[39]. One example of this view is the approach of Braddon-Mitchell and Jackson (1996, ch. 14), who defend a version of commonsense functionalism that distinguishes between personal and subpersonal levels[40]. According to them, both levels correspond to a functional description of how an agent picks up, transforms, and processes information in a way that yields intelligent behaviour. But on their view, the relation between personal- and subpersonal-level explanations is

[39] This way of characterising the personal-subpersonal distinction was also suggested by Stephen Laurence (in conversation).

[40] It must be noted that the authors do not employ a personal-subpersonal terminology, but distinguish between psychology and a functionalist account of the workings of the nervous system. Since that distinction roughly matches the personal-subpersonal distinction discussed here, and to simplify the exposition, I present their proposal in personal-subpersonal terms.

equivalent to “the relation between limited explanations and one or another underlying complete explanation” (p. 259).

So according to Mitchell and Jackson, the personal level constitutes a partial, limited version of the full story of information processing that the subpersonal level is in principle capable of conveying. And instead of being a limitation, they claim this constitutes an advantage because personal-level explanations are supposed to include the main causal generalisations and counterfactuals that matter for the purposes of explaining behaviour (e.g. by including the necessary conditions required for the causal generalisations to obtain). So even though explanations couched at the personal level are incomplete and restricted in comparison with subpersonal-level explanations, the former are vindicated because they capture the essential part of the story that matters for our understanding of the phenomena under study.

Take for example an explanation of the space shuttle Challenger accident. A complete explanation of what caused the accident would involve a massive amount of detail about the physical structure of the rocket booster that initially exploded, the chemical constitution of its propellant, the flow of gases inside and outside it, the aerodynamic forces impacting the spacecraft, etc. In contrast, a more concise explanation could be that what caused the accident was a failure in one of the mechanical gaskets of the right rocket booster, which produced a escape of high temperature gas that lead to the breakup of the spacecraft. These two explanations (the complete and partial) can be viewed as an example of the contrast between subpersonal- and personal-level explanations, respectively. Even though the latter has much less detail and precision, it is preferred because it captures the most relevant events that matter for our understanding of the accident.

In this sense, we might say that subpersonal-level explanations give a complete account of the computational processing of information in the brain, while the personal level offers a partial version of the same story but focused on processes that are more relevant for the purposes of explaining human behaviour. One obvious critique to this approach is that in the context of scientific realism we are looking for the best explanation, and not just for the most easy or tractable one. Even if a complete subpersonal account is hard to understand, or even if we are incapable to formulate it

given our current state of scientific knowledge, these are not good reasons for preferring a shortened version of it. We want the theory that best explains how the world is, not the one that is more convenient for our particular purposes.

Braddon-Mitchell and Jackson are aware that their proposal cannot be grounded solely on this sort of epistemic advantage, though. They argue that the partial (personal-level) explanation can provide us with a language with the "resources" to describe "in a non-trivial manner" the relevant counterfactuals that make clear why the behaviour under study actually happen to be the case. They go on to claim that, in this sense, the explanations couched at the personal level might be regarded as autonomous from other explanatory levels.

I believe their arguments do not support the autonomy of personal-level explanations, though. If they constitute just a partial account of what goes on at the subpersonal level, it is not clear why personal-level explanations could be regarded as having a distinctive theoretical vocabulary, or as offering more resources to deal with counterfactuals than subpersonal-level explanations. Returning to our example of the Challenger, the shortened version of the story of the accident can clearly be helpful for purposes of understanding or communicating it, but it can hardly be defended that it has more resources than the complete explanation. And even if the short story proves to be more useful for grouping the relevant data and for describing it in a more concise manner, it would still be formulated in the same theoretical vocabulary and following the same laws of the complete explanation, and thus cannot be conceived as being formulated at a different level of organisation in the way the personal level is supposed to be.

7.2.4 Normativity does not Imply Radical Autonomy

So far in this thesis, I have introduced two senses of autonomy. One is that of an autonomous agent (see 2.3), which corresponds to an embodied machine or animal capable of self-governing and self-organisation. The second is the autonomy of levels of explanation in science. As explained in 4.1, mental agents are normally studied in terms of three hierarchical levels of organisation: the psychological, the computational and the

physical levels. Each level is supposed to be autonomous in the sense of having its own theoretical vocabulary and predictions, however that autonomy is constrained in the case of higher-levels since their processes must supervene, and therefore be determined by, the processes of the next level down. Psychology, qua the top level of scientific explanation of the mind, is autonomous in this second sense.

One controversial aspect of psychology—understood as personal-level explanation in the present context—is that it has a normative dimension. This has been considered as the source of another sense of autonomy, which I shall call the *radical autonomy* of psychology. This is the view that psychology has an explanatory method that is fundamentally distinct from that of other scientific theories, to the extent that its explanations are radically incommensurable with respect to subpersonal levels of explanation (Crane, 1999). The basic idea is that normative principles of rationality apply at the personal level and constrain its explanations in a way that makes them prescriptive rather than descriptive. The point is clearly put by McDowell (1985) in the following passage:

> [Psychological] explanations [are those] in which things are made intelligible by being revealed to be, or to approximate as being, as they rationally ought to be. This is to be contrasted with a style of explanation in which one makes things intelligible by representing their coming into being as a particular instance of how things generally tend to happen. (p. 389).

In sum, according to radical autonomy the personal-level explains people's behaviour in terms of the mental states they ought to have in case they were ideally rational, and not according to behavioural data gained through empirical discovery. This is certainly at odds with the scientific picture of levels of organisation presented in this thesis, since radical autonomy opens an unbridgeable gap between the personal and subpersonal levels (see also Dennett, 1987; Davidson, 1980). In the remainder of this section, I argue that radical autonomy does not follow from the assumption that psychology conforms standards of rationality, and that personal-level explanations can be safely kept within our naturalistic framework.

The line of argumentation followed by the aforementioned authors is, at least in an important respect, transcendental: compliance to ideal standards of rationality is considered a presupposition for the very possibility of psychological explanation. In this

sense we cannot expect to predict human behaviour without starting from the ascription of beliefs and desires the person ought to have according to rational norms such as coherency and consistency. One problem with this approach relates with its idealised conception of rational norms. When discussing Dennett in 2.1.1, I argued that the imposition of ideal standards in our interpretation of behaviour is poorly revealing about the nature of cognition in general, because cognitive capacities simply lack an optimal design and normally perform below standards of full rationality. As Samuels, Stich and Faucher (2004) suggest, perhaps what the psychological evidence is suggesting is not that people are systematically irrational, but instead that the idealised standard used as a benchmark against which people's reasoning is evaluated has been wrongly conceived. For example, they have proposed to develop a "resource-relative" conception of rationality where the normative standards are relativised according to the cognitive resources people actually have, as a way to avoid imposing standards that go beyond human capacities. I return to this idea at the end of this section.

A second misguided motivation for radical autonomy relates to the alleged indeterminacy of psychological explanations of behaviour. The claim is that psychological evidence is characteristically insufficient for explaining the causes of behaviour, and that therefore a large rational pattern of mental states has to be assumed for the purposes of explanation. This would make personal-level explanations prescriptive, given that those patterns have to be framed by the interpreter in compliance with standards of rationality. In words of Davidson (1980), "we must warp the evidence to fit this frame" (p. 36). In section 2.1.1, I tackled a similar argument put forward by Dennett, according to whom since there is an unavoidable degree of indeterminacy in psychological explanations, we must presume rational patterns of mental states in order to make sense of behaviour. In response I argued that even though a certain degree of indeterminacy might always be present, it is often possible to collect a reasonable amount of behavioural evidence in order to make psychological hypothesis scientifically respectable and susceptible of further empirical verification. And even if in some cases a lack of evidence makes it hard to formulate strong psychological generalisations, it could correspond just to a case of epistemological rather than metaphysical indeterminacy (Bermúdez, 2005). So there appears to be no a priori reason for denying that psychological hypotheses could be verified in the long run after

appropriate experimental techniques caught up, overcoming our current epistemic limitations.

But what is the place of rational norms in pscyhology, then? The goal for present purposes is to account for a naturalised notion of rationality that does not lead to radical autonomy, and is at the same time compatible with a personal-level approach. At a minimum, a person's behaviour is described as rational when it is the result of a process of thinking (i.e. involving beliefs and desires) that satisfies certain normative principles of reasoning. When we try to specify the nature of those principles, however, we might wonder where those standards come from and in what sense they normatively constraint psychological explanations. I tackle these issues in the remainder of this section.

As explained above, it is problematic to impose idealised standards of reasoning that go beyond human cognitive capacities. This happens, for example, when those standards are derived from formal theories such as classical logic and probability theory. But instead of prescribing ideal standards on a priori grounds, it is much more sensitive with the psychological evidence on human reasoning to relativise those standards to human cognitive capacities and limitations, along the lines of the aforementioned resource-relative conception of normative standards (cf. Cherniak, 1986).

When it comes to determine how those resource-relative standards apply to psychological explanation, it is important to distinguish them from psychological generalisations. The latter correspond to lawful connections between people's mental states and behaviour, which allow psychological explanation and prediction. Standards of rationality, instead of attempting to predict behaviour, enter in psychological explanations as part of the story we tell about what mental states could underlie the generalisations we have previously established empirically (Fodor & Lepore, 1995). They permit us to evaluate whether people's mental states and behaviour constitute a rational means for attaining their goals, and also open the possibility of mistakes when those means are instrumentally irrational. We would expect, of course, that those rational standards also give us some predictive power; after all, they are supposed to reflect the strategies people actually follow in order to satisfy their desirable goals. However, the key idea is that rational norms do not override the descriptive nature of psychological generalisations.

