# The Threshold of Self-Consciousness

Stephane Joseph Savanah

BSc(Hons), DipEd, PGDipDataProc, PGDipPhil

ARC Centre of Excellence for Cognition and its Disorders
Department of Cognitive Science
Macquarie University
Sydney, Australia

June 2012

# Table of Contents

## Part 2: Empirical Studies of Self-Consciousness

# Preface

This thesis is about self-consciousness and how we might be able to determine its existence in non-human animals and human infants. By 'self-consciousness' I mean something very like the type of self-consciousness possessed by normal human adults. I examine the nature of self-consciousness, explore the connection between self-consciousness and concept possession, and review research into animal and infant self-consciousness. I conclude that there are ways to determine the existence of self-consciousness in animals based on observations of their behaviour, and that sufficient evidence exists to conclusively ascribe self-consciousness to chimpanzees. Furthermore, there are strong indications that self-consciousness is probably possessed by dolphins, elephants and some corvid species such as magpies and scrub jays.

This thesis is divided into two main parts. Part 1 (chapters 1-4) is mostly theoretical. In part 1 I discuss the nature of self-consciousness and how we can tell it is possessed by an organism. Part 2 (chapters 5-8) applies this analysis in the evaluation of various research paradigms on self-consciousness in animals and human infants. I conclude the thesis with chapter 9, in which I summarise the main arguments and conclusions presented and offer some thoughts about future research.

In chapter 1 I define and defend my conception self-consciousness, which I encapsulate as *an understanding of one's own existence as a psychological subject with intentional agency*. I also briefly review several research paradigms and foreshadow the conclusions reached in part 2. In chapter 2 I explore some central issues in the philosophy of self-consciousness and find a common thread, a *Fundamental Dichotomy* between *relationalism,* which sees self-consciousness as always involving a relation between a subject and a mental state, and *intrinsicism*, which regards self-consciousness as immediate and unmediated. Relationalism is the correct position for a self-concept while intrinsicism holds only for non-conceptual self-access. This position suggests the hypothesis that concept possession alone is sufficient for self-consciousness. I explain and defend this hypothesis in chapter 3 and suggest that it provides a yardstick for gauging the validity of research into self-consciousness. In chapter 4 I discuss ways in which concept possession might be determined: propositional thinking, rationality and symbol-mindedness are all indicators of concept possession. These are difficult to conclusively determine since I advocate that we must keep the standard of

evidence high. Nevertheless, in a few studies there is good reason to believe that the standard has been met, as discussed in Part 2.

The Concept Possession Hypothesis (CPH) may be considered controversial by some readers. Nevertheless, I do not rely on it exclusively in part 2 and readers who remain unconvinced by the hypothesis will still find much of interest in part 2. In chapter 5 I use CPH to argue that chimpanzees are self-conscious on the basis of their being demonstrably concept possessing. An interim conclusion is that chimpanzees are symbol-minded, which is significant in its own right. In chapter 6 I evaluate the various paradigms for studying imitation and conclude that selective imitation is evidence of theory of mind and hence self-consciousness, a conclusion that is consistent with CPH. Chapter 7 is devoted to exploring the connection between memory and self-consciousness and, based on episodic memory studies, I come to the conclusion that there is good evidence that scrub jays are self-aware. In chapter 8 I concentrate on one species, rats, and examine a range of research paradigms purporting to demonstrate rat rationality. Detailed analyses of these experiments leads me to conclude that rationality need not be invoked to explain the results, which can all be accounted for using associative and other non-conceptual theories.

# Statement of Candidate

I certify that the work in this thesis entitled “The Threshold of Self-Consciousness” has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree to any other university or institution other than Macquarie University.

I also certify that the thesis is an original piece of research and it has been written by me. Any help and assistance that I have received in my research work and the preparation of the thesis itself have been appropriately acknowledged.

In addition, I certify that all information sources and literature used are indicated in the thesis.

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Stephane Joseph Savanah

stephane.savanah@gmail.com

Dec 2012

# Acknowledgements

This work was supported by the ARC Centre of Excellence for Cognition and its Disorders (CCD) and the Department of Cognitive Science, Macquarie University, Sydney Australia.

I thank the members of my supervisory board for their guidance, feedback, support and encouragement over several years: Professor John Sutton (principal supervisor), Professor Peter Menzies, Dr. Mitch Parsell and Dr. Glenn Carruthers (associate supervisors). Before an overseas relocation forced a change, my principal supervisor was Dr. Tim Bayne, to whom I owe a debt of thanks for his initial mentorship. For their feedback on various chapters and some stimulating and thoughtful discussions, my thanks to Dr. Wayne Christensen, Dr. Stuart Palmer, Dr. Liz Schier, Dr. Ken Cheng and my good friend Mark Avery. For taking the time to correspond with me via email to answer specific questions and/or to provide insightful comments in specific areas, my thanks to Professor Gordon G. Gallup Jr., Dr. Aaron Blaisdell, Professor Endel Tulving, Dr. Judy DeLoache and Dr. Alex Byrne.

Chapter 3 (The Concept Possession Hypothesis of Self-Consciousness) was presented as a paper at the *Second Consciousness Online Conference* in Feb 2010 and subsequently published in *Consciousness and Cognition* [Savanah 2012a]. I would like to thank the assigned conference commentators, Dr. James Dow and Dr. Kristina Musholt, for their helpful and insightful commentary as well as two anonymous reviewers for the journal. A version of chapter 5 (Mirror Self-Recognition) was published as a paper in *Biology and Philosophy* [Savanah 2012b]. For their helpful feedback on an earlier version of this paper, my thanks to Professor Kim Sterelny and an anonymous reviewer.

Finally, I would like to express my deep gratitude and appreciation to my family (my wife Rita, son Sean and daughter Sinead) for their support, encouragement and tolerance throughout the term of my candidature.

# Part 1: The Nature of Self-Consciousness

# Chapter 1: The Self and Self-Consciousness

## 1.1 Introduction

Self-consciousness[1] is being studied in children and animals using a wide variety of research paradigms but there is less than universal agreement either on what constitutes self-consciousness or on what counts as sufficient evidence of its possession by a subject. When Gallup [1970] claims to have demonstrated possession of self-awareness by chimpanzees because they recognised themselves in a mirror, is this the same concept of self-awareness that Michael Lewis [1994] refers to regarding infants' self-evaluative emotions? Are organisms that demonstrate a theory of mind self-aware in the same sense as those that show the capacity for mental time travel? My approach to addressing these issues is to lay some common ground. Firstly, I formulate a baseline conception of self-consciousness. That is, I propose what should be taken as the minimum requirement for a claim of self-consciousness in a subject. Of course, not everyone will be willing to accept my conception as definitive but it will not be unfamiliar to the reader and it will at least serve to define my terminology and clarify the notion of self-consciousness that is investigated throughout this work. That notion is, simply put, the same kind of self-consciousness possessed by human adults[2]. It could be encapsulated as follows: *an understanding of one's own existence as a psychological subject with intentional[3] agency.* Below, and also in more detail in the next chapter, I contrast this with what has been called a 'primitive' form of self-awareness [Bermũdez 1998] in which the subject has access to some of its own self-specifying, non-conceptual mental states. This latter form, I argue, should not be considered self-consciousness proper. Secondly, in chapter 3, I argue that concept possession alone is sufficient for self-consciousness and in chapter 4 I examine ways in which concept possession could be conclusively determined in non-linguistic organisms.

These first four chapters constitute part 1 of the thesis, in which I lay out the theoretical groundwork. Note that part 1 is not about cognitive architecture *per se*; I rely very little on empirical considerations in this more dialectic first half for it is this very same empirical data

[1] I use the terms self-consciousness and self-awareness interchangeably in this work as further discussed in the text.
[2] I aim to clarify the limits of this simile in the text. For example, I do not wish the reader to presume that by 'human-like self-consciousness' I imply the necessity for language (this issue is addressed in chapter 4).
[3] I deliberately qualify 'agency' with 'intentional' to allow for a broader sense of the word 'agency' which applies to non-intentional subjects. That is, 'agent' might be used like 'actor' to refer to any organism that can perform actions, whether under deliberate control or just (say) reflex-like.

that I analyse in the second half using the framework presented in part 1. However, it will be seen that the ideas presented in part 1 have significant implications for cognitive science. If I am right, commonly accepted views about animal intelligence, concept possession, self-consciousness and rationality will need to be reviewed. Moreover, although I do not explicitly pursue the project here, taking these ideas further would challenge other dogmas such as the Computational Theory of Mind (though see section 4.4 for some brief notes in that direction). I acknowledge that the central idea presented in part 1, the *Concept Possession Hypothesis*, is outside current mainstream opinion and will be challenged in the literature. I heartily look forward to facing these challenges in future discourse.

In the following section I narrow in on my conception of self-consciousness. Since this is dependent on the self-conscious organism having a concept of a *self*, in section 1.3 I examine some notions of 'self'. I consolidate the many notions of the self into five broad categories and argue that two of these are essential and inseparable. This leads to a conception of self-awareness as (minimally) an understanding of the self as an intentional agent. That is, in addressing what it means to understand one's existence, I argue that the answer is an awareness of one's own intentional agency. Finally, in section1.4, I briefly sketch the many research paradigms purporting to demonstrate self-consciousness in order to demonstrate the breadth of the field. This establishes the groundwork for Part 2 of the thesis in which I examine several of these research paradigms in greater detail in order to evaluate the validity of the claims made.

**Some Terminological Points**

Throughout this work I do not make a distinction between the terms 'awareness' and 'consciousness' as some authors urge (for example, Chalmers [1995] suggests we use 'awareness' for those phenomena that are explicable in terms of computational or neural mechanisms and reserve 'consciousness' for the phenomena of experience). However, there is probably little hope of achieving consensus for these arbitrary distinctions. Of course, I do acknowledge the likelihood of different types and/or levels of consciousness but rather than assigning each a single word such as 'awareness' or 'consciousness' I find it more convenient to treat those terms as synonymous and qualify them. It is more explicit, for instance, to talk of '*phenomenal* consciousness' and '*access* consciousness' as distinct conceptions [Block

1995]. Accordingly, I also use the terms 'self-consciousness' and 'self-awareness' interchangeably throughout.

### 1.2 Self-Consciousness and the Self-Concept

The notion of self-consciousness I am interested in is one in which the subject is aware of itself *as a self*. That is, self-consciousness in the existential sense, where the subject understands what it is to be a self and knows that it is a self. Inherent in this notion is the possession of a self-concept by the subject, and, to borrow Bermũdez's [1998] phrase, is what I think of as 'full-fledged self-consciousness'[4]. This notion of full-fledged self-consciousness can be encapsulated as *an awareness of one's own existence as a psychological subject with intentional agency.*

In defining my terms above, I do not mean to discount the possibility of different types or levels of self-consciousness. For example, others may be content to use the term for any case in which an organism has access to some of its mental or physical states. Now, insofar as an organism may perceive (say) its own pain, it could legitimately be claimed that this is a case of self-awareness in that the organism is in a state of perceptual awareness and the content of that *awareness* involves the *self* or part of the self (in this case a body part). So there is certainly a *sense* in which self-perception counts as self-consciousness. But although there may be value in knowing whether a rabbit is aware of its own pain, this is not the type of self-awareness that I am most interested in and that I refer to in this work. I want to find out if the rabbit knows it is a rabbit. Or rather, I want to know if the rabbit knows that *it is*, period. What I am most concerned with is whether any animals have something like the self-awareness that human adults have. In thus using humans as the model for self-consciousness I am deliberately setting the bar high. As a species, humans have the capacity to understand that they exist. And not just that they exist in the same way as other physical objects in the world (which might be termed 'objective self-awareness': see, for example, Hart & Fegley [1994]). This type of awareness of self (as a physical object) is not the type generally considered when discussing self-consciousness in humans. Humans have the capacity to understand that they exist as psychological subjects as well as physical objects. So, for example, self-perception by organisms, where this just means perception of themselves as

[4] I discuss Bermũdez's notion of full-fledged vs. primitive self-consciousness in chapter 3.

objects in the world, does not quite make the grade for self-consciousness; it has to be perception of themselves as psychological subjects. The self-conscious organism *understands* that it exists as a *subjective self* and is often referred to as having a *self-concept* in both the psychological literature (e.g., Michael Lewis' [1994] 'idea of me') and the philosophical literature (e.g., "[self-consciousness means] the possession of the concept of the self and the ability to use this concept in thinking about oneself" [Block 1995, p235]). Exactly what should count as a concept of the self is discussed in section 1.3.

Of course there are some authors that argue for different types or levels of self-consciousness and who might wish to claim self-perception (for example) as at least a type of self-consciousness, as mentioned earlier. For example, Gennaro [1996] admits of levels of self-consciousness including one that involves no more than "*nonconscious thought awareness* of one's own mental states" (p17). I accept these as alternatives views or definitions of self-consciousness, but I wish the reader to understand that when I use the term 'self-consciousness' I am referring to (minimally) awareness of one's existence as a psychological subject. I expand on this notion below where I argue that this involves a metacognitive understanding of one's own intentional agency.

### 1.3 Conceptions of the Self

In this section, I present five distinct conceptions of self, and from these I argue that two of them are inseparable and together form a fundamental concept of self. These are the Agentive Self and the Metacognitive Self. I argue that being self-conscious can be characterised as having a concept of oneself as an intentional agent, and that this necessarily involves access to certain types of one's own mental states.

There are many and diverse conceptions of 'self'. For example, Allport [1943] lists eight (the 'ego' as: *knower; object of knowledge; primitive selfishness; dominance drive; passive organisation of mental processes; fighter for ends; a behavioural system; subjective organisation of culture);* Mitchell [1994] seven (*perceiving self and self perceived; self extended; self identified; self imagined; self objectified and intersubjective; self presented and evaluated; dissociation of the self; self evaluated by the self*); and Neisser [1988] five (*ecological self; interpersonal self; extended self; private self; conceptual self).* Many ideas

about what constitutes a self are tied to the notion of 'self-image' (i.e. self-assessment of one's qualities or self-worth). But I am not interested here in notions of the self to which higher level properties can be ascribed (such as 'I am Australian' or 'I am a fighter for ends') but rather in ideas of what it means to be a self in a more fundamental sense. There are many of these, too, but I have consolidated those that appear too similar to usefully differentiate. For example, Neisser [1988] describes a notion of self that he calls the Private Self, in which organisms have become aware of the exclusivity of their own conscious experiences. Although this notion might seem distinct from those I list below, I believe it is captured within the notion of the Metacognitive Self, which implies that the organism has developed the capacity to attend to its own mental states. The criteria for my categorisations are, firstly, that they should be sufficiently distinct from each other in their descriptions and secondly, that they each be strongly associated with emerging self-awareness rather than self-image. I classify the various notions of self into five categories, named the *Physically Extended Self*, the *Temporally Extended Self*, the *Social Self*, the *Agentive Self* and the *Metacognitive Self*. These are distinct notions of self, but, as I show below, they are not necessarily all independent.

**The Physically Extended Self**

The underlying idea here is that an organism's own body is a source of self-specifying content in awareness. The Physically Extended Self may be referred to as the 'embodied self' or what Gibson [1979] and Neisser [1988] call the 'ecological self'. At its most basic level, awareness of the Physically Extended Self is manifest in self/non-self differentiation: the ability to 'recognise' the difference between one's own body and the rest of the world. Of course, all organisms demonstrate this capability. Carnivores, for example, do not chew off their own limbs even when dying of hunger. Michael Lewis [1994] points out that even T-cells can recognise and differentiate themselves from foreign protein.