It might be objected that at the very moment we start talking about standards we are imposing some sort of idealisation, leaving out from the scope of psychology cases of systematically irrational behaviour, which even though abnormal or pathological, are actually possible. But this idealisation is innocuous insofar as the standards are not fixed a priori but result from empirical research about cognition, and behaviour that is not up to those standards can be regarded as exceptions to ideal conditions under which scientific theories normally operate. For example Fodor (1974) has compared psychology with other special sciences such as geology, whose generalisations are typically ceteris paribus, that is, they admit potential exceptions, which are normally only statable in the vocabulary of more fundamental sciences (which they supervene on). These ceteris paribus conditions correspond to a form of idealisation, however they are ubiquitous in most scientific disciplines (Rey, 1997, ch. 10).

In section 7.4 I shall return to this discussion and explore how this normative dimension characteristic of the personal level could be adapted for the purposes of differentiating agents with and without mentality.

7.3 Assessing the Distinction

So far I have introduced and explained the personal-subpersonal distinction, with the aim of identifying it with the distinction between the psychological and computational levels of explanation (identifying for present purposes the subpersonal level just with the computational, and not with the physical level, as explained in 7.2.1). This involves the assumption that personal- and subpersonal-level explanations correspond to autonomous levels of organisation, and that they form part of our scientific picture of agents endowed with mentality. This idea, however, needs some tiding up in the way of showing the legitimacy of the personal-subpersonal distinction. In the following sections I evaluate its plausibility, and for that purpose I discuss how the psychological level can satisfy two requirements for being a plausible top level of explanation of the mind (adapted from Devitt, 1991):

- *Explanation Requirement*: The personal level must perform an explanatory task that is not performed by subpersonal levels, and thus give us additional explanatory power.

- *Supervention Requirement*: The personal level must supervene on the subpersonal level in a way that makes plausible its implementation on levels lower down in the hierarchy.

Some comments on these requirements. The explanation requirement has to be understood in the context of scientific realism presented in 1.1.2. According to this view, what vindicates the personal level as a genuine metaphysical level is that is constitutes the best scientific explanation we currently have to account for certain behaviour. Even though in principle any mental phenomenon could be given a subpersonal description, the personal level is supposed to provide us with a theoretical framework capable of capturing generalisations linking mental states and behaviour that would otherwise be missed from a strict subpersonal-level viewpoint, and increase our predictive power.

The supervention requirement, on the other hand, is a constraint from naturalism (see 1.1.1). In order to be compatible with the metaphysical picture of hierarchical levels or organisation, personal-level states and processes need to supervene on the states and processes at the subpersonal level, in a way analogous to the way that computational states are implemented in physical devices. A consequence of the supervention requirement is that the autonomy of the personal level as a special science is not unrestricted. In order to justify the ascription of the personal level, there must be a plausible way of implementing it in the subpersonal level.

At this point it shall be useful to introduce the distinction between *horizontal* and *vertical explanations* (Bermúdez, 2005). Suppose we ask for an explanation for an enzymatic reaction that catalyses the conversion from molecule X to molecule Y. An explanation from chemistry would say, for example, that an enzyme binds molecule X and lowers the activation energy that this molecule requires to reach a transition state in which it is transformed into molecule Y. This is an example of horizontal explanation insofar as it describes a causal series of events between entities couched by the vocabulary of chemistry and subsumed by chemical laws. The same enzymatic reaction,

though, could be explained from the level of physics. In this case, we would have an horizontal explanation involving causal generalisations holding between entities described by the vocabulary of physics, such as those involving sub-atomic components.

Notably, there is a gap between both horizontal explanations. Each can run rather independently, in the sense that physics and chemistry constitute autonomous disciplines, however some relation must exist between them in order to make the whole picture of levels of organisation plausible. More precisely, there must be a way in which both horizontal levels are related, an account of how chemistry could be grounded in physics. This need of inter-level explanations corresponds to the supervention requirement, or what Bermúdez (2005) has called the *interface problem*. He has proposed that this inter-level relation is provided by vertical explanations, which explain how entities and processes of higher-level theories could hold at more fundamental lower-level theories. Returning to our example, a vertical explanation related with the chemical reaction could be a description of the sub-atomic structure of molecule X, or a quantum-mechanical account of the chemical process through which the enzyme accelerated the catalysis.

As can be suggested by the previous example (involving physics and chemistry), the limits between horizontal explanations are not always sharp, and when it comes to explain complex phenomena such as the mind it is common that explanations oscillate among multiple levels. Consider the case of a computational-level explanation of visual perception. It would certainly include an account of how patterns of intensity and changes in light are computed through a series of informational structures that end up generating a 3-D object-centred representation (see 1.4.2). However, note that a complete subpersonal explanation of vision typically includes elements from lower and upper levels. First, the input informational structure is the product of transducers, which convert light energy onto neural signals, a process described at the physical level. And at the other end of the computational process, the 3-D representation correspond to the activity of seeing an object on the part of a whole agent, something normally couched in personal-level terms.

This multi-level character of explanation does not contravene the idea of there being a hierarchy of autonomous levels, though. The point is that it is often useful—and perhaps necessary—to explain certain complex phenomenon by appeal to aspects of the same phenomenon couched from different explanatory levels. And note that this practice is not peculiar of explanations of mental agents, but also customary in neurobiological sciences in general (Craver, 2007).

7.3.1 Meeting the Explanation Requirement

In this section I compare how personal- and subpersonal-level explanations account for certain characteristic mental and behavioural phenomena, with the purpose of supporting the claim that personal-level explanations can satisfy the explanation requirement. More precisely, I attempt to show that (1) some explanations couched in personal-level terms are substantially different from an alternative subpersonal-level account of the same phenomenon, and that (2) often subpersonal-level explanations involve details that are largely irrelevant for grasping the phenomena described at the personal level. I present my arguments in the context of the three main stages of personal level explanation, viz. perception, reasoning and action-systems.

Suppose that John is on holidays visiting a city he does not know very well. At some point he is wandering through a street and sees the face of an old enemy. He then suddenly turns around and starts walking in the opposite direction, heading back to his hotel. After a while, however, he feels a bit lost, and decides to stop and figure out which direction to take. Even though he cannot recognise any known landmark in his surroundings, after a while he believes he is not far from the hotel, but just a few streets north. In fact he was right, and after taking a route south he manages to find his hotel.

This explanation incorporates facts about John's perception, in particular his capacity of generating a perspective of his surroundings, and focusing on particular features of it. Many philosophers, going back at least to Kant, have found appealing to say that this capacity is the one of the most distinctive features of the mind. For example, Crane (2001) asserts that what distinguishes minded from non-minded

creatures is that only the former are capable of generating "a point of view on things" or a "perspective" (p. 4). But what does such a perspective consist of?

As a first approach, to have a perspective is something more complex than to passively re-present the immediate environment, in the way a screen connected to a videocamera might be able to do. It also involves identifying elements (such as a face) in the visual field and keeping track of them. This is somewhat achieved by the perceptual capacity to represent environmental invariants, which as I explained in previous chapters many authors have identified as a hallmark of mentality. But as I have argued, the capacity to maintain an informational link with certain environmental object can easily be ascribed to non-minded computational agents. To get an idea of what is missing, consider what appears to be essential of having a perspective: the ability to find out the place where one is situated. As the previous example shows, John is not just identifying a face, but also situating it within certain frame of reference. This is certainly more demanding than keeping track of individual objects, and involves simultaneously identifying several objects and their spacial relations. Evans (1982) formulates the idea neatly in the following paragraph:

> A perceptual input—even if, in some loose sense, it encapsulates spatial information (because it belongs to a range of inputs which vary systematically with some spatial facts)—cannot have a spatial significance for an organism except in so far as it has a place in such a complex network of input-output connections. (p. 154)

The perspective proper of a minded agent then appears to consist in an integrated space of representation, from where the agent can identify and take elements for further cognitive use (for a development of this view, see Proust, 1999).

Now let us examine how this general account of perception could be explained from a subpersonal-level approach. Consider the process of seeing and recognising a human face. From a subpersonal viewpoint (and returning to Marr's standard example of a computational theory of vision), this process involves extracting information about an environmental invariant, and coding it through a series of computational stages involved in the formation of a 3-D representation of a distal environmental object. Among these intermediate computational stages, information about the human face first goes though an analysis of local geometrical structure called primal sketch, that is then

passed through a second stage called 2½-D sketch where the visible surfaces of the face are represented (see 1.4.2 for more details).

When comparing this subpersonal-account with that of the personal level, it soon becomes clear that the former has a different focus and involves information that is largely irrelevant for the personal-level explanation. Beyond doubt this subpersonal-level analysis permits us to gain a deeper understanding of the computational operations underlying the representation of a face, however to include a primal or 2½-D sketch of the face in our explanation of why John turned around would clearly be otiose. The theoretical vocabulary of personal-level explanations describes how John gains access to its environment as a rather immediate process, and its implementational informational-processing details are left out of the personal-level story. In a slightly different context, Dretske (1981) makes a similar point:

> Information about angles, lines and gradients is obviously used in the production of a perceptual belief (e.g. a truck passing by), but this information is (or may be) systematically eliminated in the digitalisation process by means of which a final semantic structure is synthesised. (p.200)

To repeat, this is not to say that informational structures such as a primal sketch are irrelevant for the purposes of explanation. For present purposes, the point is that structures of this kind do not take part in personal-level accounts involving seeing, believing, and the like. What seems more appropriate is to say that structures such as a primal sketch correspond to a subpersonal level of description, which provides enabling conditions that ground perceptual processes, in the way of vertical explanations connecting the personal and subpersonal levels. In any case, the moral is that the personal-subpersonal distinction reflects the existence of two genuinely explanatory levels of analysis.

To make another personal-subpersonal contrast, let us return to the idea that a personal-level explanation of John's perspective appeals to a dynamic and integrated space of representation. To find a subpersonal-level correlate of this explanation we can look at accounts of the computational processes that implement spatial orientation. A prominent example of this approach has been put forward by Gallistel, who claims that most animals have a complex system of navigation based on a *cognitive map*, defined as follows:

> A cognitive map is a representation of (at least some) geometric relationships among a home site, terrain surrounding the home site, goals to be visited and the terrain surrounding those goals (Gallistel and Cramer, 1996, p. 211)

According to this subpersonal-level account, a cognitive map is built by the combination of body-centred and earth-centred vectors, in the way of a map of coordinates and geometric relationships. It permits creatures to orientate themselves and navigate by path integration, which involves the ability to record and compute vector trajectories previously travelled and to calculate the vector to take in successive displacements. This job is probably carried out by domain-specific navigational capacities, and appears to be rife in the animal kingdom (see chapter 2 for path integration in wasps and honeybees).

Can this subpersonal-level explanation subsume the personal level explanation given for John's trip back to his hotel? I think here again we find a case where both explanations run separately. A personal level explanation involves something like the capacity to generate a mental map of his surroundings, including the relative position and distances between objects and places. When John figures out where he is and comes up with the belief that he is a few streets north of his hotel, he simply reaches that belief out from his navigational abilities, in the same way as the visual perception of an object appears in the visual field after opening the eyes in front of it. There is no clear way of justifying this belief by appeal to anything beyond his navigational abilities. The vector integration and calculation that underlies these capacities do not appear to be attributable to John in the same sense as we ascribe him the decision to walk back to his hotel. Those computational operations seem in fact to involve informational structures that never reach the level of belief, and instead happen in a deeply automatic and domain-specific fashion.

To conclude this defence of the distinctive character and explanatory virtues of personal-level explanations, let us abandon the example of John and focus on a pathological condition known as *blindsight perception*. Due to a lesion in the primary visual cortex, patients lack visual awareness of certain region of their visual field. But surprisingly, they remain visually responsive to light stimulus presented in the blind area of their visual field, even thought they report having not seen anything. For example, patients are asked to press a response key when are exposed with a visual

stimulus. When the investigators present stimulus in their blind area, there is a significant increase in the pressing response, demonstrating the existence of implicit processing of the unseen stimuli (Stoering, 1999).

The act of pressing a key in response to visual stimuli is a certainly a case of non-reflexive, complex behaviour, susceptible of a personal level description. However, something puzzling with these experiments is that it seems wrong to give a personal-level explanation to this, given that the relevant stimulus does not appear in the visual field, nor to integrate what we might call the perspective the subject has of its surroundings. And this does not seem to be just a matter of conscious recognition, since the stimulus does not appear to be possibly retrieved or to affect the perspective of the subject in any way (of course, it could affect the subject's representational space after being informed about the results of the experiment, however that would not be a consequence of the perception of the stimulus itself). Indeed, the stimulus in question is not considered by the subject in making a decision to press the key. In contrast, a subpersonal-level account appears to be much more appropriate, probably offering some detail about the damaged brain structures responsible for such as case of abnormal behaviour.

I believe that cases of blindsight can serve to highlight the personal-subpersonal distinction, by showing how what under normal conditions would be a typical personal-level phenomenon, has to be explained in the radically different language of the subpersonal level (in this case probably in neuroscientific terms). The case of blindsight is an abnormal case of behaviour, where the ceteris paribus character of psychology is notorious; it constitutes an exception to personal-level explanations, where we have to abandon the level of the whole-person and turn to its functional and neural components. In sum, the blindsight example supports the idea that the personal level does a different explanatory job from that performed by subpersonal-level explanations, and that it is convenient to keep this distinction to account for pathological cases where what would normally be typical personal-level behaviour has to be given a subpersonal-level account.

7.3.1 Meeting the Supervention Requirement

In 7.3 I explained how personal and subpersonal levels map onto parallel horizontal explanations, and that part of what is understood by satisfying the supervention requirement is to account for vertical explanations linking both levels. So far, I have already mentioned some ways in which those inter-level relations could be formulated. One can be by providing the grounds of certain personal level states. For example, what makes it possible seeing and recognising an object could be spelled out through a vertical explanation of the computational stages involved in the formation of perceptual constancies and the integration of percepts coming from different sensory modalities. Alternatively, if accounts of conceptual content such as prototype theories are correct, a concept we single out at the personal level might not be grounded in a single informational structure, but in an interconnected set of informational structures (e.g. Rosch, 1978).

Another example of vertical inter-level relations can be found in processes of reasoning, which from a personal-level approach correspond to inferential transitions between thoughts that figure in causal explanations of behaviour. These inferences have a formal dimension that is subject to rational norms, and which individuals are often not even aware of. The way those inferences are implemented at the subpersonal level is by means of computational process, which allow us to explain how the transitions between thoughts could be mechanised. They provide a vertical explanation that unveils the mechanics for a process that when couched in the personal-level terms simply describes inferential transitions carried out in virtue of their form[41]. Again, there need not be an isomorphism between the personal and the subpersonal level, since the machinery for what from personal level viewpoint is a single transition might in fact be instantiated in a larger series of computational steps.

A final case of vertical explanation, relates to abnormal conditions such as the case of blindsight explained in the previous section. The basic idea is that the generalisations captured by personal-level explanations are ceteris paribus in the sense of being forged within certain idealised conditions. That means that they admit exceptional situations where their purported predictions do not apply, in particular when

[41] A similar example is put forward by Davies (2000). Contrary to my approach, however, Davies takes the subpersonal level to be purely syntactic, in the sense described in 7.2.2.

it comes to pathological conditions such as blindsight. In those cases, subpersonal-level explanations fill the gap, so to speak, left by these exceptional situations that our personal-level explanations are unable to account for.

Something suggested by these examples of inter-level relations is that personal-level processes can be implemented in different configurations of subpersonal-level processes. This opens a way of spelling out the relation of supervenience that holds between both levels, by appeal to the notion of *multiple realisation*, the claim that processes couched in higher-level theories can be realised by many distinct processes described by the lower-level theory where the former theory supervenes. It is widely accepted by philosophers that the argument of multiple realisation has vindicated the metaphysical status of a functional (or computational) level of explanation, leaving it as an autonomous level of theoretical investigation about the mind (Kim, 1988).

Many authors have proposed that the notion of multiple realisation can also be applied to explain the relation between (what I am calling) personal- and subpersonal-level explanations (e.g. Horgan, 1992; Putnam, 1988). A good way to see the plausibility of this proposal can be to appreciate that something similar happens with the programming languages of some computers. In those cases, higher-level programming languages can be compiled in lower-level ones, with the possibility of their being various programs nested between them in a way that resembles a hierarchical organisation of explanatory levels. Higher-level programming languages, so the argument goes, can be realised in different compiled languages (which are directly implemented in the physical machinery of the computer) and thus be multiple realisable. A problem with this idea, though, is that again we appear to be situated within a single general computational level of description, which, however having multiple nested levels of programming, does not necessarily imply that at the top of the hierarchy we have a level compatible with a personal-level description.

One way to justify the ascription of a personal-level of description at the top of a hierarchy of multiple realisable functional levels can be adapted from what is known as *homuncular functionalism* (Lycan, 1995; see also Dennett, 1979). According to this view, explanations at all levels are analysable from a functional viewpoint. For instance, bodily organs such as the kidney have a function that can then be analysed in terms of

its sub-systems, which serve more specific functions. And if we deepen into still lower layers of analysis we find that cells are functional structures that are constituted of smaller cooperating organelles that fulfil even more specific functions. The same idea runs for subpersonal-level explanations. Computational processes can be analysed in terms of sub-processes that realise them, which can then be broken down into further nested sub-processes, and so on. Thus, according to this view, our mind consists in a continuum of levels of organisation that encompass both the computational and the physical. Moreover, this analysis also includes what we have been calling the personal level:

> To characterise the psychologist's quest in the way I have is to see them as first noting some intentionally or otherwise psychologically characterised abilities of the human subject at the level of data or phenomena, and positing—as theoretical entities—the homunculi or sub-personal agencies that are needed to explain the subject's having those abilities. (Lycan, 1995, p. 40)

Then, we can see that according to Lycan's approach we might distinguish the personal from the subpersonal level by focusing on their place in a functional analysis of the mind. The mental capacities that figure in personal-level explanations are supposed to be situated at the highest level in the functional hierarchy, while subpersonal-level explanations account for the underlying sub-processes that realise those capacities. Given that higher-levels involve a more abstract teleological description than lower ones—which get closer to a structural or implementational levels of analysis—Lycan claims that teleology comes in degrees. Then, the personal-subpersonal distinction can in principle be understood in terms of degrees of teleology. He is skeptical, however, in that a definite line could be drawn between them. These ideas are put clearly in the following passage:

> (i) At least for single organisms, degrees of teleologicalness of characterisation correspond rather nicely to levels of nature. And (ii) there is no single spot *either* on the continuum of teleologicalness or amid the various levels of nature where it is plainly natural to drive a decisive wedge, where descriptions of nature can be split neatly into a well-behaved, purely "structural", purely mechanistic mode and a more abstract and more dubious, intentional, and perhaps vitalistic mode—certainly not any spot that also corresponds to any intuitive distinction between the psychological and the merely chemical, for there is too much and too various biology in between. (Lycan, 1995, p. 45)

We might add that there is also too much computation between personal and subpersonal-level explanations. But even though the relation between levels is one of

continuity rather than sharp division, it is clear that this approach states that there is such a distinction. Namely, the personal level is the most teleological level in the hierarchy, and the one that describes typical mental capacities or functions, such as perception, memory and decision-making.