Another form of awareness of body-specifying content is awareness of the body's spatial location in the world. Insofar as an organism is able to perceive the outside world, it is in essence able to perceive its own spatial existence. The position and movement of an organism is specified by the flow patterns in the visual field, and the idea that there is an objective world cannot be separated from the idea that the subject is somewhere *in* the world [Evans

1982; Neisser 1988]. Bodily spatial locatability is apparently of some importance for the self, however it is not necessarily indicative of self-awareness in the sense I have defined. It is still a *pre-reflective* self-awareness that accompanies and shapes spatial experience, to be distinguished from *reflective* consciousness of the self [Bermũdez 1998; Gallagher & Zahavi 2006]. If there is a gradation of cognitive capacities that go to making up the development of 'full-fledged' self-awareness, then organisms with (pre-reflective) awareness only of the Physically Extended Self are in the early stages. Nevertheless it is an important inclusion since (as discussed in chapter 5) there is still debate as to whether (for example) mirror self-recognition indicates anything more substantial than this type of non-conceptual bodily self-awareness.

According to Neisser [1988] most likely some type of awareness of the ecological self is present from birth, but as very young infants have no internal representations to be conscious of, the ecological self cannot be an object of thought for them. It is widely accepted that very young infants cannot yet have a concept of the self, and so are unable to exhibit self-awareness in the psychological sense. Even so, it is entirely possible that the Physically Extended Self is an important and indispensable component of the developing 'full-fledged' self-awareness[5]. Studies on blind children show that they are slower to develop a sense of self than sighted children [Neisser 1997], possibly due to a reduced ability to explore their physical environment, including their own bodies.

Another way to regard the Physically Extended Self may be recognition of oneself as a physically extended *subject* – that is, as an *embodied* psychological self. But in this case we still need to ask what makes up the sense of a psychological self. I suggest that knowing oneself as a subject (a psychological self) *with a body* implies an understanding that the body is under one's control. If the Physically Extended Self is to be understood in these more psychological terms, then it already implies a more sophisticated conception of the self as an agent – that is, the agent controlling the physical body (I say more on this later).

In summary, awareness of the self as a Physically Extended Self might be considered a form of self-awareness only in a weak sense, to the extent that the content of awareness is the self as a physical object (one's own body). However, it is evident that most organisms, including

[5] I remain open to this possibility but whether or not it is true does not impact upon my theses in this work. Povinelli & Cant [1995] suggest that (in a very particular way) the physical body of our ape ancestors may have played a significant role in the evolution of self-awareness, as I briefly discuss in section 9.1

lower forms that are not considered self-conscious, can differentiate the bodily self from non-self. This type of non-conceptual awareness of the (physical) self is not the type we are considering when examining self-consciousness. We are interested in awareness of the self in the psychological sense; as a subject of conscious thought, not just as a physical object in the world.

**The Temporally Extended Self**

The Temporally Extended Self relates to an awareness of one's existence in time; that is, an awareness that the self has existed in the past and will exist in the future (often referred to as a capacity for mental time travel). It is to be aware of oneself as existing outside of the present moment [Neisser 1988]. The notion of a Temporally Extended Self is linked to the idea of a 'narrative self' or 'autobiographical self'; that is, a self made up of a narrative based on episodes in one's past. There are many who consider this aspect of self-knowledge to be intrinsic to the self, suggesting the possibility that what we recognise as self is what is convertible into some version of narrative. Autobiographical memory (stories of the self) and self-concept are sometimes said to be interdependent in development: adults develop a life-narrative that effectively defines the self in terms of remembered experiences [Neisser 1988; Bruner 1997; Bruner & Kalmar 1998; Nelson 2005]. From these descriptions it is clear that the narrative self is important to a self-concept in the sense of personal identity. It seems likely however, that development of a personal identity relies on an already formed sense of one's own existence: one must already know oneself to *be* a self before one can form an opinion on what sort of person one is. Hence Temporally Extended Self may be seen as a more highly developed sense of self than is required for my notion of self-consciousness.

**The Social Self**

The Social Self is a notion of self as an organism that understands it exists not in isolation but as a member of a community of conspecifics. The self recognises that there are others who are like the self but who are distinct from the self. There is no shortage of psychologists and philosophers who emphasise the importance of interpersonal relationships in the emergence of self-awareness as well as a sense of identity. It is not clear from the literature on

development of the Social Self whether awareness of the other emerges from awareness of the self ("they are like me") or if awareness of the self emerges from awareness of the other ("I am like them") or, indeed, if they emerge concurrently. To many authors, the sense of self comes to be because of and in response to a social world of many others [Neisser 1997; Walsh & Banaji 1997; Bruner & Kalmar 1998; Mascolo & Fischer 1998]. Neisser thinks self-concept originates during occasions of caretaker/child interactions in which the child is the object of attention, when the caretaker speaks to the child about the child ("that's a good girl" etc). The result being that the child takes herself as an object of thought. Pribram & Bradley [1998] contend that the requisite neurobiological organization for the development of a stable self is prompted by the interactions in the mother-infant dyad. Some others, while acknowledging the importance of social interactions for psychological well-being, do not necessarily believe that the cognitive capacities underlying self-awareness depend on social interactions [Lewis 1994; Mitchell 1994; Nichols 2005]. Yet others explicitly link development of self-awareness with the *co-development* of other-awareness [Gopnik & Meltzoff 1994; Parker & Milbraith 1994; Rochat 2003].

It is thus still an open question as to if, and in what way, social interactions are necessary for the formation of a self-concept. However, it is a reasonable assumption that if one can conceive of others as being selves, one must have a well-developed conception of *what* the self is. Thus, even if socialisation drives the development of this sense of self, such that it could not develop otherwise, the essential notion of 'self' is applied to *oneself*. As such, the realisation that one is among a community of similar others (i.e. awareness of oneself as a Social Self) cannot be fundamental as it presupposes the existence of a more essential conception of oneself. This issue is discussed in greater length in chapter 6 in the section on Theory of Mind, in which I argue that an understanding of the intentions of others depends on an understanding of oneself. Thus, an awareness of oneself as a Social Self (i.e. an understanding that others are like the self) is not fundamental as it requires a pre-existing (or, at least, co-existing) self-awareness.

**The Agentive Self**

By the Agentive Self, I refer to an organism's awareness that it is an intentional agent with causal powers in the world. The Agentive Self has the capacity to process information and

decide on the best action. An infant will initially interact with the world without self-awareness but at some stage of its development will become aware of its own agency. The dawning of its awareness of its own agency appears to be a significant developmental milestone [Stechler 1982]. It seems unlikely in this course of events that a child will develop awareness of its own existence in isolation of the world. More plausible is that the child develops awareness of self within the context of its interactions with the world.

The notion that self-awareness is awareness of agency has wide currency within psychology and the philosophy of mind. For example, Bermũdez [1998] presents three categories defining psychological subjects: self-awareness of themselves as perceivers, bearers of reactive attitudes and *agents*. According to Lucy O'Brien [2007] "...our most basic awareness of ourselves is as performers of actions, mental and physical" (p3). Both philosopher William Richards [1984] and psychologist Gerald Stechler [1982] *define* self-consciousness as consciousness of agency. I argue that awareness of the self as an Agentive Self constitutes a major component of full-fledged self-consciousness. However, as argued next, implicit in awareness of agency is metacognition.

**The Metacognitive Self**

Metacognition can be characterised as the capacities for monitoring and control of one's own mental states [Nelson 1997; Fernandez-Duque et al. 2000]. In a sense, 'monitoring' can occur in the absence of the ability to 'control', if one allows direct experiential access as a case of 'monitoring'. However, the connotation attached to 'monitoring' is that it is a *deliberate action* of the agent rather than involuntary (such as the sensation of pain). Thus, the capacities for both *monitoring* and *control* together are implicit in metacognition. Given the characterisation of the Agentive Self above, it is clear that self-awareness must involve metacognition since awareness of self implies this kind of access to one's own mental states. The Metacognitive Self, then, applies to an organism that is able to monitor and control its mental states and this is the other major component of full-fledged self-consciousness along with the Agentive Self. Though distinct, these are overlapping conceptions and together constitute a fundamental concept of self, as discussed next.

### A Fundamental Concept of the Self

I believe the five conceptions of self presented above do a satisfactory job of covering the different notions of the self in self-awareness. Although there are many more ways to form descriptions of self, these other descriptions are encompassed within one of my five conceptions, or else they are not relevant to self-awareness. For example, Mitchell's [1994] 'perceiving self and self perceived' fits comfortably under my description of the Physically Extended Self as does his 'self identified', while his 'self extended' (relating to one's material possessions and social position) relates more to personal identity than to self-awareness. Similarly Neisser's [1988] 'interpersonal' self is covered by the description of the Social Self while his 'private self' fits within the broader Metacognitive Self.

Regarding the question as to what type of self is implicit in self-awareness I argue that the answer is the combination of the Agentive Self and the Metacognitive Self. These are inseparable in that they always coincide, and together constitute a fundamental concept of the psychological self. An organism that understands itself to be an intentional agent has the capacity to control its own immediate intentions to act, implying the application of metacognition. Also, knowingly performing metacognitive acts of monitoring and controlling one's own cognition is the deliberate action of an intentional agent. Andrew Brook makes a similar point in this way:

> When one is aware of oneself by doing cognitive and perceptual acts, one is aware of oneself as spontaneous, rational, self-legislating, free – as the doer of deeds, not just as a passive receptacle for representations... [Brook 2001, p21]

I suggest that being conscious of oneself as an Agentive Self, with its implicit metacognition, represents the most fundamental concept of self and is sufficient to meet the criteria for self-consciousness as I earlier defined it – that is, awareness of one's own existence as a psychological subject.

## 1.4 Self-Consciousness Research Paradigms

In this section I present a brief overview of the research programs relevant to self-consciousness. In each of these, the behaviour under scrutiny is linked in some way to a theory or view of self-consciousness. The reviews presented below are intended to highlight

the main issues under consideration in summary form; it is sufficient for my purposes at present to only demonstrate the breadth of research and the variety of opinions on self-awareness. The research areas examined below are *mirror self-recognition; Theory of Mind; episodic memory; self-evaluative emotions* and *metacognition.* In Part 2 I have selected three research paradigms to examine in detail, which I consider the main contenders for self-consciousness studies in animals. These are: mirror self-recognition (chapter 5); imitation (as an example of Theory of Mind, chapter 6); and episodic memory (chapter 7). My reasons for these choices will be made apparent in the brief discussions below. In chapter 8 I take a different approach, focusing instead on one species, rats. As lab rats are extensively used as subjects in many different paradigms, this approach allows me to address the question as to whether rats as a species are self-aware while at the same time evaluating several more research paradigms.

**Mirror Self-Recognition**

Probably the most studied behaviour with respect to self-awareness is mirror self-recognition (MSR). Since Gallup's pivotal 1970 paper on MSR in chimpanzees a very large number of studies have been conducted on human infants and animals using the technique. Not all chimpanzees are able to recognise themselves in a mirror, but it is well established that as a species they have this cognitive capacity. Gallup [1975] considers the mirror test to be a definitive indicator of self-awareness as one needs a self-concept in order to recognise that the image in the mirror is oneself. A less inflationary view is that MSR indicates little more than bodily self-awareness. An organism seeing its own reflection might view it in some way as an extension of the body, so that seeing the reflection is perceived similarly to seeing one's own actual body. One strong proponent of this type of view is Mitchell [1993, 1994, 1997a, 1997b], who insists that MSR can be explained by kinaesthetic-visual matching, wherein the organism matches the visual experience of its image in the mirror with the proprioceptively experienced movement of its own body. These issues are examined closely in chapter 5, in which I argue that kinaesthetic-visual matching actually argues in favour of the view that MSR indicates self-consciousness. Thus, mirror self-recognition is indeed (as *Scientific American Mind* called it) the 'gold standard' for self-awareness [Mossman 2007].

## Theory of Mind

Theory of Mind (ToM) refers to one's ability to understand a conspecific's mental state. In prominent theories of ToM this capacity implies the existence of self-awareness in a subject. The essence of most theories of ToM (including simulation theory, as I argue in chapter 6) is that in order to infer a conspecific's mental state, an organism must somehow use its own mind as a model. Since one can have no direct access to a conspecific's mental state, one must infer it by interpreting the conspecific's behaviour. To do this one must project onto the conspecific associations of behaviours and mental states that have been modelled in some way on the self. I argue this point and cover ToM theories generally in more detail within chapter 6.

There are many different research programs studying self-awareness in human infants and animals based on the theory of Theory of Mind, including *imitation, shared attention, gaze following, pretence, deception* and *false belief.* The majority of these I cover only briefly below, but I select only one representative ToM paradigm to examine in detail in Part 2: imitation.

Probably the most basic demonstration of ToM is a set of capacities that can be grouped under the general heading 'gaze-directed behaviour', including **gaze cueing**; **gaze following** and **shared attention**. In these studies infant or non-human primate subjects are observed for their ability to redirect their attention based on another's attention. If a subject sees an experimenter gazing at an object and then turns to look for the object of the experimenter's attention, then this may be evidence that the subject understood that the experimenter was performing the action of looking. Therefore, in some fundamental sense, the subject demonstrated a ToM about the experimenter; he understood that the experimenter was engaged in the mental act of observation and that the object under observation was perceived via the eyes. The most basic of these, *gaze cueing*, is stimulus-driven rather than goal-oriented and occurs when an experimenter's averted gaze triggers automatic attentional orienting responses in the subject; the more sophisticated *gaze following* enables the subject to isolate the object being gazed at [Parsell 2009]; while *shared attention* appears to operate at a yet higher level [Parsell, personal communication]. Povinelli & Prince's [1998] view of the available experimental data is that human infants interpret the actions of others in terms of a mental state we call attention around 18 months of age. Prior to this age, they point out

that things are less clear, even in experiments where mother and infant are both looking at the same object in unison: *joint* attention does not ensure *shared* attention.

Povinelli & Prince [1998] caution us against an anthropomorphic reaction to attention and gaze-following behaviour in animals: "Surely a chimpanzee that spins around to look where you are looking must have understood that you were looking at something behind him. Well, maybe – but maybe not" [Povinelli & Prince 1998 p58]. Although chimpanzee gaze-following experiments initially appear to confirm a high-level model (they interpret line of sight as a projection of attention), there is an alternative interpretation that this is just a part of the learning mechanism. Some evidence for this interpretation is that the 'learned' behaviour is not retained – a year or so after the experiments were done the same chimps' behaviours had reverted to chance. Given these concerns, shared attention and gaze following paradigms might not be conclusive enough to count as evidence of ToM.

**Imitative behaviour** comes on a variety of forms, some of which appear to be reflex-like and others of which appear to rely on conscious deliberation. As such the issue becomes which, if any, imitative behaviours indicate ToM and therefore self-awareness. Reflex-like behaviour, such as fainting when seeing another person faint, may be subserved by 'mirror neurons' [Savanah 2006], although mirror neurons have also been used to explain goal-oriented imitative behaviours [Rizzolati 2005]. This question and others related to imitation is explored in detail in chapter 6, where I argue that Selective Imitation provides a good paradigm for self-awareness. In Selective Imitation only some actions are copied, in such a way as to indicate that the imitator has understood the intention of the model.