A problem with this account is how to spell out what is meant by having more teleology. Lycan acknowledges that he leaves this term rather unexplained, however he roughly characterises it in two ways. One is that more teleological means "more abstract" in the sense that it is far from the "grittily concrete" and "purely mechanical" realisation reached at the bottom of the subpersonal level. Secondly, he subscribes to the notion of teleology normally used by philosophers of biology, which in turn understand teleological capacities as biological mechanisms that have certain functions. So, and leaving evolutionary considerations aside, Lycan's notion of teleological capacity is that of abstract mechanisms that are aimed to satisfy functions which are typically ascribed to minded creatures.

But where should we draw the line then? I believe that the right level of abstraction is the one that allows us to formulate a proper personal-level explanation. The main goal of this chapter so far has been precisely to characterise personal-level explanations, and vindicate them in terms of their explanatory advantages and plausible insertion in a metaphysical picture of hierarchical levels of organisation. In the following sections, I attempt to put the personal-subpersonal distinction to work, with the purpose of stating some conditions we would expect computational agents would have to satisfy in order to realise a personal level of organisation.

7.4 The Agent Level: A Proposal Towards Drawing the Line

In this section I attempt to use the personal-subpersonal distinction examined in this chapter to spell out the distinction between (merely) computational- and psychological-level explanations, and in this way present a proposal about how to draw the line between computational agents with and without mentality. The key claim is that we are justified in ascribing mentality to agents who instantiate computation when their

behaviour can be properly explained from a personal-level perspective, which is equivalent to what I have called the psychological level of explanation.

The term personal-level is problematic for present purposes, though, because it is conceptually associated with humans (i.e. persons) whereas our aim is to apply personal-level explanations to non-human animals and even to machines. Looking for a better term, I propose to use the term *agent level* in place of personal level. I believe that this denomination is general enough to encompass what I have called autonomous agents (viz. embodied computing systems capable of self-governing and self-organisation) and that can be tailored to capture the basic features of the personal level. In the remainder of this section, I elaborate the notion of agent level by discussing how it can accommodate the three characteristics of the personal-level put forward in section 7.2, viz. its subject matter, theoretical vocabulary and normative dimension.

7.4.1 Subject-matter

In the way that the objects of study of the personal level are whole-persons, agent-level explanations are about whole-agents, and so it would be a mistake to identify agent-level descriptions with one component or module of a computational agent. Then agent-level explanations have to be formulated with a degree of generality appropriate for being attributable to the agent as a whole and not merely to its parts. This certainly restricts which computational architectures could be candidates for agent-level descriptions, given that not just any complex assembly of multiple computational systems can be properly described as a genuine agent.

In this context, genuine agent-level explanations should be distinguished from "as if" ones. What I mean with the latter are explanations that talk about whole-agent behaviour in a metaphorical fashion, often for pragmatic purposes, without the intention of ascending into a higher (metaphysical) level of explanation. For example, Marr's theory of vision (Marr, 1982) adopts the strategy of first defining a higher-level of description about "*what* the device does and why", concerned with the tasks carried out by the visual system in order to extract information from the environment and the constraints under which it operates. He sees this higher-level description of the task

performed by the visual system as an unavoidable step towards exploring in detail how it is carried out through informational structures and algorithms. It would be mistake, though, to interpret Marr's higher-level description as equivalent to a personal-level approach. As McClamrock (1991) suggests, Marr's approach to the visual system is centred on a computational (i.e. subpersonal) level of analysis, and so this higher-level (which he indeed calls "computational") works more as a means to elucidate the computational details of the system rather than to map onto an autonomous higher-level of description.

The same principle can be applied to the study of artefacts. Suppose an army captures a weapon from its enemy and sends it to a group of scientists for study. It would certainly be useful for them to start figuring out "what the device does and why" before going into its internal machinery. But this should be considered just an heuristic strategy and not to involve the attribution of agency. The machine is not supposed to be capable of doing things in the way of agents, any more than a vending machine is supposed to sell beverages or to give the right change. Therefore, a non-trivial agent-level description is supposed to map onto a particular domain of entities, that satisfy certain constraints that make them accountable as agents. But what are those constraints?

A first constraint can be advanced in the context of the whole-part distinction. It seems natural to say that part of what makes an assembly of multiple computational systems an agent, is that those systems are integrated and function in a coordinated way towards achieving goals that concern the agent as a whole. Recall the example of the digger wasp mentioned in chapter 2. A particular characteristic of this insect is that even though it exhibits sophisticated and flexible navigational capacities, at the same time the way in which it drags food back to its nest is rather rigid and stereotypic. As it seems to be the case of most insects, its computational capacities are highly domain-specific and modular in the sense of not being transferable to other domains. Then it appears that the wasp, after deploying its navigational module for flying and reaching its nest, switches onto a non-computational system to drag the food into the nest. Given that both systems are not integrated, and indeed their processes are couched from different levels of explanation (i.e. the computational and the physical, respectively), it would be odd to describe the transition in the wasp's behaviour from an agent-level perspective. It would

be more appropriate to say that the wasp passed from one subsystem to another, rather than to situate the transition at the same whole-agent level of description. Of course, we could give a general description of the food-gathering tasks performed by the wasp, describing it from a higher-level perspective of an integrated agent. However, as noted above, this talk would be metaphorical, and not revealing of the instantiation of a genuine agent-level in the wasp.

7.4.2 Theoretical Vocabulary[42]

Since with the agent level we intend to map onto the natural domain of mental entities, we have to adopt the theoretical vocabulary of psychology as it figures in personal-level descriptions. As explained in 1.4.2, even the simplest version of psychology is characterised by the use of mentalistic concepts such as beliefs and desires. They conform some of the basic elements for being a thinker, and must interact causally to produce behaviour. It is widely recognised that thoughts are structured by recombinable parts, called concepts. The basic idea is that when an agent thinks about X in different ways, e.g. believing that X is big and believing that X is white, it is entertaining the same concept X. As Fodor (1994) points out—reflecting the standard view of CTM—"concepts are the least complex mental entities that exhibit representational and causal properties; all the others [e.g. beliefs, desires, etc.] ... are assumed to be *complexes* whose constituents are concepts" (p. 96).

Concepts, then, are an integral part of the vocabulary of agent-level explanations. But which animals have concepts? It can be tempting to identify concepts with the informational structures described by computational-level explanations, however, as I have argued throughout this thesis, there is no reason for supposing that all informational structures could fit the bill. Many of them appear to be deeply entrenched at subpersonal stages of processing that do not figure at all in agent-or personal-level explanations. In addition, concept possession also presupposes mastering certain cognitive abilities. For example, pigeons can be trained to sort pictures into categories of tree or person, but these findings do not warrant the conclusion that they have concepts. Pigeons may be just grouping together common visual elements into a

[42] Parts of this section are adapted from my article *Do honeybees have concepts?*

single internal representation, without being able to make further recognitional distinctions and inferences that are characteristic of possessing abstract concepts such as those of a tree or a person (Allen & Hauser, 1996). A common way to frame this idea is by stating that to possess concepts an agent must satisfy the *generality constraint*, which was first formulated by Evans (1982) as follows:

> We cannot avoid thinking of a thought about an individual object x, to the effect that it is F, as the exercise of two separable capacities; one being the capacity to think of x, which could be equally exercised in thoughts about x to the effect that it is G or H; and the other being a conception of what it is to be F, which could be equally exercised in thoughts about other individuals, to the effect that they are F. (p. 75)

The main idea is that genuine thinkers should be capable of producing and entertaining an unbounded set of novel well-formed combinations of concepts. This capacity is closely related with the so-called systematicity and productivity of thought, which have been proclaimed by proponents of the computational theory of mind as elemental features of thought (Rey, 1997, ch. 8). Then it appears that to be couched in agent-level terms the computational capacities of an agent must satisfy the generality constraint.

This immediately appears to be problematic for modular cognitive architectures. As in the mentioned case of the wasp, the computational capacities of many animals, including rats and birds, appear to be massively modular (see Shettleworth, 1998, for a review). For instance, they have been studied in artificial environments that offer limited kinds of information that can be used by them to orientate. In those experiments animals proved to be able, not only to use these different environmental clues to navigate, but to deploy them in a way that requires computation. However, some kinds of information appear to be perceived and used independently, without the capacity to integrate them with other visual clues. All this suggests that they process the various kinds of spatial information by dedicated cognitive modules, that exhibit the hallmarks of domain-specificity, computational processing and isolation[43]. Interestingly, robots built by AI researchers also have modular architectures (Carruthers, 2006, ch. 1). In fact, the Mars rovers I put forward in previous chapters have their total computing systems divided up amongst task-specific modular structures (Cichy, 2010). But if informational structures between modules are not combinable with one another, how

[43] The classic definition of a cognitive module comes from Fodor (1983).

could animals and other computing agents satisfy the generality constraint and be considered genuinely conceptual? In case they cannot, many animals would be out from the scope of agent-level explanations.

Carruthers (2009) has addressed this issue and argued that even though they have a massively modular architecture, most animals are capable of conceptual thought. To justify his claim, he formulates a "weak" version of the generality constraint where all that matters is to be able to make at least *some* combinations between the concepts an agent possesses, while the capacity to make all possible combinations of thoughts constitutes an ideal, he suggests, that perhaps only humans can get close to achieving. Carruthers contends that something like the weak version of the constraint actually appears to be satisfied inside the workings of single modules, as for example in those involved in navigation in honeybees and other insects. For example, imagine that certain navigational module of honeybees has among its domain-specific repertoire of representations those of colours green and yellow, while they also possess a module for flower recognition that besides representations of green and yellow, also incorporates representations of red. So even though a honeybee cannot have thoughts involving elements from both modules, as would be to entertain 'the hive is red' from the navigation module, the fact that within this module it can think 'the hive is green' and 'the hive is yellow' shows the insect has the basic capacity to recombine its concepts, and therefore at least in such a modest way can satisfy the weak generality constraint.