**Deception** is another key field of study in the area of ToM, especially in non-human primates. For example, gorillas play games wherein they display awareness that they can mislead other animals [Parker & Milbraith 1994]. When a primate engages in deliberate deception of its conspecifics, for example by acting in a manner that leads others away from a source of food in order to keep all the food to itself, this may be evidence of ToM. It has had to assume that its victims would interpret the meaning of its misleading behaviour and it would have to have predicted how they would behave as a result. This is a fairly sophisticated form of ToM in that not only does the subject project itself onto others, but it also realises that others are able to do likewise on itself. That is, the subject needs to understand that its victims also have a theory of mind. It appears that many different primate species engage in what is technically denoted 'tactical deception', in which an organism employs an act from

its normal behavioural repertoire in a unique context so that it serves to manipulate another individual. But sceptics are still unconvinced, as much of the evidence is anecdotal in nature; plus, there are plausible explanations of seemingly intentional deception that are based on a paradigm of associatively learned behaviour [Povinelli & Prince 1998]. As such deception may not offer a convincing paradigm for ToM.

There is a large contingent of researchers investigating **false belief** as an indicator of ToM in children. In one version [Baron-Cohen, Leslie & Frith 1985] infants observe one actor placing an object inside a container and then leaving the room. They then observe a second actor remove the object and place it in a different container. When the first actor returns the infants are asked where she will look for the object. Younger children tend to indicate the second container (where the object actually is) rather than the first (where the first actor must falsely believe it to be). This demonstrates that they do not understand that the first actor might have a false belief. When children develop a ToM, it is argued, they are then able to understand that others may have false beliefs and will pass this test. Bloom & German [2000] (among others) dispute the idea that failure at the false belief task indicates a lack of ToM, as they argue that failure could be the result of other factors such as the relative complexity of the task. However, they agree that it can be used as a positive test. The 'magic age' for acquisition of false belief understanding in the early literature was 4 years old, though as early as 1998 Chandler & Carpendale [1998] claimed that under optimising conditions, 2-3 year-olds could also succeed. More recent non-verbal experimental paradigms (e.g., using eye-trackers) have lead to claims for false belief in infants as young as 25 months [Southgate, Senju & Csibra 2007] or even 15 months [Onishi & Baillargeon 2005].

Other studies on ToM in infants and animals include **empathy** and **teaching**, in both cases of which it is assumed for fairly obvious reasons that these behaviours rely on the subjects having a ToM. Regarding empathy, Mitchell [1994] reports that 10-12 month old infants become distressed by another's distress but can only offer comfort around 18 months: 'Our empathic responses are apparently a communicative portrayal of the other's expressed feelings…' [Mitchell 1994 p92]. In the case of teaching, Custance & Bard [1994] inform us that teaching is a key factor in the development of imitative abilities, which develop in social interactions.

To summarise, ToM offers several paradigms for the investigation of self-awareness. To represent this category I have selected imitation for further analysis in chapter 6. Imitation as

a paradigm is suitable for animal studies and the topic provides an opportunity to examine the relatively recent discovery of ‘mirror neurons’.

**Episodic Memory**

Episodic memory is intrinsic to a particular conception of self-awareness described in section 1.3 as the Temporally Extended Self. In this conception, the self is aware of its own persistence through time. It has been argued that self-awareness is not possible without this self-knowledge as one’s very identity seems dependent on a remembered past and a projected future [Bruner 1997], but as I earlier argued this relates more to personal identity than self-awareness. It is perhaps less controversial to argue that episodic memory is not possible without self-awareness, for one must have a concept of oneself in order to have memories of oneself. If this is right then demonstration of episodic memory in an organism would simultaneously demonstrate the existence of self-awareness. In human infants, it appears that episodic memory does not emerge until about 3 years of age [Neisser 1988; Povinelli & Prince 1998; Nelson 2005]. This finding is based on studies that show 3 year old infants have difficulty remembering time-sequenced past events even though they fare well on tests of semantic (factual) memory.

According to Tulving [2005] memory experiments on animals cannot yet distinguish between episodic memory and semantic memory. Tulving himself suspects that only humans have the capacity for episodic memory. Nevertheless, Tulving suggests the possibility of designing experiments to detect episodic memory in animals. Tulving believes the ability to remember time-sequenced events allows forward as well as backward mental time travel. In other words, it allows for future planning. Therefore, an experiment that tests for the ability to plan ahead will provide evidence of episodic memory in the test subjects. In fact there is intriguing recent evidence of future planning in scrub jays (e.g., Raby, Alexis, Dickinson & Clayton [2007]). As such this topic is analysed in depth in chapter 7, wherein I conclude that there is strong evidence of self-consciousness in scrub jays.

## Self-Evaluative Emotions

Self-evaluative emotions such as embarrassment, envy, pride, shame and guilt are sometimes called the secondary emotions to distinguish them from the so-called primary emotions of anger, fear, disgust and joy [Lewis 1994, 1997]. They are also sometimes referred to as the self-conscious emotions. The primary emotions appear much earlier than the secondary emotions and according to some authors the secondary emotions can only appear in infants after they have attained self-awareness [Lewis 1994; Nelson 2005]. Under 18 months, the child exhibits the primary emotions, but it is not until around 18 months that the child visibly acquires the secondary emotions [Lewis 1994].

All of the secondary emotions relate to the infant evaluating himself against a social standard. Whereas the primary emotions may be directly evoked as a reaction to a stimulus without an intervening self-appraisal, the secondary emotions are always evoked following a consideration of the self. One can only feel pride, say, after evaluating one's own achievement as 'good' or 'worthy'. Similarly, shame and guilt arise only as a response to an evaluation of the self as having acted 'badly'.

Determining the existence of self-evaluative emotions in animals is difficult as they are unable to report their emotions and their observed behaviours will be open to alternative explanations. For example, when a dog has its tail between its legs this need not necessarily imply shame. It might (for instance) simply be a response to an action it took that it has come to associate with a subsequent punishment. Some self-evaluative emotions may be detectable in apes, at least in sign-using apes that can to some degree report their own mental states. For example, Mitchell [1994] reports that the beginnings of reflective self-awareness are present in some of the sign-using apes when they evaluate their actions as 'good' or 'bad'. Nevertheless, Mitchell is unwilling to place too much weight on this as evidence of human-like self-awareness. Given the difficulty of using this paradigm for animals, I have not selected it for further examination.

## Metacognition

I earlier argued that metacognition and awareness of agency are linked and implicit in self-consciousness. Thus a demonstration of metacognition could be taken as evidence of self-

consciousness in a test subject. However, metacognition is difficult to demonstrate in non-linguistic organisms. Currently metacognition experiments rely on a 'bail-out' paradigm. For example, an experiment described by Smith [2005] investigates the ability for rhesus macaques to judge their own ability at visual density discrimination tasks. After a suitable period of training, the monkeys are 'asked' in a series of trials to judge whether a box of pixels on a computer screen is dense or sparse according to a set threshold, and are rewarded for a correct answer. For a lesser but guaranteed reward the subject also has the option to decline the test, in effect (it is claimed) to answer with a response of 'uncertain'. Judgements of this type are deemed by the experimenters to be acts of metacognition because the subject is making a decision not on the density of the box but rather on knowledge of his own ability to succeed at the task. Analogous experiments have been conducted with rats [Foote & Crystal 2007]. However, I argue in chapter 8 that the results in these 'bail-out' paradigms can be explained by associative theories without the need to assume metacognition in the subjects.

## 1.5 Conclusion

In researching self-consciousness in animals, the question I address is whether any animals are self-aware in more-or-less the same way we humans are. Although it is interesting to know if an animal has *any* mental states like ours, it would not be surprising to discover that (for example) some animals can experience pain or even joy. It *would* be surprising for many to discover that an animal is aware that it exists in much the way that we know this fact about ourselves. I have argued that the most fundamental aspect of this conception of self-consciousness is awareness of one's own existence as an intentional agent, and that this necessarily involves the capacity for metacognition.

I have provided a brief overview of many research paradigms aimed at detecting the existence of self-awareness in animals and human infants. Several of these are analysed in depth in part 2 of this work, in which I argue that there is strong evidence for the existence of self-awareness in chimpanzees, dolphins and elephants (based on mirror self-recognition studies); and in scrub jays based on episodic memory studies. A detailed analysis of key rat experiments reveals no evidence for self-awareness in rats.

In the rest of part 1, I set the theoretical groundwork for the analyses of empirical data conducted in part 2. In the next chapter I present a critical dichotomy in the philosophy of self-consciousness, between a type of self-access that is direct, unmediated and non-conceptual versus a type of access that is relational and concept-dependent. I argue that only the latter type indicates self-consciousness proper and that this view is consistent with generally accepted theories of self-consciousness. Following on from this, in chapter 3 I argue that the key component separating organisms that are self-conscious from those that are not is concept possession. If so, then determining the existence of concept possession in organisms provides evidence of self-awareness. Of course this is itself no simple task and I devote chapter 4 to a discussion on how to determine concept possession in animals.

# Chapter 2: Introspection and the Fundamental Dichotomy

## 2.1 Introduction

The central idea proposed in this work (and a key component of the analyses performed throughout Part 2) is the Concept Possession Hypothesis of Self-Consciousness (CPH), as explicated in chapter 3. Briefly, the claim made in CPH is that concept possession alone is sufficient for self-consciousness proper. In chapter 1 I defined and defended my notion of 'self-consciousness' as *an understanding of one's own existence as a psychological subject with intentional agency.* In the current chapter I present the thesis that there is a fundamental dichotomy in the philosophy of self-consciousness. This dichotomy is foundational to CPH as it forms the basis for a simple taxonomy of consciousness that is a key element in the argument for CPH. This taxonomy, presented in the next chapter, separates self-conscious organisms (the highest level) from organisms that are conscious but *not* self-conscious (the next level down). The latter category of organisms at best only has the so-called 'primitive form of self-consciousness' (mentioned briefly in chapter 1), while the former has self-consciousness proper. In this chapter I (i) explain what the fundamental dichotomy is, (ii) argue for which side is correct, (iii) compare the consequences of this position with generally accepted theories of self-consciousness, and (iv) in so doing lay the groundwork for the CPH argument.

A contentious debate within the philosophy of self-consciousness is about how we can become aware of the self in a way that indicates self-consciousness. In chapter 1 I described self-consciousness as being aware of the *self as such* – not being aware of just any physical or mental state one is in, but *knowing* that one *is* a self, in the sense of knowing oneself as an intentional agent. Becoming aware of the self in this way requires some kind of access to the self. In one view this self-access is 'privileged' and unlike the access we have to other persons or anything else in the world. In this model access is direct and unmediated. The opposing view models self-access on perception, such that we perceive our selves in a roughly analogous way to the way we perceive others. These two views form a dichotomy, which can be generalised as that between an *intrinsicist* position (as in the privileged access model) and a *relationalist* position (as in the perceptual model). I show that this dichotomy is rife throughout the philosophy of self-consciousness and I refer to it henceforth as the Fundamental Dichotomy of Self-Consciousness.

I argue below that the correct model for access to the self is the perceptual model, that is, we should adopt the relationalist side of the Fundamental Dichotomy. I do this by showing that access to the self as such can only be attained through the deliberate action of introspection. Other types of self-access can be direct and immediate, but these are not the type relevant for self-consciousness – that is, for access to the self *as such*. For example, the experience of pain can be considered a type of direct and immediate self-access, but it is likely that pain can be experienced by organisms that are not self-conscious. This example could be seen to represent a 'primitive' form of self-consciousness, in the sense that it at least describes access to some types of self-specifying (though non-conceptual) mental states. But, as discussed in chapter 1, such primitive forms do not meet the standard for my notion of self-consciousness proper. Self-consciousness means access to the self *as such* and implies a self-concept. As discussed below, while the relationalist side of the Fundamental Dichotomy requires concept possession the intrinsicist side does not.

Kriegel [2007] issued a challenge to self-consciousness theorists: that the 'peculiarities of self-consciousness' must be accounted for in the context of a general theory of self-consciousness. The Fundamental Dichotomy is examined below in that spirit. If, as I claim, the relationalist side of the Fundamental Dichotomy is correct then it should be consistent with known aspects of self-consciousness. It is therefore incumbent on me to demonstrate that, with respect to self-consciousness proper, relationalism is the preferable position. The

peculiarities that Kriegel refers to are such matters as the Essential Indexical, the Elusiveness Thesis, Immunity to Error – and many other classic curiosities and debates regarding the nature of self-consciousness. In section 2.3 I show how the Fundamental Dichotomy permeates all of these, arguing that there is a correspondence between all the debates on self-consciousness such that each argument fits into one or the other side of the Fundamental Dichotomy. During this exposition I argue for the relationalist position over the alternative (intrinsicist) view. The breadth of this section is wide, encompassing many apparently diverse aspects of self-consciousness. However, my intention is not to examine each aspect of self-consciousness in depth but only in sufficient detail so as to show how it maps onto the Fundamental Dichotomy and also to defend the relationalist position. Nevertheless, if all goes well, it should help meet another Kriegel [2007] challenge: to determine which of the alleged peculiarities in fact obtain. The result will not be an exhaustive general theory of self-consciousness but may at least form the kernel of one.

What emerges from this discussion is the importance of concept possession for self-consciousness. This idea is developed in the next chapter, where I argue that concept possession is, in effect, the defining characteristic of self-conscious organisms.

**The Fundamental Dichotomy**

On one side of the Fundamental Dichotomy the nature of self-consciousness is seen as *intrinsic*; on the other it is viewed as *relational*. I refer to these positions as *intrinsicist* and *relationalist* respectively, although the relationalist position should be seen primarily as anti-intrinsicist. That is, if a position is not intrinsicist, then it is necessarily relationalist. When an organism experiences certain types of mental states we can easily think of these as separate things: the mental state and the subject experiencing that mental state. For example, an organism may be undergoing a visual perception. In this situation, the two things stand in a certain relation to each other; the subject is *experiencing* the visual perception. This is what I mean by 'relational' in the context of the Fundamental Dichotomy: there are two separate entities (the subject and the object – in this context a mental state) standing in a relation to each other.

Most of the time there is no problem in conceiving a relation between a subject and the subject's mental states – thoughts, perceptions, etc. But when we consider a state of self-consciousness the situation is not so clear cut. The phenomenology of self-awareness seems fundamentally different to awareness of non-self things. We may be aware of (say) objects in the world through perceptions, but our awareness of our selves seems to occur without perception. Indeed, it may even seem that our awareness of our selves exists constantly 'in the background' no matter what other mental activity we may be engaged in. Self-consciousness can thus be construed as an intrinsic *way of being* rather than as a separate mental state or property of a subject. This is what drives the *intrinsicist* side of the Fundamental Dichotomy.

The relationalist view of self-consciousness would have it that there is a *relation* between (say) a self-referring thought and the self-conscious organism *having* the thought, while the intrinsicist view would consider the self-referring thought as inseparable from (i.e., intrinsic to) the self-conscious organism (i.e. the self-referring thought is what *constitutes* the self-consciousness of the organism). What I argue in section 2.2 is that self-conscious organisms are distinguishable from non-self-conscious organisms in that the former are able to perform certain mental feats; specifically, the act of introspection[6]. I describe this as a kind of internal dialogue or interrogation, in which the self-conscious agent acts in a dual capacity, as the interrogator and the interrogated. That is, the agent is not self-conscious *intrinsically*, as a way of being, but by having the ability to introspect, which I see as *relational* (a relation between the introspector and the introspected). I argue that the act of introspection requires conceptual capacities, and that these capacities are unavailable to the non-self-conscious agent.

Even though many of the streams in the philosophy of self-consciousness might appear distinct and unconnected, I believe there is a correspondence between them in that they all in some way manifest the Fundamental Dichotomy. Call this my Correspondence Thesis. The various debates in self-consciousness are far too broad to all be discussed in detail here but I present sufficient background to explicate the main issues involved and to enable me to carry out my aim, which is to map them onto the Fundamental Dichotomy (however, as a test case, I have singled out one topic for somewhat deeper examination, that of the essential indexical). I will present each argument as falling into one of two opposing camps: the

[6] Later I define my usage of 'introspection' more clearly and compare it with other conceptions.

intrinsicists and the relationalists. So, for example, while *immunity to error through misidentification relative to the first person pronoun* seems to have no bearing on, say, *the essential indexical*, in fact I suggest that there is a correspondence: in each case, one side of the argument will support the intrinsicist view, while the other side will support the relationalist view.

## 2.2 Access to the Self

A central issue in regard to self-knowledge is *mode of access* – that is, the *way* in which we gain self-knowledge. It is significant that there is a lack of consistency in describing exactly *what* we are accessing to gain this self-knowledge. Authors write variously of accessing the *mind* [Ryle 1966/1994], *mental states* [Armstrong 1968/1994; Davidson 1987/1994; Rosenthal 1986], *thoughts* [Burge 1988/1994], the *self* [Shoemaker 1968/2001; Chisholm 1969/1994; Evans 1982], or even '*internal psychological organization*' [Van Gulick 1988]. Sometimes, a combination of terms is used in the same piece. Despite this, there is some commonality in what all these authors are talking about. Self-knowledge is gained by access to the *self* in terms of the content of the *mind* – which is to say the content of one's *thoughts* (which are themselves *mental states*). Later I argue that access to certain types of mental states (e.g., perceptual states) does not count as self-consciousness. For now though, we can assume that these authors have broadly the same general intention - despite the inconsistency in the precise description of *what* is accessed - which is to explain the *nature* of this access.

### Perceptual Model versus Privileged Access

Many of the authors mentioned above believe we have direct, unmediated access to our selves [Davidson, 1987/1994; Burge, 1988/1994; Van Gulick, 1988; Rosenthal, 1986, 2004; Shoemaker, 1968/2001; Chisholm, 1969/1994; Evans 1982]. In these views, there is no need to look at oneself 'from the outside'. We are simultaneously *in* mental states and *conscious* of them. We are conscious of our thoughts just by thinking them. This view of self-access, that it is unmediated, is intrinsicist in that there is no separation between the thought and the act of thinking or between the thought and the thinker. The view is commonly known as the *privileged access model* of self-consciousness.

Opponents of the privileged access model argue that a thought cannot be its own object. According to Ryle [1966/1994] self-consciousness should not be "described as a torch that illuminates itself by its own light" (p39). Armstrong [1968/1994] put it this way: "A mental state cannot be aware of itself any more than a man can eat himself up" (p110). According to Armstrong, although the objects of perception are situations in the physical world while objects of introspection are current happenings in our own mind, introspection may still properly be compared to sense-perception. Armstrong goes on to say that although the introspecting and the thing introspected are both mental states, it is impossible that they should be one and the same mental state. Ryle and Armstrong are correct: as argued in more detail later, access to the content of a mental state requires the existence of some additional higher order thought whose object is the mental state in question.

The antithesis to the privileged access model is sometimes referred to as the perceptual model, because without the benefit of unmediated access, one must somehow *perceive* the inner self, perhaps by the action of an 'inner sense' analogous to the traditional five 'outer senses'. In such models, which fall into the relationalist side of the Fundamental Dichotomy, the mode of self-perception is much the same as that for the perception of others:

> There are respects in which it is easier for me to get...knowledge about myself than to get it about someone else; there are other respects in which it is harder. But these differences of facility do not derive from, or lead to, a difference in kind between a person's knowledge about himself and his knowledge about other people. [Ryle 1966/1994, p31]

Shoemaker [1968/2001] and Davidson [1987/1994] mount assaults on the perceptual model based on the apparent asymmetry, touched upon by the Ryle quote above, between access to our own mental states and those of others. Shoemaker claims there is a sense in which "my self is accessible to me in a way in which it is not to others." Davidson agrees that although we can treat our own thoughts the way we treat others', the reverse is not true. This asymmetry appears to pose a problem for the perceptual model if we maintain that perception of the self is analogous to our perceptions of others. However, I do not think the asymmetry needs an explanation. Access to the self need not be *perfectly* modelled on sense perception to be a valid alternative to privileged access. The self can be perceived by the self, with a conceptual gap between perceiver and perceived. But perceiving one's inner mental states must be somehow different to perceiving anything in the world. These are different actions for two very different objects of perception. One can agree with Shoemaker [1986/1994] that

there is no ‘organ of introspection’ without acceding to a conclusion that therefore inner sense cannot exist. After all, our sense organs exist to receive signals from the outside world and retransmit them in a different form to our brains for processing. If the signals originate from the brain itself, as is the case for mental states, then a dedicated organ – other than the brain itself – is not necessary for detecting and retransmitting those signals.

**Introspection**

Access to the self for the gaining of self-knowledge, can be encapsulated in a single word: *introspection*. My central thesis in this section is that this type of self-access is a deliberate *action* and not a *state of being*. Some authors prefer to use ‘introspection’ to refer to self-access that is not conscious or deliberate [Engelbert & Carruthers 2009; Carruthers 2010]. An objection to my usage might, for example, take the form “but I can become *involuntarily* aware of my mental tiredness.” This is indeed the case, but I do not think this type of self-access should be referred to as introspection; rather it indicates what I call (after Gallagher & Zahavi [2006]) *self-givenness*, as discussed below. One can have pre-reflective access to certain types of mental states, but to gain self-knowledge (knowledge of self *as such*), one must deliberately reflect on oneself. It is in this latter sense that I use ‘introspection’, but in any case the main point to be made is the distinction between these two modes of self-access (reflective and pre-reflective), which reflects the Fundamental Dichotomy.

My model for self-consciousness is the human being. That is not to say that non-humans are precluded from self-consciousness, but that it is the type of self-consciousness we humans are familiar with that is under discussion. In humans gaining self-knowledge is a deliberate act of self-examination. By contrast, there are other forms of self-access for certain types of internal states that need not lead to self-*knowledge* as such. It is uncontroversial that at least some animals that we would not consider self-conscious have access to their internal states such that those states can motivate behaviour. That access need not be deliberate; for example, a state of hunger might be experienced and may motivate certain behaviour without the subject deliberately paying attention to its internal state. Indeed, the behaviour of a carpet-cleaning robot may be similarly driven by its (obviously non-deliberate) access to its internal state. Of course, humans also sometimes exhibit that kind of non-deliberate access to internal states, but humans have the further capacity to gain knowledge of the self *as such*. By this I mean

conceptual understanding of the self, i.e., a self-concept. Conceivably, I may have access to my *current activities* in a non-conceptual way; I may have access to my *emotional state* in a non-conceptual way (in the same way as, say, a mouse represents that it is in a state of fear); I may have access to my *current perceptions*[7] in a non-conceptual way; but to have access to my own mental state in a way that signifies self-consciousness it must be a state that is about the self in a conceptual way. For example, I may be in a state of fear and, like a mouse, I may represent in a non-conceptual way that I am in a state of fear. But if I am asked *why* I am in a state of fear I will need to have a concept of fear in order to comprehend the question (and the answer). And to discover the answer, I will need to introspect. I will need to perform the action of turning my attention inward, onto my self, in order to gain conceptual information about my internal state. Thus, introspection is exemplary of self-consciousness in action.

On this conception of introspection, the privileged access model is inadequate, for it likens introspection to a *state of being:* self-knowledge is not gained through an action, it exists as an intrinsic property of the self. If introspection is an action it requires a subject to perform the action upon an object, implying a relation between subject and object. The perceptual model best fits this conception, as it implies the existence of a subject (the perceiver) and an object (the self perceived), even if they are both one and the same self. An action requires a separation between agent and object. In the case of introspection the agent is the subjective self (the thinker) and the object is also the subjective self, but, to use Kriegel's [2004] terminology, there is a *conceptual gap* between the self-as-subject (the agent) and the self as object of introspection. Introspection requires the application of conceptual thought. It is a type of interrogation, if you will, of the self in which inferences about the self can be made. To appropriate a quote from Dennett [1991]: "If I couldn't talk to myself, I'd have no way of knowing what I was thinking" (p315) [8].

To summarise my point, we are self-conscious organisms because of our ability to perform the *action* of introspection, not by simply *being* in a 'self-conscious state'. Self-consciousness by nature is not intrinsic – it is not a state of being. It requires the application of conceptual thought. The view that one can be conscious of oneself simply by being conscious *at all* has been referred to as *self-givenness* [Gallagher & Zahavi 2006]. In this view one does not need

[7] I argue the case for the existence of non-conceptual perceptual states in chapter 3.
[8] In using this example (and the previous example of being *asked* about fear) I do not mean to imply that concept possession requires language. I remain neutral on whether this is or is not the case. The language issue is addressed in chapter 4.

a deliberate act of introspection to be aware of one's existence – that knowledge always seems to be there 'in the back of our minds.' I can agree with this idea of a 'background self-awareness' up to a point, but I argue below that this so-called self-givenness is not the same as self-consciousness.

**Self-Givenness**

Sellars [1956/1997] famously attacked what he called the 'myth of the given', where 'given' here means foundational (i.e. epistemically independent) knowledge – in other words, unanalysable conceptual mental content. Givenness in Sellars' sense has been described as 'self-presenting, self-authenticating inner episodes' (e.g., see Parsell [2011]) and accordingly I concur with Sellars in dispelling this intrinsicist conception. But there is a type of givenness that does not depend of concept possession. This idea of self-givenness, as usually espoused by phenomenologists, can be viewed as a sort of 'always on' access to the self but it is a pre-reflective, non-conceptual self-access [Gallagher & Zahavi 2006] and hence does not satisfy my notion of self-consciousness. Consider (say) a dog. A dog is, in a sense, 'aware' of some of its mental states, for example, that it is hungry or experiencing pain. It can also be aware that it is eating. Gennaro [2009] rightly remarks that we should not withhold concepts such as PAIN and DESIRE from animals given that they have even a *partial* understanding. But in animals it is difficult to determine even a partial *understanding*: can we ascribe the concept DESIRE to a dog just because we observe that it seeks food (as opposed to simply experiencing hunger)? What about a cockroach? Does the carpet-cleaning robot desire to clean carpets? In a certain sense a dog may be 'aware' that it is eating but this is probably in a non-conceptual way – it does not necessarily know what eating is. Plausibly, a dog is 'aware' of its own interactions with the world in the same way that it is 'aware' of (is able to detect) affordances [Gibson 1979]. Affordance does not rely on concept possession[9]. It is likely that there are some organisms, perhaps including dogs, that can *only* be aware of their interactions with the environment in this non-conceptual way.

[9] "You do not have to classify and label things in order to perceive what they afford" [Gibson 1979, p134]. Categorisation is a key element of conceptualisation, so the preceding quote implies the possibility of non-conceptual perception of affordances.

This non-conceptual awareness of some types of mental states is how I view the notion of self-givenness. I used animals as examples to illustrate the point, but humans, who we know are capable of self-consciousness, also have this givenness, this pre-reflective self-access. A human does not need to reflect upon an experience of pain to experience the pain. But an organism that is *only* aware of itself in this non-conceptual way cannot be aware of itself *as a self*. To be aware of itself as a self the organism must have a *concept* of the self. There is no reason to believe that just because Sartre [1943/2000] knew, when asked, *that* he was counting cigarettes that this implies he had self-consciousness. Of course we already know he *did* have self-consciousness, being an adult human, but this is not a conclusion to be drawn simply by the givenness of his access to this type of mental state (the state of performing a certain action). This constant background non-conceptual self-access does not constitute self-consciousness as I defined it in chapter 1.

Some authors, as in the examples below, make the mistake of equating self-consciousness with access to *any* mental state. The error more easily arises in authors who study the nature of occurrent self-consciousness rather than the threshold of self-consciousness. These authors examine self-consciousness in organisms we already take for granted as being self-conscious: adult humans. So, they are starting from a position that already assumes the existence of self-consciousness. The question for them is whether human self-consciousness is best modelled on privileged access or on sense perception. For adult humans access to mental states is access to a self-conscious organism's mental states and it is tempting to consider any such access as sufficient for a description of self-consciousness. Thus, when Chisholm [1969/1994] asks if a man might be aware of himself as experiencing without thereby being aware of himself, his answer is no: "…in being aware of ourselves as experiencing, we are, *ipso facto*, aware of the self or person – of the self or person as being affected in a certain way" (p105). And to Burge [1988/1994], "Knowing one's thoughts no more requires separate investigation of the conditions that make the judgement possible than knowing what one perceives" (p72). It is implicit in these statements that they are referring to adult human beings. But when we are talking about organisms that might or might not have self-consciousness (animals or human infants) we must tread more carefully. While in these cases we cannot assume the existence of self-consciousness, we must still consider that in all likelihood many species can nevertheless be aware of (for example) their phenomenal experiences. It is not difficult to conceive of a dog experiencing pain or tail-wagging joy. However, having access to only this type of mental state need not imply self-consciousness.

What is required for self-consciousness is access to the self as such – access that resolves into conceptual self-knowledge. The 'self-givenness' we have is not a givenness of *self as such* - it is just a givenness of certain types of non-conceptual mental states, the content of which might happen to involve the self (for example, an emotional state).

Consider this view from Van Gulick [1988]:

> Just as an organism can acquire and apply information about its external environment through a reflex-mechanism without having a perceptual experience or being able to report what it detected, an organism can also acquire and apply information about its internal psychological organization without having an inner-directed experience or being able to communicate that information publicly. (p163)

The view expressed here is an argument for the privileged access model of self-consciousness: an organism acquires information about its internal psychological organisation without an inner-directed experience. Thus Van Gulick here denies the perceptual model of self-consciousness. But if we accept my point about the difference in types of mental states, and that direct access to the type mentioned by Van Gulick is not of the type that supports self-consciousness, then we can see his assertion in a different light. What Van Gulick intends is the identification of self-consciousness with possession of "reflexive meta-psychological information." But the analogy Van Gulick actually makes here could easily be interpreted as an organism's detection of Gibsonian affordances (the acquisition and application of information "about its external environment though a reflex-mechanism"). He likens this ability with what I would call self-givenness – direct access to "internal psychological organisation." In that context I agree with this view; as previously mentioned I accept this level of (non-conceptual) self-givenness. My disagreement with Van Gulick is that, unlike him, I do not accept that this indicates self-consciousness in the organism.