But, should we accept this weak version of the generality constraint? Is it too modest? I believe it is, partly because it is not up to some more fundamental aspects of concept possession the generality constraint is supposed to reflect. Following its original formulation, the generality constraint is intended to ensure that when a creature really has the concept F, we are committed to the view that when it has any thought that deploys this concept (e.g. Fa, Fb, etc.) it is exercising the same conceptual capacity (see Evans, 1982, pp. 101-105). However, this does not seem work with honeybees. Let me explain this with an example.

Recall the previous example of the two modules for flower recognition and for navigation. The honeybee would be able to think 'the flower is yellow' in the first module, while 'the hive is yellow' in the second. Contrary to what the generality

constraint proclaims, the conceptual capacities deployed to think about the concept of YELLOW in both cases are different, thus raising doubts about whether the insect is really able to entertain the concept of YELLOW. It could be argued that both modules share the same conceptual capacities, but the nature of cognitive modules seems to count against this idea. Cognitive modules are often conceived as "mental organs" in analogy with the organs of the body, since they evolved functionally specialised mechanisms in same way as the heart or the lungs (Pinker, 1997). It is a natural consequence of this specialisation that the functions performed by these organs correspond to distinct biological capacities, not recombinable with one another. If cognitive modules are also highly specialised, both in terms of the symbolic structures they can process and in the processing mechanisms (i.e. programming languages) they use, it can be called into question whether they combine their symbolic structures by exercising equivalent processing capacities. Ultimately, it is an empirical matter to determine how compatible the computing mechanisms between modules are. However, the present critique at least shows that this compatibility should not be taken for granted, and that further research is required to determine whether certain massively modular systems could meet the generality constraint required for agent-level processing.

7.4.3 Normative Dimension

Rationality has traditionally been conceived as a rather demanding notion. Many follow the idea that genuinely rational agents entertain their thoughts within a *space of reasons*, which is (roughly) a framework of logical relations that allows us to weight reasons and decide what to do (McDowell, 1994a). These logical relations typically involve constraints of consistency and coherence between thoughts. But as I have argued in 7.2.4, to impose such high standards as a benchmark for judging whether an agent is rational is mistaken, given that even human cognitive capacities normally perform below those standards. It is much more plausible to formulate a resource-relative conception of rationality, relativised to the cognitive capacities and limitations of the agents under scrutiny and allowing that we can be rational without complete consistency and coherence.

There is a danger of a slippery slope here, though. How inconsistent and incoherent can a rational agent be? Clearly there have to be certain margins. As Block (1994) notes, "the attribution of irrational beliefs cannot go on without limit; eventually, one loses one's grip on the content of what one has attributed" (p.111). Indeed, the point relates with the foundations of CTM itself. Part of what makes a computing system capable of performing certain task is that its computational operations have the right inferential structure—i.e. program—situated within certain margins of internal coherence. An adding machine, for example, succeeds as such insofar as its program can preserve the numerical values of its symbols and manipulate them according to adding functions. If minor alterations are made to the program, it might still be possible for the machine to perform its job, however if the alterations accumulate they will reach a point when the machine fails in performing its adding function for any numerical value. As a result, it would not be an adding machine anymore (cf. Haugeland, 1981; Cummins, 1989).

An analogue case can be made for the mind. Its capacity to think depends on its having the right inferential structure, which can perform rational operations insofar as its structure is kept within certain margins of internal coherence. A computing system that can no longer carry out the cognitive functions that characterise the mind (e.g. computing symbols standing for conclusions from symbols standing from premisses), would then cease to possess mentality. And with respect to the opposite end of the spectrum, a computational mechanism might be fully logical (e.g. consistent) in the sense of manipulating symbolic structures in conformity to logical rules, however be merely computational—i.e. lack mentality. In other words, not all autonomous agents capable of performing inferences are capable of reasoning (see chapter 2).

In any case it seems plausible to assume that rationality could come in degrees, and that it largely depends on the possession of inferential-computational capacities. What remain as open questions are the extent of internal coherence and the cognitive architecture that could make genuine rationality possible. Hurley (2003)[44] has formulated a proposal that deals with these issues. She claims that it can be possible for some animals to have rational capacities that are context-bound, in the sense that they are not transferable from one task to another, however these capacities must satisfy

[44] Unless otherwise noted, all page quotations in this section correspond to this paper.

some minimum requirements such as holism and normativity. With holism Hurley means an integrated network of relations between perceptual and central cognitive states, such that the animal has available certain space of representations of possible means to attain its goals. What makes this network context-bounded is that it could be realised in domain-specific cognitive systems, which would still count as holistic because, at least within the domain-specific system, means and ends are related in a flexible and transferable way, for example allowing the animal to alternate different means to satisfy certain desires. In the words of the author, they form "islands of practical rationality", bounded to specific cognitive domains that do not generalise (p. 238).

After arguing that this sort of holism is sufficiently general to support reasoning processes, Hurley goes on to account for how normativity could come about in domain-specific cognitive systems. She sees as a defining feature of normativity the "possibility of mistake", in the sense that there must be a way of judging certain processes and courses of action as right or wrong according to standards of rationality. At this point, she appeals to complexity and teleology:

> Normativity admits of different kinds and levels and degrees. But the kind of mistake possible for a relatively complex and flexible, teleologically embedded system seems to me adequate to meet the normativity condition for correctly attributing practical reasons to an intentional agent ... (p. 244)

With the requirement of complexity the author means processing (or we might say computational) sophistication, such as the capacity of yielding flexible responses and exerting feedback control over its own processes. But as the author acknowledges (and concordantly with my own proposal), complexity of this sort can be possibly found in certain machines (e.g. robots) to which we should not ascribe rationality, and so even though complexity appears to be important it is not itself sufficient for the sort of normativity at stake here. As mentioned in the quote, the next requirement is teleology. The kind of teleology Hurley has in mind corresponds to an etiological notion of function, which defines functions in terms of their history (typically how they evolved by natural selection).

I believe that this move to teleology weakens Hurley's proposal. As mentioned in section 2.3.2, an etiological notion of function is problematic, at least for present

purposes, since it would rule out the possibility that a human-built robot could ever be endowed with mentality (or normativity, in the present context). And even if we understand cognition in teleological terms, the most plausible approach to functions according to the present naturalistic framework would be dispositional instead of etiological, with functions being defined in terms of their current roles in our best scientific theories of how the mind works (cf. 5.5.1 where I discuss further problems that emerge from situating etiological and dispositional functions within the same account of the mind). However, by adopting a dispositional approach to functions nothing would prohibit us from ascribing teleology to subpersonal-level processes, as for instance to the face-detection mechanisms run within a cognitive module. Therefore what is at issue does not seem to be whether there is normativity, but what could make certain normative processes genuine cases of reasoning. So, contrary to Hurley's view, I suggest that it is not teleology but computational sophistication that can give us the key for understanding what sort of computational architecture might possibly realise cognitive capacities accountable from an agent-level perspective (see below).

An interesting thing to note regarding Hurley's view is that she recognises that we are only justified in attributing rationality to an animal when that attribution is formulated from an agent-level perspective (which she calls the "animal-level"). In other words, if an agent is described as acting for reasons, they must correspond to the agent's own reasons, whereas to ascribe reasons to its subsystems would lead us nowhere. But when it comes to clarifying what requirements computational processes would have to meet in order to be attributable to an agent as a whole, Hurley's proposal is disappointing. Apart from appealing to certain degree of holism (see above), she argues that what a cognitive process requires to qualify as reasons for an animal is that they explain action in the context set by "at least primitive forms of practical rationality" (p. 233). Therefore, she situates having reasons as a precondition for being accountable from an agent-level perspective, leaving unexplained what is required for the agent-level in the first place. I propose, in contrast, that it is the capacity to realise an agent-level cognitive architecture that makes possible the emergence of rationality, and not the other way round.

Let me conclude this discussion of how the normative dimension could be adapted to an agent-level approach by focusing on cognitive architecture. A key point is

whether modular—"island-like"—domain-specific computational systems could be accountable for reasoning. One might answer affirmatively to this, by appeal to the computational sophistication of some of these modular systems. Note that animals can be very smart in certain specific domains. In addition to the examples of insect navigation given elsewhere in this thesis, another could be the scrub-jays's capacities of encoding and retrieving temporal information about caching. These birds can not only retrieve information about where and when they stored food, but also selectively retrieve food items depending on their decaying time (that the birds had previous learned) and the time elapsed since they were stored (Clayton & Dickinson, 1998). Cases like these sometimes match and even outstrip human cognitive capacities. But the same could be said of artefacts such as chess-playing computers, or the automatic pilots of some airplanes. Something all these computational systems have in common is that they are highly task-specific, and there operations cannot generalise to new situations or tasks. This makes controversial that the behaviours they trigger are attributable to the whole-agent, instead of their modular subsystems, making them poor candidates for being the locus of the agent's reasons.