We can appropriate Van Gulick's analogy to support my point about access to the self. According to Van Gulick an organism can access its own internal psychological organisation in the same way as it acquires and applies information about its environment. Of course, Van Gulick was arguing that this access could be via a reflex-mechanism and therefore not a deliberate action. And, as argued above, I accept this insofar as the internal mental states accessed are devoid of conceptual content (and therefore not representing a case of self-consciousness, since they must therefore be devoid of a *self*-concept). But we can extend the

analogy to an organism that *is* self-conscious and is interacting with its environment in a *conceptual* way – meaning, not in the mechanistic, non-conceptual mode of affordance detection. In other words, although some organisms may interact with their environment in a non-conceptual way (via affordance detection), other organisms (such as humans) can also interact with their environment in a conceptual way. That is, they can *understand* aspects of their environment. A self-conscious organism must be concept-bearing (as it has at least a *self*-concept) and is therefore able to make judgments about its environment, make inferences about it. And just as it can do that with its external environment, it can do it with its 'internal psychological organisation'. Viewed this way, Van Gulick's analogy is consistent with the view that while there can be direct, unmediated access to certain types of mental states (such as emotions or raw perceptions) that do not involve conceptualisation and do not indicate self-consciousness, there can *also* be access to mental states in a *conceptual* way to make inferences about the self, and this does indicate self-consciousness.

To summarise the main points so far: the constant 'always on' background access to the self (self-givenness) does not represent self-consciousness proper. It is a capacity (possessed by both self-conscious organisms as well as some non-self-conscious organisms) for access to some types of (non-conceptual) mental states such as raw emotions or perceptions. Self-consciousness proper requires the ability to gain self-*knowledge*, meaning a conceptual understanding of the self's mental states. I note briefly now that this view is similar to some versions of the 'dual reasoning' hypothesis (i.e. the system 1/system 2 dichotomy [Sloman 1996] and, perhaps, the personal/subpersonal dichotomy [Frankish, 2009]) yet potentially different in an important way. System 1 is described as a more primitive faculty of 'reasoning' that is intuitive, automatic and associative (among other characteristics) while system 2 is reflective and controlled (etc.) [Carruthers 2012][10]. Thus, system 1 seems to parallel self-givenness and system 2 parallels introspection. However, the primary distinguishing factor separating the two sides of the Fundamental Dichotomy is *concept possession*. As argued further below, introspection, which is a hallmark of self-consciousness, involves concept possession while the more primitive self-givenness does not. If system 1 is truly a *reasoning* system and a faculty for *belief*-formation [Carruthers 2009c; 2012] then it involves concept possession and therefore would not quite match my notion of self-givenness. However, system 1 could rather be viewed as the faculty subserving non-

[10] Carruthers does not completely endorse the system 1/system 2 dichotomy, arguing instead that system 2 reasoning is realised in cycles of system 1 operations.

conceptual dispositions to behave in certain programmatic ways, and if so the dual reasoning hypothesis is a reasonable reflection of the Fundamental Dichotomy. The same may be said for some formulations of the personal/subpersonal distinction (i.e. introspection reflects the personal level and self-givenness reflects the subpersonal) and indeed on some views system 1/system 2 maps directly onto subpersonal/personal (e.g. Frankish [2009]). Naturally, the same caveat applies: formulations that appear to attribute conceptual abilities to subpersonal systems (e.g. Frankish, [2009]) would be inconsistent with CPH.

**Introspection and Concept Possession**

There is a close relationship between the relationalist view of introspection and concepts in that there can be no introspection for an organism that is not concept-bearing. Consider this passage from Seager [2001]:

> ...the evident fact that animals and children...are incapable of introspection is neatly explained by their lack of the proper field of mentalistic concepts...[t]hey consciously perceive the world, but they don't know that that is what they are doing and it is this lack of conceptual machinery that precludes introspection (p260).

Introspection in this view requires conceptualisation. Self-knowledge is acquired by making inferences about the self based on self-perception. Inferential thought requires concepts, so an organism that is not concept-bearing is unable to perform introspection. Such an organism only has non-conceptual access to its own mental states, perhaps in an analogous way (as in the earlier Van Gulick analogy) to how it interacts in a non-conceptual way with its environment.

The intrinsicist view of self-consciousness, by contrast, does not rely on the existence of concepts. In the intrinsicist view, based on the privileged access model, an organism can be self-conscious without performing an *act* of introspection, since access to the self on this view is immediate and unmediated. Here introspection is erroneously considered a *state of being* rather than an action. In this view there is no need for concepts – there is no inferential thought taking place; just the mental states themselves *being* self-conscious. As argued earlier, an organism may have direct and unmediated (and non-conceptual) access to some of its mental states (such as emotions or perceptions of pain, etc.), but this should not be

considered self-consciousness proper. Of course, an organism that is self-conscious need not always be introspecting and may *also* access those mental states in a direct and unmediated way. In other words, like the non-self-conscious organism, it can also undergo experiences in which concepts are not involved. My claim is that organisms that are conscious but *not* self-conscious can *only* access their mental states in this way.

### 2.3 The Correspondence Thesis

I have claimed that the relationalist side of the Fundamental Dichotomy is the correct position and that intrinsicism is incorrect as applied to self-consciousness. Intrinsicism is true only of the so-called primitive form of self-consciousness, which I discount as a case of self-consciousness proper. In this section I examine how this position measures up against the many and sometimes curious putative properties of self-consciousness. What emerges is a correspondence between the opposing views of self-consciousness on the one hand, and the Fundamental Dichotomy on the other. I structure this section into subsections addressing, firstly, some established dichotomies. In these there is a clear opposition between two sides (as we saw with the Perceptual Model versus the Privileged Access model). I show how these dichotomies correspond to the Fundamental Dichotomy. I then address some of the so-called 'peculiarities' of self-consciousness, such as *immunity to error through misidentification relative to the first person pronoun*. In these discussions I show how the phenomenon in question supports one or the other side of the Fundamental Dichotomy. Every case either does or does not involve a relational aspect; those that do not can be called *intrinsicist* while those that do will be considered *relationalist*. I then present arguments that the intrinsicist cases are incorrect – that is, that they do not in fact obtain in the relevant way. I aim to show that the relationalist conception of self-consciousness forms the basis of a self-consistent framework.

I am primarily classifying the *views* themselves rather than their proponents. Philosophers might never before have looked upon self-consciousness in terms of this Fundamental Dichotomy and will therefore not consider themselves either intrinsicists or relationalists. However, they will naturally fall into either camp based on how their views map onto the Fundamental Dichotomy.

**Transitive versus Intransitive Self-Consciousness**

I begin with the distinction made by Kriegel [2007] between transitive and intransitive self-consciousness because it is paradigmatic of the Fundamental Dichotomy. Here is Kriegel's own explanation:

> Compare "I am self-conscious of thinking that *p*" and "I am self-consciously thinking that *p*". In the former, transitive form, self-consciousness is construed as a relation between me and my thinking. In the latter, intransitive form, it is construed as a modification of my thinking...The adverb "self-consciously" denotes a way I am having my thought that p. No extra act of self-consciousness takes place after the thought that p occurs. Rather, self-consciously is how the thought that p occurs. [Kriegel 2007, sec 2]

What Kriegel calls the 'intransitive form' of self-consciousness above is an example that fits the intrinsicist view. According to Kriegel, in this case the thought is itself the state of self-consciousness – there is no gap between the thought and the state. By contrast, the 'transitive form' implies a relation between two numerically distinct mental states: the thought and the state of being self-conscious. Here there is an implied gap and one entity stands in a relation to the other.

The issue at hand is whether the transitive or intransitive characterisation is correct, and this depends on the interpretation of the two sentences in the quotation above. In a similar example, Chisholm [1969/1994] describes the sentence 'I have a depressed feeling' as *equivalent to* 'I feel depressed'. Chisholm thinks that the former statement gives an erroneous impression that there are two entities, one of them had by the other, but that the true sense is derived from the latter statement, which describes a certain undergoing. Thus Chisholm appears to be expressing a preference for the intransitive characterisation – an intrinsicist view. On the other hand, given the equivalence in meanings, it is just as arguable that (as I maintain) the *former* statement provides the correct sense, self-consciousness is transitive and thus relational in nature.

In characterising this (and later, other) distinctions as a dichotomy, I am implying that only one or the other should hold; these are not merely different ways to describe the same aspect of self-consciousness or different aspects that could both obtain. One is committed to one or the other point of view. I argue below that self-consciousness is transitive and there is no

such thing as intransitive self-consciousness[11]. Burge [1988/1994] by contrast, is apparently an intrinsicist, for he claims that there is no way to create a gap between what we think and what we think about - the object of reference just is the thought being thought. In the context of self-conscious thought, this seems to map onto the intransitive view. There is the *state* (of thinking) and then there is the *thought* itself (the object of reference, which must represent the self in the case of self-conscious thought), but Burge sees no gap between these entities, considering them one and the same. From this we might legitimately conclude that Burge would consider self-consciousness as intransitive in nature to the exclusion of the possibility of transitive self-consciousness. Kriegel himself, however, would not agree that the distinction between transitive and intransitive self-consciousness implies that only one or the other can hold, as this passage seems to illustrate:

> ...transitive self-consciousness is involved in cases where the subject is focally aware of being in [mental state] M, whereas intransitive self-consciousness is involved in cases where the subject is only peripherally aware of being in M. [Kriegel 2004]

To Kriegel *transitive* and *intransitive* are two different possible forms of self-consciousness. There seems to be a significant difference between two types (or levels) of self-consciousness in this passage: being *focally aware* and being *peripherally aware*. This could be interpreted as implying that being focally aware of being in a mental state means truly being aware of the *self* (i.e. the self as the focus of attention). That characterises the transitive form as self-consciousness proper. But being peripherally aware seems very much like what I described earlier as self-givenness. As argued in section 2.2, self-givenness should not be considered a form of self-consciousness as it is not consciousness of the self *as such*. Accordingly, 'intransitive self-consciousness' should not be taken to imply consciousness of the self *as a self*, but might be taken to mean the more 'primitive' form of background self-access described earlier as self-givenness.

[11] I will, however, allow 'intransitive self-specifying mental states', and it could be that this is what is sometimes meant by the term 'intransitive self-consciousness' (as I think the upcoming passage by Kriegel illustrates). But if so, the usage of self-consciousness here is a very weak sense, akin to what Bermũdez [1998] calls a 'primitive' form of self-consciousness. As discussed in chapter 1, this conception of a primitive form self-consciousness does not fit within my usage of self-consciousness. In chapter 1 I gave my reasons and justifications for my definition of self-consciousness.

**Self-as-Subject versus Self-as-Object**

A distinction is often made between consciousness of self as either subject or object. Usage of these terms is not always consistent between authors so we need to step carefully through the argument. In the first place it is necessary to define what is meant by 'subject' and 'object' in this context. In one usage consciousness of self-as-subject is characterised as consciousness of oneself as a psychological subject, as the thinker of thoughts and experiencer of phenomenal experiences, as compared with consciousness of oneself as an object in the world, as a flesh-and-blood object. Thus Shoemaker [1968/2001] gives an example of self-as-object usage in language as 'that is my leg I see in the mirror' (i.e. a reference to the physical body) and an example of self-as-subject as 'I feel a pain in my leg' (referring to the subject undergoing the experience). In this usage there is no dichotomy; one can be aware of both one's physical body (or body parts) and also be aware of oneself as a psychological subject – these are two different *things* to be aware of. What I want to explore is the usage in which we are only discussing awareness of *one* thing: the self *as such* – as a psychological subject – and how this may be viewed in two different *ways*. For this we need to consider another sense of 'object': not as a reference to a *physical* entity but rather as the object of a thought. Think of this usage as analogous to the linguistic usage of *subject* and *object* in proper sentences.

According to Kriegel [2007] there are two modes of presentation under which a subject may be conscious of herself. In one the subject thinks of herself as the thinker (self-as-subject); in the other she also thinks of herself, but *not* as the thing doing the thinking (self-as-object). Kriegel says in the latter case there is a 'conceptual distance' between the thinker and the object of the thought (i.e. the self), while in the former case there is not. This view brings us a little closer to mapping this issue onto the Fundamental Dichotomy. We can characterise consciousness of self-as-object, where there is a conceptual distance or gap between thinker and object-of-thought, as falling into the relationalist camp. Consciousness of self-as-subject, where according to Kriegel there is no such gap, falls into the intrinsicist camp. But for this to be valid as a dichotomy the two have to be mutually exclusive. To clarify where the dichotomy exists, then, consider the two sides in the following way. Let us consider the self-as-subject side of the dichotomy as implying the *absence* of awareness of self-as-object. In terms of the Fundamental Dichotomy then, an intrinsicist would argue that consciousness of the self is as self-as-subject to the exclusion of self-as-object (call this self-as-subject-*only*), while the relationalist would argue *either* that (1) consciousness of the self is as self-as-object

to the exclusion of self-as-subject, *or* that (2) the two modes are not mutually exclusive and that consciousness of the self must include awareness of the self-as-object. I argue below for alternative (2).

The following is an example of the self-as-subject-*only* view. Brook [2001] claims that awareness of self-as-subject is unique in being not 'experience-dividing': there is no way to distinguish the self presented to oneself in any representation (as there is only ever one and the same self with nothing other than that self to compare with). From this he concludes that when one appears to oneself as oneself it is *not as the object* of a representation. "To represent something as an object is to place it vis-a-vis other objects, and usually to ascribe properties to it. If so, to appear in a representation as subject is *not* to appear as an object of any kind..." (p25; my emphasis). This looks like a denial of the possibility of consciousness of self-as-object, in which case it is a position favouring the intrinsicist side of the Fundamental Dichotomy.

I argue that in a certain (important) sense all intentional mental states must have an object in an analogous way to grammatical sentences: a sentence must always have a subject that is the 'doer' and an object that is acted upon. In the case of self-conscious thoughts, the object is the self, so awareness of self-as-object *must* always hold. That does not mean that the subject is unable to think of herself as *being a subject*. When a subject has a thought about herself as such, she becomes the object of the thought – that is, she is both the subject (the thinker) and the object of the thought. Thus, I argue against a separate and distinct notion of consciousness of self-as-subject taken in isolation from self-as-object, in the way Brook describes. While it may be true that one can think of oneself as a thinker, and as the thinker of the thought being had (that is, a subject), the thinker is still the *object* of the thought. Although the object of such a thought might be the self (a thinker, a subject) it is still always true that what is thought about is the thought's object. Thus there is never escaping the fact that the self in self-conscious thoughts is self-as-object, even if the self is thought about as being the subject having the thought.

In summary, there can never be a 'self-as-subject' thought such that it is therefore *not* a self-as-object thought. Thus there are no self-conscious thoughts where there is no 'conceptual distance' between the thought and the thinker. One may know oneself to be a subject, be able to think of oneself as thinking, and know that one is thinking of oneself, but this in no way eliminates the implied separation between the thinker and the object of the thought. This view

fits into the relationalist side of the Fundamental Dichotomy and comes into play throughout the discussion as to the nature of self-consciousness.

**Unitary versus Non-Unitary Self**

An intrinsicist might argue that if the self is unitary then there is no room for a gap between self as subject and self as object. Rather than two numerically distinct entities bearing a relation to each other, these might be described as two descriptions of one and the same entity – the unitary self. The relationalist might get away with emphasising the notion of this gap being only *conceptual* in nature, as discussed in the foregoing subsection: there is no need for an *actual* distance between the self as thinker and self as thought about. However, even without using this rebuttal, there is evidence that the self might *not* be unitary, which would undermine the intrinsicist position further.

We humans tend to believe of ourselves that we are a single, psychologically unified entity persisting over time. This certainly appears to be the case with regards to multiple concurrent experiences (phenomenal mental states): we can be simultaneously aware of visual, auditory, tactile and olfactory experiences. Along with these perceptual experiences are "...proprioceptive experiences, experiences of agency, affective and emotional experiences, and conscious thoughts of various kinds" [Bayne 2004, p219]. When experienced together, these all appear phenomenally unified. Bayne uses the term 'co-consciousness' to express the phenomenon of unified conscious experience. So it seems true that there is a unified consciousness in our phenomenal experience as there appears to be only a single unified subject of experience at any one time. This seems to support a view of self as being unitary and therefore seems to fall into the intrinsicist side of the Fundamental Dichotomy.

But this *experience* of a unified consciousness need not be inconsistent with the relationalist view. For instance, it could be true that the self is divisible into separate co-existing entities and it might be that only one of these entities is capable of 'inhabiting the subject' at any given time – that is, of becoming the single subject of experience. It may be, as I explore below, that the entity inhabiting the subject at any one time is capable of perceiving another co-existing entity that potentially *could* inhabit the subject. These entities may be viewed as multiple manifestations of self, each of which is a potential subject – that is, each of which is

capable of becoming *the* self – the subject of consciousness. This non-unitary view of self fits under the relational side of the Fundamental Dichotomy. In this case there is a way for a self to 'relate to itself' in that the single subject of consciousness might become aware of another entity that is a separate manifestation of the self. Later, I briefly describe a model of self-consciousness that may fit this conception: Dennett's Multiple Drafts Model. But first, let's examine some empirical evidence that supports this view.

The phenomenon of split brains presents evidence that selves can be split. Specifically, the indication is that *consciousness* can be split, which undermines the notion of the unity of self. Split brains refer to epilepsy patients who have had their corpus callosum severed thereby separating the left and right hemispheres of the brain. Although in everyday activities these individuals typically appear normal, under some experimental conditions they exhibit behaviours as if there were two separate consciousnesses each having control over different parts of the body [Nagel 1971]. This gives new meaning to the expression 'the left hand doesn't know what the right hand is doing'. More prosaically put, different perceptual inputs fed separately into each brain hemisphere elicit independent and conflicting responses. It is tempting to think the mind itself has been split although the two halves are still able to coordinate normal behaviour. Nagel claims that it is our understanding of the unity of consciousness that is at fault. According to Nagel we are unable to ascribe any whole number of individual minds to the subjects of the experiments, and as such this calls into question the concept of a single subject of consciousness as it applies to ordinary persons. These experiments demonstrate that multiple consciousnesses are possible within the same individual, although only one of these can reach the level of 'subject'. That is, they represent entities that can potentially be the subject of conscious experience (inhabit the subject) but only one at a time. Thus, the experiments show that separate consciousnesses are able to exercise control over the body, but there is still only one *experienced* consciousness by the subject. Another pathological case to consider in this context is a phenomenon that was once referred to 'multiple personalities' in which apparently distinct personalities take control of the individual's behaviour at different times. This is now known as Dissociative Identity Disorder (DID), characterised by lack of a single coherent personality. DID would seem to be a paradigm example of the non-unitary self where in this case alternative selves quite *literally* inhabit the subject.

In ‘hidden observer studies’ [Kirsch & Lynn 1998] subjects under hypnosis have been observed to behave in ways suggesting a division of consciousness within an individual. In one example of the technique a subject was first placed into a hypnotic state and instructed through the hypnotist’s suggestion to be (hypnotically) deaf. Once the subject’s deafness was confirmed using standard procedures (lack of startle responses to loud noises) the hypnotist accessed a ‘hidden observer’ within the subject’s psyche by saying “although you are hypnotically deaf, perhaps there is some part of you that is hearing my voice...”. Sure enough the subject confirmed this to be the case. The experiment has also been successfully conducted for cases of analgesia to experimental pain (the hidden observer felt the pain) and negative hallucination (the hidden observer saw the object that the subject was hypnotically blind to). These experiments suggest at least the capacity for multiple consciousnesses within the same individual. It is almost as if one self has inhabited the subject that is under hypnotic trance but another is still accessible to the hypnotist nonetheless. The results are of course open to many interpretations. However, the salient feature of these results as far as the Fundamental Dichotomy is concerned, is the separability of the entities involved – the hypnotised subject and the hidden observer. These entities are separate, or separate *enough*, to allow an interrelationship, thereby providing support to the relationalist side of the Fundamental Dichotomy.

Yet we need not only depend on pathological or abnormal cases to find evidence of split selves. Even we normal human adults are wont to exhibit behaviours indicative of a lack of unity of the self. Who has never wrestled with their conscience? When that happens, who are we fighting with? Yes, the glib answer is ‘oneself’ but the point is that a tussle requires more than a single participant – even an internal conflict with oneself. There is a kind of internal competition here in which two entities interrelate. What does one make of self-deception – who is deceiving whom? In these examples, even normal human adults experiencing themselves as a single subject at a particular point in time must admit to perceiving themselves manifested as an ‘other’ self. Gilbert Ryle describes the dawning of this duality of self as follows:

> At a certain age the child discovers the trick of directing higher-order acts upon his own lower-order acts. Having been separately victim and author of jokes, coercions, catechisms, criticisms and mimicries in the interpersonal dealings between others and himself, he finds out how to play both roles at once. He has listened to stories before, and he has told stories before, but now he tells stories to his own enthralled ear...He finds that he can give orders to himself with such

> authority that he sometimes obeys them, even when reluctant to do so. [Ryle 1966/1994, p38]

This passage supports the view that a single, apparently unified, self can still manifest a duality between which there exists a relation.

In his defence of the representational theory of mind, Seager [2001] divides the self into two: a 'secret self' and a 'constructed self'. The secret self is 'invisible to the subject' and known only through postulation, while what we perceive as our selves is the constructed self. Seager describes the secret self in the way I might describe the 'self inhabiting the subject': "...it might be thought of as the *centre* of my world, utterly invisible because everything is seen from its vantage point" (p256)[12]. Seager goes on to say "...we should call the self we know by introspection the *constructed* self and [this] reveals the *relation* between the two selves" (p257, second emphasis mine). Seager's ideas are similar to mine in this respect and falls firmly into the relationalist side of the Fundamental Dichotomy.

**Multiple Drafts Model**

Even though we experience a *sense* of unity of self, at any one time there may be multiple potential selves harboured within us, each capable of inhabiting the subject. This idea may be best explained by something like Dennett's [1991] 'multiple drafts' model, which itself harks back to Nagel's [1971] conception of consciousness as the integration of physiological control systems. Here are some words of explanation from Dennett:

> …once a particular 'observation' of some feature has been made, by a specialised, localized portion of the brain, the information content thus fixed does not have to be sent somewhere else to be rediscriminated by some 'master' discriminator…at any point in time there are multiple 'drafts' of narrative fragments at various stages of editing in various places in the brain. (p113).

In Dennett's theory, selective pressure is applied to competing hard-wired brain functions vying for dominion. According to Dennett, there is no 'captain of the crew' - only a virtual captain made up of temporarily dominant coalitions of these specialised brain circuits. The specialised circuits have come about via natural selection during the course of evolution.

[12] The secret self may be seen as corresponding somewhat to the notion of the elusive self, which is described later in this section.

Thus, for example, we have inherited the instinct to duck when something looms, or to heighten our vigilance if the possibility of an emergency is discerned. These, and much more sophisticated, built-in responses to stimuli are all part of the brain's inventory of available functions. But there is no central control system to coordinate these brain functions and generate a conscious experience. Rather, from the 'pandemonium' of competing brain activities a dominant combination wins out, and this is what constitutes consciousness. Relating this to my conception, Dennett's 'dominant combination' of specialised brain circuits constitutes the entity that inhabits the subject. The unsuccessful competing coalitions represent the alternative entities that potentially could inhabit the subject.

**Higher Order Thought**

Higher Order Thought (HOT) theories are not generally about *self*-consciousness *per se*, they are usually just about consciousness. HOT theories intend to explain how it is that a mental state can be conscious (for a summary see Carruthers [2009a]). There are many flavours of HOT theories and I do not explicate them in any detail; my aim is to see if they fall as a whole into the intrinsicist or relationalist side of the Fundamental Dichotomy. The general idea of HOT theories is that a mental state is conscious in virtue of a higher order mental state that represents it. So, I am conscious of being in pain if I have a thought about being in that state of pain. The consciousness of pain (i.e. the perceptual awareness of pain), as in this example, need not be considered indicative of *self*-consciousness (as I argued in section 2.2). However, at one time David Rosenthal [1986] did consider such consciousness to be self-consciousness. In his seminal paper on the subject of HOT theory, Rosenthal says:

> If a mental state's being conscious consists of having a higher-order thought that one is in that mental state, being in a conscious state will imply having a thought about oneself. But being conscious of oneself is simply having a higher-order thought about oneself. So being in a conscious mental state is automatically sufficient for one to be conscious of oneself [Rosenthal 1986, pp343-344].

As I have already argued in chapter 1, and again in a different form earlier in this chapter, self-consciousness should be regarded as consciousness of oneself *as such*. Rosenthal's description in the above passage need not necessarily be of *self*-consciousness. Nevertheless, we must consider HOT theories with respect to self-consciousness as in these theories introspection (which I earlier argued exemplified self-consciousness) is having a *yet higher*

*order thought* about the higher order thought about a mental state (e.g., Rosenthal [1986]; Gennaro [2005]).

Because of the implied separation between the HOT and its target, I argue that HOT is a version of relationalism. There is a sense in which the target mental state is perceived by the HOT, rather than the notion of the HOT and its target being one and the same state. However, Rosenthal [1986] wants to use the idea of HOTs in a way that apparently supports intrinsicism. Rosenthal thinks awareness of conscious mental states is immediate and stipulates that a contemporaneous thought is not mediated by any *inference* or *perceptual* input. But this only reflects Rosenthal's intuition, not reality. When Paula Droege [2005] tries to back up this claim, she provides pain as an example:

> When I am conscious of my pain, I am not conscious of any inferential process preceding the consciousness. I do not think to myself, "Wow, I just hit my knee really hard, I must be feeling pain." Similarly, if a doctor informed me that my knee surgery would likely involve lingering pain, this information alone would be insufficient to make my pain conscious. No conscious inferential or observational process intervenes between a mental state and the higher-order thought about it (sec 3).

In an earlier section I gave pain as an example of certain types of mental states that are accessed directly, but I argued that this type of 'self-givenness' does not represent an awareness of the self as such. Droege then goes on to describe higher order thoughts about mental states in relationalistic terms:

> ...consciousness is a *relational property* of mental states; it is not intrinsic to their nature...a mental state is conscious by virtue of standing in a representational relation to a higher-order thought (sec 3).

One can infer from Droege's description above that she sees the same dichotomy as I do between relationalism and intrinsicism. However, Droege describes Rosenthal's view of consciousness as 'not intrinsic to their nature' despite earlier saying that for Rosenthal awareness of mental states is immediate and unmediated – exactly the terminology I have used to describe intrinsicism. The apparent inconsistency is resolved if we adhere to the distinction I made earlier separating 'self-givenness' (awareness of certain types of mental states such as pain) from self-consciousness proper (awareness of oneself as a psychological subject). HOT theories come into play again in the Elusiveness Thesis, discussed later, which also supports the relationalist side of the Fundamental Dichotomy.

**Immunity to Error through Misidentification (IEM) and Reference Failure**

Sydney Shoemaker [1968/1994] is credited with naming this peculiarity of self-consciousness, although he himself credits Wittgenstein with first noticing it (and Brook [2001] sees similar themes discussed in the writings of Kant). The basic idea is that there are certain usages of the first person pronoun that are immune to error through misidentification – that is, they can never be in error as to the referent. When 'I' is used in public language or its mental equivalent, it always refers to the self (the utterer). An example of this type of usage is 'I am in pain'. This is unlike other usages of first person pronouns where there is a possibility of error, such as in 'my legs are crossed'. It is conceivable (perhaps due to a trick with mirrors) that the utterer is mistaking someone else's crossed legs for his own, but there is never any doubt about who is in pain. IEM is similar, and perhaps related to, another peculiarity noted by several authors, immunity to reference failure [Shoemaker 1968/1994; Castañeda 1969/1994; Strawson 1994]. A correct usage of 'I' is apparently guaranteed against both a lack of reference and against a mistaken reference. 'I' must always refer to the utterer.

I see a correspondence between IEM and the self-as-subject-only view of self-consciousness. For example, Shoemaker describes the usage of 'I' (as in 'I am in pain') as a *subject* usage while the use of 'my' (as in 'my legs are crossed') is an *object* usage. This correspondence aligns IEM with the intrinsicist side of the Fundamental Dichotomy, which on further examination seems to make sense. There is a certain immediacy involved in IEM; an irreducibility of the 'I'. It is immune to error through misidentification because it is inseparable from the self that it refers to. In using 'I' one refers to a *state* that the self is in as compared to a usage of 'my' which refers to a property owned by the self. In using 'my' one can separate the self owning the property and the property being owned (say, the self as physical body or body part). Let's assume that a proper usage of the first person pronoun is an act of self-consciousness. Then, this apparent irreducibility of the 'I', this seeming immunity to error through misidentification, must correspond to an intrinsicist view of self-consciousness: there is no conceptual gap between the utterer of 'I' and the referent. If IEM is real, then this argues against the relationalist side of the Fundamental Dichotomy. It is therefore incumbent on me to present a case that calls into question the validity of IEM.

Possible counter-examples refuting IEM are certain pathologies such as thought-insertion and hearing voices [Frith 1992, p66]. Subjects experiencing these intrusive thoughts may misidentify usages of 'I', by attributing them to an external entity rather than to themselves, even though they originate from within their own minds. For example a sufferer may experience a voice making statements such as "I want you to boil the bunny" and may disown or lack a sense of ownership of the voice. Here, the sufferer takes the instance of 'I' in the foregoing sentence to identify someone else, while it is the instance of 'you' that refers to herself, even though the voice originated from within her own head.

In another example, given the earlier discussion of a hidden observer, 'I am in pain' spoken by a hypnotised subject will leave some doubt as to the referent. That is, it could be the entranced subject or it could be the hidden observer. In split brain patients, too, there appear to be two consciousnesses, either of which might be able to use 'I', so that an external observer will be in doubt as to which consciousness made the utterance[13].

In summary, where there is some doubt as to the unity of self-consciousness, as previously discussed, there opens up the possibility of a multiplicity of attributions of 'I' utterances. Where this occurs, a misattribution is also possible, thereby calling into question IEM. The first person pronoun, it seems, is not really immune to error through misidentification after all. The intuition that underlies IEM, however, remains and may be a facet of what Ryle [1966/1994] calls the elusiveness thesis. As I argue next, the elusiveness thesis, too, supports the relationalist side of the Fundamental Dichotomy.