To see further why domain-specific computation is problematic to account for reasoning recall that both animals' and computers' inner architectures typically consist in multiple modules. This opens the possibility that their operations come into conflict, in the sense of leading to inconsistent sets of conclusions. Then it becomes controversial to consider each of these modular operations as contributing with reasons that serve the goals of a whole individual, since such a disunited set of operations and goals can hardly support the ascription of instrumental rationality to a whole agent (Saunders & Over, 2009). Indeed, modular processes appear to make more sense from an evolutionary perspective, in particular as serving the biological goals of survival and proliferation of our genes, rather than as the actual reasons of an agent. Stanovich and West (2000) make the same point by appeal to the distinction between evolutionary adaptation and instrumental rationality. According to them,

> The key point is that for the latter (variously termed practical, pragmatic, or means/ends rationality), maximization is at the level of the individual person. Adaptive optimisation in the former case is at the level of the genes. In Dawkins's (1976; 1991) terms, evolutionary adaptation concerns optimization processes relevant to the so-called replicators (the genes),

> whereas instrumental rationality concerns utility maximization for the so-called vehicle ... which houses the genes. (p. 660)

Stanovich and West propose that what makes plausible the ascription of rationality to an agent is its possession of a domain-general processing system capable of overriding domain-specific predispositions and integrating them, pursuing the interests of the agent taken as a whole. This domain-general system constitutes a level of abstraction that makes it possible to carry out inferences about domain-specific mechanisms in a causal-logical fashion, in a way attributable to the whole agent. This metacognitive ability thus appears to be required for yielding genuine instrumental reasoning, and interestingly, according to a recent review (Penn & Povinelli, 2007) there is no conclusive evidence that non-human animals are capable of such a metacognitive capacity[45].

7.5. Conclusions

I have argued that the personal-subpersonal distinction offers a plausible framework for determining what makes certain computational systems capable of mentality. After exploring how the main aspects of agent-level explanations could be realised in a computational architecture, it becomes apparent that the notion of agent-level points in the direction of cognitive access, generality and integration. These constraints should be operative in explanations involving rational links between perception, central-cognition and actions systems, as well as those processes dealing with multiple modular structures. Thus, animals having a massively modular computational architecture (as is presumably the case of arthropods, for example) should be considered poor candidates for being accountable from an agent-level perspective, and so it would be unjustified to ascribe mentality to them. However, to further explore what kind of computational architecture would make it for an agent-level description goes beyond the possibilities of this thesis.

---

[45] This metacognitive ability is closely related with a *dual-process cognitive architecture*, which is claimed to be have evolved much more recently than modular—domain-specific—architectures, and thought by most theorists to be uniquely human (see Evans, 2003, for a review). To go into the details of the dual-process approach to reasoning goes beyond the scope of this chapter, though.

My purpose in this chapter has been to make a comprehensive characterisation of the personal-subpersonal distinction, and to adapt this distinction to the case of non-human entities through the notion of agent-level. I have articulated a certain philosophical perspective on this notion, and however I used examples from animal cognition and artificial intelligence as support to my arguments, it has only been my purpose to illustrate an important distinction that could be adopted to draw the line between computational agents with and without mentality. Further revision of ethological data and empirical work may well be needed to determine whether agent-level descriptions could be applied to explain the behaviour of non-human animals or actual robots.

Concluding remarks

The aim of this thesis has been to explore the minimum conditions computational agents have to meet in order to possess a mind, and to put forward an adequacy criterion for drawing the line between agents with and without mentality. After contextualising the debate within the framework of naturalism and the computational theory of the mind, in chapter 2 I argued that computational explanations of behaviour do not entail mentality. On the contrary, explanations couched from the computational level are appropriate for non-mental computers and thus capture a natural domain that is distinct from—however overlapping with—the domain of mental agents. A consequence of this idea is that we have to count the computational level as an alternative, autonomous level of explanation, on pain of falling into a false dilemma in case of considering psychology as the only alternative to proper physical-level explanations of behaviour. I then undertake an exhaustive critical survey of the existing proposals about the minimum conditions computational agents have to meet in order to be explainable from the psychological level.

Chapters 3 & 4 tackled the informational approaches of Fodor and Dretske. They are part of the standard account of how computational agents pick up and transform information into a structure suitable for thought and reasoning. In particular, Fodor attempts to specify the informational relations that symbolic structures must have with their referents. I argued that his view was poorly revealing of the minimum conditions under which those relations hold, and therefore was not very useful for the present purposes. Dretske, in turn, proposes that the key for understanding what makes certain computing systems capable of mentality lies in the way information from the world is coded into a symbolic structure. I contended that even though this coding process is important, it does not suffice as a minimum condition for possessing mental symbols. Overall, I conclude that the capacity to develop symbolic structures bearing informational relations with the environment (of the sort described by Fodor and Dretske) does not seem to be something exclusive to mental agents. What these informational approaches describe as constitutive of mentality appears to be just part of

the computational processes that enable a system to possess mental symbols, but not the whole story.

Chapter 5 discussed the view of Burge, who adds teleology in his account of perceptual systems. But contrary to what he proposes, I argued that the capacity of certain information-gathering systems to accurately detect distal environmental properties is not necessarily a psychological (i.e. genuinely perceptual) capacity. Even if characterised in teleological terms, these detector-functions might still be ascribed to a merely computational agent. Consequently, Burge's view becomes susceptible to the same counterexamples as the informational approaches mentioned above. Though centred on perception, Burge also attempts to give a more holistic flavour to his approach by characterising perceptual functions as serving the purposes of a whole agent. But despite being on the right track, his proposal fails due to complications with his overall picture of dispositional and etiological functions merging within the same organism.

Chapter 6 addressed two distinct proposals about minimal forms of psychological explanation. Bermúdez attempts to do without the standard—inferential—model of psychological explanation and puts forward a version of success semantics that defines mental symbols in operational terms. But besides inheriting problems traditionally associated with operational accounts (such as behaviourism), his view fails to provide the causal mechanisms required for explaining complex animal behaviour. Carruthers, on his side, follows a more traditional version of psychological explanation and proposes a cognitive architecture that captures what he sees as the core of mentality. I believe Carruthers is right in situating the burden of psychological explanation within a whole-agent perspective, however I argue that his view is unsatisfactory as an account of what is paradigmatic of mentality, or is at least incomplete in comparison with what I characterise as a personal-level approach to psychological explanation. Indeed, Carruthers' core cognitive architecture also becomes an easy target for the counterexamples of mindless machines given in several parts of this thesis.

The final chapter developed a proposal towards an adequacy criterion for drawing the line between agents with and without mentality. My strategy consisted in identifying the contrast between psychological- and computational-level explanations

with the distinction—well entrenched in the philosophical literature—between personal and subpersonal levels of explanation of the mind. I spelled out what a personal-level approach consists of, explaining that it takes whole agents as their subject matter, uses a distinctive theoretical vocabulary, and is constrained by norms of rationality. I argued that this approach is compatible with a naturalistic framework, and that it provides an especially satisfying way of identifying the paradigmatic aspects of psychological explanation.

My proposal can be understood as a top-down approach, which takes psychological explanations of human behaviour as the paradigm for judging whether other computational agents have minds. In order to avoid anthropomorphic concerns related with defining the mind in terms of persons, I developed an agent level of description that focuses just on the essential aspects of the personal-level approach, so as to adapt it for the purposes of explaining the behaviour of animals and even machines. By means of this agent-level approach I attempted to explore the minimum conditions a computing system requires for possessing a mind. Paramount among those conditions is the possession of symbolic structures that function in a coordinated way towards achieving goals that concern the agent as a whole, that satisfy the generality constraint, and that take part in metacognitive processes that integrate different domain-specific areas of the agent.

As any account committed to scientific realism, my proposal is a blend of epistemology and metaphysics. It assumes that it is through our best psychological theories that we can gain (observer-independent) knowledge about the nature of the mind. More precisely, I contend that the agent-level approach provides the most convenient way to formulate psychological explanations and, therefore, that it can reveal to us the main features and constraints a computational agent must comply to possess mentality. Admittedly, this thesis presents a first pass through those constraints. However, my purpose has been to show that the agent-level approach has advantages over other criteria offered by the literature, and I have given concrete examples of the sort of computational architecture a mental agent is supposed to have. A further task would be to explore in more detail the particular computational architectures of the different animal species in the evolutionary tree of life, and figure out which of them appear to be the most basic ones that could be the target of agent-level explanations.

Bibliography

Adams, F. (2003). The informational turn in philosophy. *Minds and Machines*, *13*, 471–501.

Aguilera, B. (2011). Do honeybees have concepts? *Disputatio*, *4*(30), 59–71.

Aguilera, B. (2013). Is perception representational? Tyler Burge on perceptual functions. In P. Hanna (Ed.), *An anthology of philosophical studies vol. 7* (pp. 59–71). Athens: ATINER.

Allen, C. (2009). Teleological notions in biology. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from <http://plato.stanford.edu/archives/win2009/entries/teleology-biology/>

Allen, C., & Bekoff, M. (1997). *Species of mind: The philosophy and biology of cognitive ethology*. Cambridge, MA: MIT Press.

Allen, C., & Hauser, M. D. (1996). Concept attribution in nonhuman animals: Theoretical and methodological problems in ascribing complex mental processes. In M. Bekoff & D. Jamieson (Eds.), *Readings in animal cognition* (pp. 47–62). Cambridge MA: MIT Press.

Antony, L. M., & Levine, J. (1991). The nomic and the robust. *Meaning in mind: Fodor and its critics* (pp. 1–16). Oxford: Basil Blackwell.

Bechtel, W., Graham, G., & Abrahamsen, A. (1998). The life of cognitive science. In W. Bechtel & G. Graham (Eds.), *A companion to cognitive science* (pp. 1–104). Oxford: Basil Blackwell.

Bermúdez, J L. (2003). *Thinking without words*. Oxford: Oxford University Press.

Bermúdez, José Luis. (1995). Nonconceptual content: From perceptual experience to subpersonal computational states. *Mind and Language*, *9*, 333–369.