**Elusiveness Thesis**

Gilbert Ryle's [1966/1994] Elusiveness Thesis is that there will always be an elusive 'I' in self-referent sentences that cannot be replaced by 'my body'. This elusiveness of 'I' is a peculiarity of self-consciousness that supports the relationalist side of the Fundamental Dichotomy. It is rooted in the view that for a mental state to be self-conscious, it needs to be accompanied by a higher order thought. If I am conscious of the smell of cooking this could

[13] Possibly, it *could only* be the left brain consciousness, as we might preclude the possibility of the right hemisphere acquiring sufficient language skills. However, even in this case it might be possible for the right brain consciousness to express the self-reference mentally using a language of thought – that is, without necessarily using *natural* language in thoughts. Note that I do not take a stand here on the existence of a language of thought; either way my arguments stand.

be a low-order mental state, in this case a perception. If I then think 'what am I smelling?' that thought is of a higher order, operating on the lower order mental state. The elusiveness comes in with regard to the consciousness of the highest order state. This state cannot operate on itself, though it can be operated on by a yet higher order mental state. Ryle [1966/1994, p39] refers to this as the systematic elusiveness of the 'I'. The following passage by Ryle illustrates his point and also reflects some of the ideas I have presented earlier regarding self-conversations:

> A higher-order action cannot be the action upon which it is performed. So my commentary on my performance must always be silent about one performance, namely itself, and this performance can be the target only of another commentary. Self-commentary, self-ridicule, and self-admonition are logically condemned to eternal penultimacy. (pp39-40)

To use the terminology I have previously introduced, one can think of the highest order act as being undertaken by the self inhabiting the subject. The lower-order mental states are the objects of the thought had by the subject. Only the highest order mental state is the state of the subject. That state is itself in turn accessible for the operation of yet another mental state, but it thereby loses its status as the highest order state, for it then becomes an object of the new (higher order) mental state. That new higher order mental state then becomes the mental state of the self inhabiting the subject. The lower order mental state can be thought of as a manifestation of the self, and being the object of a higher order thought can be considered a self-as-object. The self-as-subject having the higher order thought cannot itself be an object of thought – or rather, it can *become* the object of thought, but only by sacrificing its place of inhabiting the subject – it becomes an object of the thought had by that manifestation of the self that then inhabits the subject. There can only be one subject of experience at any one time. This is how I view the notion of the elusive self: the manifestation of the self inhabiting the subject cannot itself simultaneously be the object of its own subjectivity. An attempt to do so results in a forced separation of the manifestations of the self as subject and self as object. As Sartre [1943/2000] observed, as soon as we turn our attention onto our own self, by examining a thought or mental state we are in, we need to 'jump back' from that state in order to examine it. We will then also be in a mental state of self-examination, which itself is available for access but to access that state we need to jump back from it again. Consider another quotation from Gilbert Ryle:

> If I perform a third-order operation of commenting on a second-order act of laughing at myself for a piece of manual awkwardness, I shall indeed use the first

> personal pronoun in two different ways… 'I was laughing at myself for [my] being butter-fingered.' [Ryle 1966/1994, p42]

Notice that 'myself' and 'my' are first personal pronouns that can be replaced by 'my body' but that there will always be that elusive 'I' that can't. We could add a fourth order, that I was thinking about how I laughed at myself for my being butter-fingered. Now the previously elusive 'I' (in 'I laughed') *can* be replaced by 'my body' but we've added a new (elusive) 'I' to perform the highest-order act. It seems one always has to 'step back' from oneself in order to examine the self. In other words, to introspect, we need to introduce the conceptual gap between the introspecting self and the introspected self. Once again, this conceptual gap implies a relation between the introspector and the introspected, and as such provides support for the relationalist side of the Fundamental Dichotomy.

In the foregoing example, we see the presence of an elusive 'I' that cannot be replaced (for example, by 'my body') – at least, not without reintroducing 'I' elsewhere within the sentence. Apparently, the utterer of self-referential sentences is unable to disengage from the 'I'. The elusive 'I' is reminiscent of the 'I' in IEM, which seemed to be immune to error as it was 'locked into' the self. Thus 'I' seems to be essential in the sense of being somehow indispensible, at least in self-referential utterances. The fact that 'I' is essential in this way has been considered of some import for self-consciousness (e.g., Castañeda [1966/2001]; Perry [1979/1984]) and could be viewed as supportive of the intrinsicist side of the Fundamental Dichotomy. I argue next that in fact the so-called 'essential indexical' is *not* essential (at least, in thoughts) and so does not threaten the relationalist side of the Fundamental Dichotomy.

**The Essential Indexical**

In *The Myth of Mental Indexicals* Ruth Millikan [2001] argues that so-called essential indexicals in thought are indeed essential but not indexical. In this section I argue quite the reverse: they are not essential but they are indexical. After presenting a bit of background on the notion of the essential indexical and how the terms are used, I show why the 'essentiality' aspect of the essential indexical (if it were correct) supports the intrinsicist side of the Fundamental Dichotomy. I then present a case that although the essential indexical exists in public speech, in thoughts the essentiality evaporates. Further, I argue that in thoughts the

indexicality remains and that together these points support the relationalist side of the Fundamental Dichotomy.

An indexical is an expression whose reference is only fixed by context. Examples are ‘here’, ‘this’, and ‘I’. If I say ‘here is the boat pond’ on 4 July 2009 and then repeat the sentence the next day, the reference could be completely different: the second time I could be standing next to a different boat pond. Similarly, two persons could both utter that same sentence at exactly the same time and still be referring to completely different things, if they are nowhere near each other. The referent is indexed according to the *context* of the sentence: ‘here’ is indexed to the location of the speaker; ‘this’ is indexed to the object indicated by a pointing finger and ‘I’ is indexed to the speaker.

Usually indexicals can be substituted with a definite description. Thus, if I point to Clark Kent and say “He is Superman” that is equivalent to the sentence “Clark Kent is Superman.” ‘He’ in this sense can be substituted by the name or some other definite description of the indicated object. However, some indexicals (especially, ‘I’) are ‘essential’ in that they apparently *cannot* be substituted with a definite description in a sentence without losing equivalence in meaning. In the case of ‘I’ this linguistic peculiarity leads to some interesting conclusions about self-consciousness.

The most famous description of the essential indexical appears in John Perry’s [1979/1984] supermarket example. Perry notices that someone is spilling sugar from their trolley but at first doesn’t realise it is himself. In this situation, the ‘I’ in the sentence “I am making a mess” cannot be substituted for anything else (such as ‘John Perry’) without changing the meaning in a fundamental way. Although the meanings of “I am making a mess” and “John Perry is making a mess” appear to be the same if spoken by John Perry, the second sentence loses its ability to explain Perry’s behaviour:

> Suppose I had said, in the manner of de Gaulle, ‘I came to believe that John Perry is making a mess.’ I would no longer have explained why I stopped and looked in my own cart. To explain that I would have to add, ‘and I believe that I am John Perry’, bringing in the essential indexical again. [Perry 1979/1994, p169]

To give a complete and definitive explanation of his meaning, Perry is unable to avoid using the essential indexical ‘I’. As Castañeda [1966/2001] puts it, the first-person pronoun is an unanalysable category. By this he means that a usage of ‘I’ as in the example above “...cannot be analysed in terms of any other type of referring mechanism” (i.e. other personal pronouns,

proper names, demonstratives and definite descriptions)[14]. The 'I' appears to be essential; irreducible; unsubstitutable.

The notion of 'I' as an essential indexical, like the idea of immunity to error through misidentification, supports the intrinsicist side of the Fundamental Dichotomy. In IEM, so the argument goes, 'I' cannot fail to refer to the self because there is no separation between them – thus there is no room for error in the reference. The essentiality of 'I' has the same general effect: 'I' is necessarily *locked in* to the self. Indeed, unlike other indexicals such as 'this', there is no explicit identification necessary; no need for a pointing finger: 'I' intrinsically indexes the self. What I argue, though, is that although the essential indexical exists in public speech, it does not exist in 'private speech' (thoughts). 'I' usage in thoughts are *not* essential in the manner exemplified by the Perry passage above. Comparing 'I' usage in public speech and thoughts, 'I' usage in public speech has more to do with linguistics and communication than with self-consciousness, whereas 'I' usage in thoughts has more relevance to self-consciousness. As such, it is the 'private speech' usage that we need to evaluate in regard to the Fundamental Dichotomy. If I am right, that 'I' usage in thoughts is not essential in the sense previously described, then this notion of the essential indexical provides no support for the intrinsicist view.

In the Perry example quoted above, by using 'John Perry' in place of 'I', Perry has to subsequently make clear *to the listener* that he believes he is John Perry in order to clarify his meaning. This qualification would not be necessary for Perry's own sake – that is, he would not need to make clear to *himself* that he believes he is John Perry; that would be absurd. I do not mean it is absurd that he might not know he is John Perry, for he might be suffering amnesia. But in such a case, where he was ignorant that he was John Perry, he would not have known that "John Perry is making a mess." It is clear in the passage presented that John Perry is aware he is John Perry and is aware that he is referring to himself in the third person. As such he need not convince *himself* of who he is; he only needs to reaffirm that to *others*.

[14] In fact it was Castañeda [1966/2001] who had earlier noticed this peculiarity when examining particular uses of 'he'. 'He' can be used in the sense 'he himself', for which there is no single word in the English language. To signify this particular usage, Castañeda writes it as 'he*'. In my example regarding Clark Kent, 'he' was *not* used in this sense. An example of 'he*' usage would be: (a) 'Clark Kent knows that he* is Superman'. He* is another example of an essential indexical. For example, we could attempt to substitute 'Clark Kent' for 'he*' in (a) to produce (b) 'Clark Kent knows that Clark Kent is Superman'. But (b) is not identical to (a), for Clark Kent might know that Clark Kent is Superman without knowing that he is himself Clark Kent (he might be amnesic about his own identity while still remembering that fact about Superman's identity). Just as with Perry's example, to make (b) identical to (a) we would need to add 'and Clark Kent knows that he* is Clark Kent', bringing in the essential indexical again.

For example, if I look in the mirror and think to myself "Stef Savanah, you handsome devil!" I do not need to add "and I believe I am Stef Savanah" just because I used the third person in self-reference. Similarly, in the Perry example quoted above, he would *not* need to add "and I believe that I am John Perry" if there was only himself as audience. In thoughts, then, 'I' does seem to be substitutable with a definite description (such as the name of the thinker) without losing any meaning or losing any explanatory power for motivation. So the essentiality of the first person pronoun applies only to public usage. In thoughts, the *essential* indexical does not exist; it only exists as a linguistic quirk in public usage. As the *essential* indexical does not exist in thoughts, it loses its force to support the intrinsicist side of the Fundamental Dichotomy. Although in thoughts 'I' is not *essential*, I argue that it is still *indexical*, and being so supports the relationalist side of the Fundamental Dichotomy, as I explain next.

How strangely unsettling it is when people speak about themselves in the third person. I always get a niggling doubt as to whether they actually know they are in fact talking about themselves! That doubt disappears when they use the first-person pronoun, for that is the reason for using the word 'I'. 'I' is used to tell other persons facts about the self. But what are the reasons for non-public tokens of 'I'? That is, when having I-*thoughts*, what is the purpose of the *mental* usage of 'I'? If we do use 'I' in thoughts, it cannot be to communicate facts about ourselves to others, as others are obviously not privy to our thoughts. We might presume usage of 'I' in thoughts must be to communicate facts about ourselves to *ourselves*. But if this is so why would we need to use the word 'I'? There seems no need to index ourselves when having thoughts about ourselves. In thoughts, 'I' would seem to be dispensable. It is for this reason, I believe, that Ruth Millikan [2001] argues there is no such thing as a mental indexical. Her point is that there is no interpreter of 'I' in thought for whom it would be indexical (since it could not be indexical for the self). In the following passage, Millikan uses '@RM' (her own initials) instead of 'I' to represent her self-referential thoughts:

> Perhaps we are supposing the relevant context to be the *mind* '@RM' appears in. ... Are we supposing then, that in my language of thought, in my inner system of representation, tokens of '@RM' might appear in your head so that I must check whose head '@RM' appears in before identifying its content? Or, are we supposing, perhaps, that my mental language is some sort of universal language, one selfsame language that all people speak in their heads, so that rather than '@RM', I must think 'I,' the self-name in universal Mentalese? But even if that were the case (maybe Jerry Fodor thinks that it is), in what sense would the self

> name be *indexical?* Certainly there would be no interpreter for *whom* it would be indexical. (p173)

The final sentence in the above passage shows that Millikan believes there is no interpreter for whom an 'I'-thought would be indexical, therefore it cannot be indexical. But the natural language usage of 'I' clearly *is* indexical – it obeys the rule for indexicals, that its reference is context-dependent: when you say 'I' you refer to a different entity than when I say 'I'. By this reasoning it seems that how we make a *mental* self-reference cannot be with the natural language 'I'. I take Millikan's '@RM' as a way to describe the mental tokening of a self-reference without recourse to the (natural language) 'I'. The point is, of course, that such mental tokenings of a self-reference may have completely different properties to the natural language tokening, in particular *not* being indexical. But using '@RM' disguises an important fact: that even if there truly does exist a language of thought[15] in which the tokening of a self-reference is indeed non-indexical as Millikan asserts, it does not preclude a mental usage of the natural language 'I'. We *can* and *do* experience streams of thought that most definitely do use the natural language 'I'. That is, there are times when we think using the same language we use for public speech. That means we do use the actual word 'I' in private thoughts, for example, when we engage in self-conversations using natural language (at least, I know I do). So, if '@RM' is meant to represent a pure Mentalese tokening of a self-reference distinct from the natural language 'I', then it must co-exist with the natural language 'I' in thought. When I think to myself "how will I explain my ideas in the next section?" I literally do think using those words, in English, just as I would when conversing with another person. The 'I' as used in the quoted thought is a natural language usage just the same as if I had spoken the sentence out loud to an audience. As such, and contrary to Millikan, *that* usage remains indexical, even though it is a mental tokening.

Now, if a mental tokening of 'I' remains indexical, then by Millikan's own reasoning there must be an interpreter for whom it is indexical. But there is no external interpreter privy to that thought; there is only the self having the thought. This is just the paradox Millikan has raised, but it is resolved if we allow the possibility of an *internal* interpreter for whom it is indexical. That is, the self or a manifestation of the self acting as interpreter of the thoughts expressed by the self. This state of affairs is conceivable in the sense previously discussed, where in self-conversations there is a conceptual separation of the self manifested as talker

[15] As previously mentioned I do not take a stance here on the existence of a language of thought. Whether or not it exists does not affect my argument that we can and do (also) think in natural language. I give some thoughts on language issues in chapter 4.

and the self manifested as listener. Both manifestations represent the same subject, but the self manifested as listener is the interpreter for whom a usage of 'I' – even in thought – is indexical. I emphasise again that I do not view this as if there were literally two subjects; there is only ever one subject of conscious experience. But the individual is made up of many parts or manifestations of the self that interrelate, and while only one of these can inhabit the subject at any one time, another can be the object of that subject's thought. When you mentally talk to yourself you force a conceptual gap between yourself as the talker, and yourself as the audience of your thoughts. This separation can account for the indexicality of natural language 'I' usage in thoughts. And, once again, the existence of a conceptual gap between 'talker' and 'listener' supports the relationalist side of the Fundamental Dichotomy.