Bermúdez, José Luis. (2005). *Philosophy of psychology: A contemporary introduction*. New York: Routledge.

Block, N. (1994). Advertisement for a semantics for psychology. In S. Stich & T. Warfield (Eds.), *Mental representation: A reader* (Vol. X, pp. 81–141). Cambridge: Blackwell.

Boden, M. (2001). Life and cognition. In J. Branquinho (Ed.), *The foundations of cognitive science* (pp. 11–22). Oxford: Clarendon Press.

Botterill, G., & Carruthers, P. (1999). *The philosophy of psychology*. Cambridge: Cambridge University Press.

Braddon-Mitchell, D., & Jackson, F. (1996). *Philosophy of mind and cognition*. Oxford: Blackwell.

Brooks, R. (1991). New approaches to robotics. *Science*, *253*, 1227–1232.

Burge, T. (2010). *Origins of objectivity*. New York: Oxford University Press.

Byrne, J. (1990). Learning and memory in Aplysia and other invertebrates. In R. P. Kesner & D. S. Olton (Eds.), *Neurobiology of comparative cognition* (pp. 293–315). New Jersey: Lawrence Erlbaum.

Carruthers, P. (2004a). On being simple minded. *American Philosophical Quaterly*, *41* (3), 205–220.

Carruthers, P. (2004b). Review of Bermudez's Thinking without words. *British Journal for the Philosophy of Science*, *55*(4), 807–810.

Carruthers, P. (2006). *The architecture of the mind*. New York: Oxford University Press.

Carruthers, P. (2009). Invertebrate concepts confront the generality constraint (and win). In R. Lurz (Ed.), *The philosophy of animal minds* (pp. 89–107). Cambridge: Cambridge University Press.

Chalmers, D. J. (1996). Does a rock implement every finite-state automaton? *Synthese*, *108*(3), 309–333.

Chater, N., & Heyes, C. (1994). Animal concepts: Content and discontent. *Mind & language*, *9*(3), 209–246.

Cherniak, C. (1986). *Minimal rationality*. Cambridge MA: MIT Press.

Chomsky, N. (1959). A review of B. F. Skinner's Verbal behavior. *Language*, *351*, 26–58.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: MIT Press.

Churchland, P. M. (1979). *Scientific realism and the plasticity of mind*. New York: Cambridge University Press.

Churchland, P., & Sejnowski, T. (1999). *The computational brain*. Cambridge MA: MIT Press.

Cichy, B. (2010). Keynote talk at the 2010 workshop on spacecraft flight software (FSW-10). Retrieved from http://win-dms-ms1.caltech.edu/five/Viewer/?peid=476727664f1b4d8390d3ab37670ababd

Clayton, N. S., & Dickinson, A. (1998). Episodic-like memory during cache. *Nature*, *395*(6699), 272–273.

Copeland, J. (1993). *Artificial intelligence: A philosophical introduction*. Oxford: Blackwell.

Crane, T. (1995). *The mechanical mind: A philosophical introduction to minds, machines and mental representation*. Harmondsworth: Penguin Books.

Crane, T. (1999). The autonomy of psychology. In R. A. Wilson & F. C. Keil (Eds.), *The MIT encyclopedia of the cognitive sciences* (pp. 64–65). Cambridge MA: MIT Press.

Crane, T. (2001). *Elements of mind*. New York: Oxford University Press.

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

Cummins, R. (1983). *The nature of psychological explanation*. Cambridge, MA: MIT Press.

Cummins, R. (1989). *Meaning and mental representation*. London: MIT Press.

Davidson, D. (1980). *Essays on action and events*. Oxford: Clarendon.

Davidson, D. (1984). *Inquiries into truth and interpretation*. Oxford: Oxford University Press.

Davies, M. (2000). Persons and their underpinning. *Philosophical Explorations*, *3*(1), 43–62.

Day, T., & Kincaid, H. (1994). Putting inference to the best explanation in its place. *Synthese*, *98*, 271–295.

Dennett, D. (1982). Beyond belief. In A. Woodfield (Ed.), *Thought and object* (pp. 1–95). Oxford: Clarendon Press.

Dennett, Daniel. (1969). *Content and consciousness*. London: Routledge.

Dennett, Daniel. (1979). *Brainstorms*. London: Penguin Books.

Dennett, Daniel. (1984). *Elbow room: The varieties of free will worth wanting*. Cambridge MA: MIT Press.

Dennett, Daniel. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Dennett, Daniel. (1991). Real patterns. *Journal of Philosophy*, *LXXXVII*, 27–51.

Dennett, Daniel. (1994). Dennett, Daniel C. In S. Guttenplan (Ed.), *A companion to the philosophy of mind* (pp. 236–244). Oxford: Blackwell.

Devitt, M. (1991). Why Fodor can't have it both ways. In B. Loewer & G. Rey (Eds.), *Meaning in mind: Fodor and his critics* (pp. 95–118). Cambridge: Blackwell.

Dretske, F. (1980). The intentionality of cognitive states. *Midwest Studies in Philosophy*, *5*(1), 281–294.

Dretske, F. (1981). *Knowledge and the flow of information*. Cambridge, MA: MIT Press.

Dretske, F. (1986). Misrepresentation. In R. Bogdan (Ed.), *Belief: Form, content, and function* (pp. 17–36). Oxford: Oxford University Press.

Dretske, F. (1988). *Explaining behavior: Reasons in a world of causes*. Cambridge, MA: MIT Press.

Dretske, F. (1994). Dretske, Fred. In S Guttenplan (Ed.), *A companion to the philosophy of mind* (pp. 259–264). Oxford: Blackwell.

Dretske, F. (1999). Machines, plants and animals: The origins of agency. *Erkenntnis*, *51*, 19–31.

Egan, F. (1995). Computation and content. *The Philosophical Review*, *104*(2), 181–203.

Evans, G. (1982). *The varieties of reference*. Oxford: Oxford University Press.

Evans, J. S. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*(10), 454–459.

Fine, A. (1999). Realism and antirealism. In R. A. Wilson & F. C. Keil (Eds.), *The MIT encyclopedia of the cognitive sciences* (pp. 707–709). Cambridge MA: MIT Press.

Fitzpatrick, S. (2008). Doing away with Morgan's canon. *Mind & Language*, *23*(2), 224–246.

Floridi, L. (2011). Semantic conceptions of information. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from <http://plato.stanford.edu/archives/spr2011/entries/information-semantic/>

Fodor, J, & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*, 3–71.

Fodor, J. (1968). *Psychological explanation*. New York: Random House.

Fodor, J. (1974). Special sciences. *Synthese*, *28*(2), 97–115.

Fodor, J. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.

Fodor, J. (1980). Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and Brain Sciences*, *3*, 63–109.

Fodor, J. (1983). *The modularity of mind: An essay in faculty psychology*. Cambridge, MA: MIT Press.

Fodor, J. (1986). Why paramecia don't have mental representations. *Midwest Studies in Philosophy*, *10*(1), 3–23.

Fodor, J. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. London: Bradford.

Fodor, J. (1990). *A theory of content and other essays*. Cambridge MA: MIT Press.

Fodor, J. (1991). Replies. In Barry Loewer & G. Rey (Eds.), *Meaning in mind: Fodor and its critics* (pp. 255–319). Cambridge: Blackwell.

Fodor, J. (1994). Concepts: A potboiler. *Cognition*, *50*(1-3), 95–113.

Fodor, J. (1998). *Concepts: Where cognitive science went wrong*. Oxford: Clarendon Press.

Fodor, J. (2000). *The mind doesn't work that way*. Cambridge, MA: MIT Press.

Fodor, J. (2003). Review of Bermúdez's "Thinking without words". *London Review of Books*, *25*(19), 16–17.

Fodor, J., & Pylyshyn, Z. W. (1981). How direct is visual perception: Some reflections on Gibson's "ecological approach". *Cognition*, *9*, 139–96.

Fodor, Jerry, & Lepore, E. (1995). Is intentional ascription intrinsically normative? In B. Dahlbom (Ed.), *Dennett and his critics: demystifying the mind* (pp. 70–82). Oxford: Blackwell.

Franklin, S., & Graesser, A. (1996). Is it an agent, or just a program?: A taxonomy for autonomous agents. *Third International Workshop on Agent Theories, Architectures, and Languages* (pp. 21–35). Springer-Verlag.

Gallistel, C R. (1990). *The organization of learning*. Cambridge, MA: Bradford Books/ MIT Press.

Gallistel, C R. (2009). The foundational abstractions. In M. Piattelli-Palmirini, J. Uriagereka, & P. Salaburu (Eds.), *Of minds and language: A dialogue with Noam Chomsky in the Basque Country* (pp. 58–73). New York: Oxford University Press.

Gallistel, C R, & Cramer, A. E. (1996). Computations on metric maps in mammals: Getting oriented and choosing a multi-destination route. *The Journal of experimental biology*, *199*, 211–217.

Gallistel, C.R., & Gibbon, J. (2001). Computational versus associative models of simple conditioning. *Current Directions in Psychological Science*, *10*(4), 146–150.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Glymour, C. (1999). Realism and the nature of theories. In M. H. Salmon, C. Earlman, C. Glymour, J. G. Lennox, J. Machamer, J. McGuire, J. Norton, et al. (Eds.), *Philosophy of science* (pp. 104–131). Indianapolis: Hackett Pub Co Inc.

Godfrey-Smith, P. (1992). Indication and adaptation. *Synthese*, *92*(2), 283–312.

Godfrey-smith, P. (1993). Functions: Consensus without unity. *Pacific Philosophical Quaterly*, *74*, 196–208.

Grice, P. (1957). Meaning. *Philosophical Review*, *66*, 377–88.

Haselageer, W. F. G. (1997). *Cognitive science and folk psychology: The right frame of mind*. London: Sage.

Haugeland, J. (1981). Semantic engines: An introduction to mind design. In J. Haugeland (Ed.), *Mind design: philosophy, psychology, and artificial intelligence* (pp. 1–34). Cambridge, MA: MIT Press.

Haugeland, J. (2003). Syntax, semantics, physics. In J. M. Preston & M. A. Bishop (Eds.), *Views into the Chinese room: New essays on Searle and artificial intelligence* (pp. 379–392). New York: Oxford University Press.

Hawkins, R. D., & Kandel, E. R. (1984). Is there a cell-biological alphabet for simple forms of learning? *Psychological Review*, *91*(3), 375–91.

Healy, S. D. (1998). *Spatial representation in animals*. Oxford: Oxford University Press.

Horgan, T. (1993). Nonreductive Materialism and the Explanatory Autonomy of Psychology. In S. Wagner & R. Warner (Eds.), *Naturalism: A Critical Appraisal* (pp. 295–320). University of Notre Dame Press.

Horst, S. (2009). The computational theory of mind. In E. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from http://plato.stanford.edu/archives/win2009/entries/computational-mind

Hurley, S. (2003). Animal action in the space of reasons. *Mind & Language*, *18*(3), 231–256.

Jamieson, D., & Beckoff, M. (1996). On aims and methods of cognitive ethology. *Readings in animal cognition* (pp. 65–78). Cambridge MA: MIT Press.

Kim, J. (1982). Psychophysical supervenience. *Philosophical Studies*, *41*, 51–70.

Kim, J. (1988). *Mind in a Physical World*. Cambridge, MA: MIT Press.

Kim, J. (2006). *Philosophy of mind*. Cambridge, MA: Westview.

Koons, R. C. (1998). Teleology as higher-order causation: A situation-theoretic account. *Minds and Machines*, (8), 559–585.

Levin, J. (2010). Functionalism. In E Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from <http://plato.stanford.edu/archives/sum2010/entries/functionalism/>

Lipton, P. (2004). *Inference to the best explanation*. London: Routledge.

Loewer, B, & Rey, G. (1991). Editor's introduction. *Meaning in mind. Fodor and his critics* (pp. xi–xxxvii). Oxford: Basil Blackwell.

Lycan, W. (1995). *Consciousness*. Cambridge MA: MIT Press.

Margolis, E., & Laurence, S. (2007). The ontology of concepts—Abstract objects or mental representations? *Noûs*, *41*(4), 561–593.

Marler, P., & Hamilton, W. J. (1966). *Mechanisms of animal behavior*. New York: Wiley.

Marr, D. (1982). *Vision*. New York: W. H. Freeman.

McClamrock, R. (1991). Marr's three levels: A re-evaluation. *Minds and Machines*, *1*, 185–196.

McCulloch, W., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology*, *5*, 115–133.

McDowell, B. (1985). Functionalism and anomalous monism. In E. Lepore & B. McLaughlin (Eds.), *Actions and events: Perspectives on the philosophy of Donald Davidson* (pp. 387–398). Oxford: Basil Blackwell.

McDowell, J. (1994a). *Mind and world*. Harvard: Harvard University Press.

McDowell, J. (1994b). The content of perceptual experience. *The Philosophical Quaterly*, *44*(175), 190–205.

Menzel, R., Brandt, R., Gumbert, A., Komischke, B., & Kunze, J. (2000). Two spatial memories for honeybee navigation. *Proceedings. Biological sciences / The Royal Society*, *267*(1447), 961–968.

Menzel, R., Greggers, U., Smith, A., Berger, S., Brandt, R., Brunke, S., Bundrock, G., et al. (2005). Honey bees navigate according to a map-like spatial memory. *Proceedings of the National Academy of Sciences USA*, *102*, 3040–3045.

Menzel, R., Lehmann, K., Manz, G., & Fuchs, J. (2012). Vector integration and novel shortcutting in honeybee navigation. *Apidologie*, *43*(3), 229–243.

Miller, G. a. (2003). The cognitive revolution: A historical perspective. *Trends in Cognitive Sciences*, *7*(3), 141–144.

Millikan, R. (1984). *Language, thought, and other biological categories*. Cambridge: MIT Press.

Millikan, R. G. (1989). Biosemantics. *The Journal of Philosophy*, *86*, 281–297.

Neander, K. (1995). Misrepresenting & malfunctioning. *Philosophical Studies*, (79), 109–141.

Newel, A., & Simon, H. A. (1981). Computer science as empirical enquiry: Symbols and search. *Mind design: philosophy, psychology, and artificial intelligence* (pp. 35–66). Cambridge MA: MIT Press.

Newell, A. (1960). Physical symbol systems. *Cognitive Science*, *183*, 135–183.

Papineau, D. (1987). *Reality and representation*. Oxford: Blackwell.

Papineau, D. (2009). Naturalism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from <http://plato.stanford.edu/archives/spr2009/entries/naturalism/>

Penn, D. C., & Povinelli, D. J. (2007). Causal cognition in human and nonhuman animals: A comparative, critical review. *Annual review of psychology*, *58*, 97–118. doi:10.1146/annurev.psych.58.110405.085555

Pinker, S. (1997). *How the mind works*. New York: Norton.

Preston, B. (1998). Why is a wing like a spoon? A pluralist theory of function. *The Journal of Philosophy*, *5*, 215–254.

Price, C. (2001). *Functions in mind: A theory of intentional content*. Oxford: Oxford University Press.

Proust, J. (1999). Mind, space and objectivity in non-human animals. *Erkenntnis*, *51*, 41–58.

Putnam, H. (1975). Minds and machines. *Mind, language and reality - Philosophical papers vol 2* (pp. 362–385). Cambridge: Cambridge University Press.

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

Pylyshyn, Z. W. (1980). Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences*, *3*, 111–132.

Pylyshyn, Z. W. (1984). *Computation and cognition*. Cambridge MA: MIT Press.

Rey, G. (1997). *Contemporary philosophy of mind: A contentiously classical approach*. Oxford: Blackwell.

Rives, B. (2010). Jerry Fodor. *Internet encyclopedia of philosophy*. Retrieved from http://www.iep.utm.edu/fodor/

Rosch, E. (1978). Principles of categorisation. *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Salmon, W. C. (1989). Four decades of scientific explanation. In P. Kitcher & C. Salmon, Wesley (Eds.), *Scientific explanation* (pp. 3–219). Minneapolis: University of Minnesota Press.

Samuels, R., Stich, S., & Faucher, L. (2004). Reason and rationality. In I. Niiniluoto, M. Sintonen, & J. Wolenski (Eds.), *Handbook of epistemology* (Vol. 9, pp. 1–50). Dordrecht: Kluwer.

Saunders, C., & Over, David, E. (2009). In two minds about rationality? In J. S. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 317–334). Oxford: Oxford University Press.

Searle, J. (1990). Author's response. *Behavioral and Brain Sciences*, *3*, 450–456.

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

Shettleworth, S. (1998). *Cognition, evolution, and behavior*. New York: Oxford University Press.

Shettleworth, S. J. (2010). Clever animals and killjoy explanations in comparative psychology. *Trends in Cognitive Sciences*, *14*(11), 477–481.

Smithers, T. (1997). Autonomy in robots and other agents. *Brain and cognition*, *34*(1), 88–106.

Stalnaker, R. (1984). *Inquiry*. Cambridge MA: MIT Press.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *The Behavioral and Brain Sciences*, *7*(5), 645–65.

Sterelny, K. (1990). *The representational theory of mind: An introduction*. Oxford: Basil Blackwell.

Stich, S. (1981). Dennett on intentional systems. *Philosophical Topics*, *12*(1), 39–62.

Stich, S. (1983). *From folk psychology to cognitive science*. Cambridge MA: MIT Press.

Stoering, P. (1999). Blindsight. In R. A. Wilson & F. Keil (Eds.), *The MIT encyclopedia of the cognitive sciences* (pp. 88–90). Cambridge MA: MIT Press.

Swerdlow, N. R., Caine, S. B., Braff, D. L., & Geyer, M. a. (1992). The neural substrates of sensorimotor gating of the startle reflex: A review of recent findings and their implications. *Journal of Psychopharmacology*, *6*(2), 176–190.

Tetzlaff, M., & Rey, G. (2009). Systematicity and intentional realism in honeybee navigation. In R. Lurz (Ed.), *The philosophy of animal minds* (pp. 72–88). Cambridge: Cambridge University Press.

Thomas, E. a, Sjövall, H., & Bornstein, J. C. (2004). Computational model of the migrating motor complex of the small intestine. *American journal of physiology. Gastrointestinal and liver physiology*, *286*(4), G564–72.

Touretzky, D. S., & Saksida, L. M. (1997). Operant conditioning in Skinnerbots. *Adaptive Behavior*, *5*(3), 219–247.

Turing, A. (1937). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *s2-43*, 544–546.

Turing, A. (1950). Computing machinery and intelligence. *Mind*, *59*, 433–60.

Warfield, T. A., & Stich, S. (1994). Introduction. *Mental representation: A reader* (pp. 3–8). Cambridge: Blackwell.

Weng, J. (2004). Developmental robotics: Theory and experiments. *International Journal of Humanoid Robotics*, *1*(2), 199–236.

Wood, J. D. (2011). *Enteric nervous system (the brain-in-the-gut)*. Princeton: Morgan & Claypool Life Sciences Series.

Wooldridge, D. (1971). *The machinery of the brain*. New York: McGraw-Hill.

Wright, L. (1973). Functions. *The Philosophical Review*, *82*, 139–168.