## 2.4 Conclusion

The two alternative models for access to the self, the perceptual model and the privileged access model, form a dichotomy that I argue can be generalised to a Fundamental Dichotomy pervading the philosophy of self-consciousness. The perceptual model can be generalised as the *relationalist* position in the Fundamental Dichotomy while the privileged access model can be generalised as the opposing *intrinsicist* position. I argued that the relationalist position is the correct characterisation of self-consciousness proper (or 'full-fledged' self-consciousness). The intrinsicist position is characteristic of what might be considered a primitive form of self-consciousness, a form that earns that nomenclature only because it represents access to some self-specifying mental states, but which does not meet the standard I set in chapter 1. The standard I set is deliberately high in being humanlike self-consciousness, for I am interested in whether any non-human animals have an understanding of their own existence like humans do, that is, whether they possess a self-concept.

This positioning of relationalism and intrinsicism was measured up against the many and diverse known and supposed 'peculiarities' of self-consciousness. I showed how the various arguments naturally fell into one side or the other of the Fundamental Dichotomy and I argued that in each case relationalism is the preferable position. Consolidating those peculiarities consistent with relationalism forms the kernel of a coherent framework for exploring the philosophy of self-consciousness. Although I do not claim by any means to have formulated a complete general theory of self-consciousness in this work, the

Fundamental Dichotomy does play a key part in the central thesis argued herein: the Concept Possession Hypothesis. Concept possession is the key factor that separates the relationalist position (which applies to self-consciousness proper) from the intrinsicist position (which at best only applies to a 'primitive' form of self-consciousness that I prefer not to describe as self-consciousness at all). This two-tiered view reflects the top two levels in a phylogenetic hierarchy where the highest level consists of full-fledged self-conscious organisms and the next level down consists of organisms that are conscious but *not* self-conscious. This hierarchy is more fully developed in the next chapter in which I argue for the Concept Possession Hypothesis.

# Chapter 3: The Concept Possession Hypothesis of Self-Consciousness


## 3.1 Introduction[16]

I offer a straightforward hypothesis about the underlying nature of self-consciousness: concept possession implies self-consciousness and vice versa. In chapter 1 I encapsulated my notion of self-consciousness as *an understanding of one's own existence as a psychological subject with intentional agency.* This can be viewed as a description of what it is to possess a *self*-concept; thus by very definition self-consciousness entails the possession of a self-concept. The key idea in the Concept Possession Hypothesis of Self-Consciousness (henceforth CPH) is that the possession of *any* concept implies self-consciousness. I argue below that a subject must already possess a *self*-concept in order to have any other concept. If so, then a convincing demonstration that the standard for self-consciousness has been met is any display of concept possession by the subject. What should and should not count as a display of concept possession is examined in chapter 4. This chapter is devoted to arguing for the plausibility of CPH.

It is intuitively evident to many that there is some sort of correlation between intelligence and self-consciousness (e.g., David Chalmers [personal communication]). By 'intelligence' in this context I mean flexibility of behaviour, to be contrasted with stimulus–response type behaviour. The main characteristic of intelligence is the ability to decide on an action, to *deliberate*, rather than to act 'automatically' to a given stimulus. We can see from everyday

[16] A version of chapter 3 has been published as a paper (Savanah [2012a]). The differences between the versions are mostly minor with the exception of the following. In the version here I make clear how CPH was derived from the Fundamental Dichotomy as presented in chapter 2. Also in this version I include a discussion on the theory of non-conceptual content, which in the paper was taken to be implicit from the discussion on Bermũdez [1998].

speech the intuitive connection between flexibility of behaviour and self-consciousness. We often use the terms 'deliberate' and 'conscious' synonymously (where by context we can assume that by 'conscious' it is '*self*-conscious' that is meant). For example, "I didn't do it consciously" means "I didn't do it deliberately." When someone uses that expression they mean they did not exercise conscious control over their actions – rather, they reacted 'automatically', which is to say they responded (unthinkingly) to a stimulus.

To be clear: in making the connection between intelligence and self-consciousness, I do not mean to imply that human beings with higher IQs are 'more' self-conscious than those with lower IQs. What I mean to say is that 'phylogenetically advanced' species are more likely to be self-conscious than less advanced species. And when I say 'advanced' I mean *behaviourally* advanced in the sense that a greater degree of species flexibility is evident. The question to be answered is whether the behaviour observed is a manifestation of 'intelligence' or the result of 'automatic' (stimulus-response type) mechanisms. The essence of intelligence is conceptualisation, as concept usage involves generalisation which is what allows a subject to engage in flexible behaviour (I say more on all this later).

In section 3.2 I chart a hierarchy of consciousness as inspired by the Fundamental Dichotomy presented in chapter 2. Here I divide living beings into three very broad levels characterised by both mode of behaviour and associated level of consciousness as I view it. The purpose of this high level taxonomy is to demonstrate the apparent connection between behavioural flexibility and self-consciousness. As discussed in chapter 2, the factor that separates the top two levels (between organisms with *self*-consciousness at the top and those with only consciousness at the next level down) is concept possession, which is the essence of CPH. The argument thus turns on the proposition that there are organisms capable of conscious mental states that are able to interact and thrive in their habitat *without* possessing any concepts. Therefore, in section 3.3, I argue in favour of non-conceptual mental states by offering the case of perceptual states as one example.

In Section 3.4 I present three considerations in support of CPH. These are an analogy with perception, in which the self is presented as central to both perception and conception; an argument based on the notion of a 'web of concepts' that *necessarily* includes the self-concept; and a discussion of how José Luis Bermũdez's [1998] ideas in *The Paradox of Self-Consciousness* provides support for CPH.

### 3.2 A Yardstick for Self-Consciousness

My intention in this section is to establish the *a priori* plausibility of CPH by highlighting the connection between intelligence and self-consciousness. I want to be clear that I do not equate intelligence with self-consciousness; intelligence is about flexibility of behaviour while self-consciousness is about awareness of the self *as a self*. These are different things, and yet intuitively they appear to go hand-in-hand. Indeed, the same animals we suspect of higher intelligence are the first we test for self-awareness (chimpanzees, elephants, dolphins, etc).

The one thing we can be absolutely sure about is that adult humans have the capacity for self-consciousness. So, to begin with, consider the difference between humans and non-humans. To motivate the discussion, I pose the following question. What, essentially, is the difference between each of the following:

1. The metal jaws of a triggered bear trap clasping its victim's ankles
2. A frog flicking out its tongue to snare a flying insect
3. A pouncing leopard clamping its maw upon the neck of its prey
4. Me chewing on a fried chicken leg

In purely mechanistic terms, the four descriptions above are similar: in each case one entity is stimulated to attack another. A casual Martian observer might fail to detect any difference at all, though we humans know there is a big difference, at least between the two extremes. A bear trap does not have a mind and is not conscious (far less *self*-conscious) but at the other end I'm pretty certain that I am conscious and self-conscious. The two in the middle are a little trickier, though. How can we tell if a frog is self-conscious, or even conscious? And what about the leopard, which at least has a more advanced brain? I would want some good evidence before accepting that the leopard is self-conscious, let alone the frog.

Now, it is often said that what separates humans from other species is our capacity for abstract thought. A key difference between me and the leopard is that I have the ability to reason inferentially, to deliberate – but again I would want a lot of good evidence before I'm prepared to accept that the leopard has this ability. *I argue that our special capacity for reasoning is intimately tied up with self-consciousness.* What I argue is that those organisms that are conscious but not self-conscious can only behave in a programmatic way (either through genetic hard-wiring or by associative learning), while self-conscious organisms are able to deliberate. If so, then perhaps the causal role of self-consciousness is to provide the

organism with the ability for abstract thought such as inferential reasoning. Based on this account, behaviours of animals and human infants that can definitively demonstrate this ability should be taken as conclusive evidence of self-consciousness in the subjects. This, then, provides the yardstick for the evaluation of self-consciousness research.

**Three Levels of Development**

We can identify three broad layers of development on the road to self-consciousness, phylogenetically speaking: in the first level are organisms with no consciousness at all; in the second are organisms with consciousness but no self-consciousness and in the third level are organisms with self-consciousness. The two highest levels here reflect the Fundamental Dichotomy as discussed in chapter 3: the relationalist side of the Fundamental Dichotomy properly characterises the nature of self-consciousness (the highest level) while the intrinsicist side does not. The intrinsicist side can, however, be applied to the next level down where consciousness is present but *not* self-consciousness as I have defined it.

Level 1 is for organisms with no mind. It might not be so easy to always know whether an organism has a mind or not, but it is safe to say that there are at least some simple organisms that do not have a mind, such as single-celled organisms or bacteria. In these organisms, not having a mind, any representations can only be physically instantiated; there are no mental representations because again there is no mind to have a mental anything. No mind means no consciousness. These organisms are not conscious of their surroundings except perhaps in some very loose or metaphorical sense, in the same mechanistic way as might be said of a mousetrap or chess-playing computer programs: they 'perceive' their environment as when certain stimuli trigger (conditioned or genetically) pre-programmed responses. The organisms' ability to interact with the world is wired into their control systems (brains, or nervous systems or whatever) and not mediated by a mind.

Level 2 is for organisms that do have a mind and are capable of consciousness but *not* self-consciousness. Again, it's hard to know which animals to include in this category, but (perhaps due to some kind of taxonomic class bias on my part) I tend to assume that all mammals, with their relatively advanced brains, reach at least this level. Like humans, non-human mammals have sense organs and a not too dissimilar brain, so it is a reasonable

assumption that they and we have at least some mental states in common; for example, visual and other perceptual states. It is also reasonable to assume that even if there are some non-human animal species that in fact *are* capable of *self*-consciousness there must be some that do *not* reach that level of development and level 2 is intended to encompass this group. Thus I include in level 2 any organisms that stop short of *full-fledged* self-consciousness but may be capable of non-conceptual self-specifying mental states (which in earlier chapters I grudgingly characterised as a 'primitive form' of self-consciousness but not self-consciousness proper as I defined it in chapter 1).

Many authors accept that there are different levels and/or types of consciousness (e.g., Block [1995]; Gennaro [1996]; Nelkin [1996]; Kriegel [2007]) but for our purposes we need not specify all of these and how they differ. We need only acknowledge that level 2 organisms are capable of some or all of these conscious states but not self-consciousness. At the very least, level 2 organisms have perceptual consciousness; that is, they are conscious *of* their immediate environment. We need also acknowledge that in having minds their perceptions will result in mental representations. For example, a visual image of a seen object is a mental representation of the object. The same could be said of other perceptual modes, such as audition. Of course there is a physical basis for these mental representations; they are somehow encoded in the dynamic physiology of the brain. So, associated with a mental representation is a physical representation (perhaps a neural correlate). How the physical encoding of representation in brain physiology leads to the experience associated with a mental representation is beyond the scope of this work[17]. However, I want to emphasise that possessing mental representations, such as visual images, does not imply that the organism is capable of *understanding* what those images represent. I argue this point in more detail in section 3.3; for now I proceed by way of an example.

To some, the claim that level 2 organisms have mental representations may imply they are capable of intentional behaviour (that is to say behaviour that goes beyond stimulus–response reactions). But this is not necessarily the case; certainly not in the (perhaps sparser) way I am using 'mental representations'. I suspect that my usual example, a visual image, is common enough in the animal kingdom. I can see no reason why an organism that is capable of experiencing a visual image must be considered capable of intentional behaviour. I can

[17] This is commonly known as the 'mind-body problem' but it has been rebadged many times over, such as Joseph Levine's [1983] 'explanatory gap'; David Chalmers' [1995, 1996] famous 'hard' problem of consciousness; and Jaegwon Kim's [2005] problem of physical causal closure, to name but a few.

imagine cases even in humans where mental representations lead to actions that are not intentional. For example, as in the movie *A Clockwork Orange* a person might react involuntarily to the sight (or, 'visual image') of violence by vomiting. Thinking of a melody (a mental representation of an auditory kind) might make a person spontaneously weep, before realising that the cause of the sadness was a memory associated with the tune. If this is the case for humans it is conceivable that for some or all animals *only* stimulus–response actions are triggered by their mental representations. The behaviour of level 2 organisms may be more complex than level 1, but it is still underwritten by either genetic hard-wiring or by associative learning. Within level 2 I place organisms that are not intelligent in the way I have been using that term: they do not act flexibly to new circumstances[18], rather they still follow a stimulus–response paradigm. An example might be the stalking and attack response of a predator to the stimuli of hunger and perception of prey. This behaviour does not demonstrate what I mean by intelligence. The actions of level 2 organisms are still based on a stimulus-response paradigm although in this case they may be mediated by mental representations.

Being conscious does not necessarily mean being self-conscious[19]. An organism may be perceptually conscious of objects in its immediate environment – perhaps including its own physical body – without necessarily being aware of its own existence as a psychological subject. So now we come to the highest level in this simple taxonomy, level 3. Humans are a paradigm example of level 3 organisms. In humans we get both self-consciousness and intelligent behaviour. Humans are able to engage in the greatest level of flexible behaviour, with the ability to deliberate and solve complex problems. Humans, we know, are also self-conscious in the relevant way – they are aware of their own existence. Let this dual capacity characterise level 3.

[18] I mean 'flexible' here in the sense of the *individual's* capacity. The species as a whole may be considered 'flexible' in a certain sense - that it can co-evolve with its changing environment through natural selection. And we might also say that level 2 organisms are 'more flexible' than level 1 organisms, in the sense of possessing a *wider range* of (species-specific) behaviours.

[19] Some other authors, however, do equate the two. For example, ethologist Euan MacPhail [2000] suggests that animals such as cats and dogs are conscious in the sense that they experience feelings and (he claims) a self-concept is necessary for the experience of feelings (p267). To Dickinson & Balleine [2000] there are only two levels to consider rather than my three: "... S-R [stimulus-response] robots and cognitive creatures that are also endowed with intentional representations, affective experience, and the ability to integrate the two in consciousness..." (p201). Uriah Kriegel [2004] argues that there is no consciousness without self-consciousness and Rocco Gennaro [1996] claims that consciousness entails self-consciousness. These may amount to no more than terminological differences in how 'consciousness' and 'self-consciousness' are viewed. For example, in chapter 1 I argued that self-perceptions (such as consciousness of one's own pain) might be considered self-consciousness only in a very primitive sense but certainly not in the way I defined it in chapter 1.

This three-level framework provides the basis for my arguments in the next two sections in which I argue that concept possession (the essence of 'intelligence') is what separates self-conscious organisms (level 3) from conscious-but-not-self-conscious organisms (level 2). But first I present a brief discussion about the transition between these levels.

**Intermediate States**

In the simple breakdown of developmental stages into three broad levels the obvious question is: *is there an important level missing?* More specifically, is there a level in between levels 2 and 3 in which we find only intelligence or only self-consciousness, but not both together? If that were the case then it would undermine CPH, which effectively says the two are co-dependent. But if there is such an intervening level, which of intelligence or self-consciousness characterises it? The difficulty in answering this question is one more good reason to suspect that they in fact co-evolved. This simple three-level taxonomy is not meant to imply that there is a quantum jump between levels; in all likelihood the transition is not all-or-nothing but graduated. The simplification of three distinct levels is for the purpose of explication. But if it can be argued that there may exist transitional stages whereby an organism possesses only (say) 'proto-concepts'[20] (provided such a conception could be rendered explanatorily useful) then I would argue the organism has correspondingly attained only proto-self-consciousness, and it is in this way that I see the two capacities as co-evolving. But it is reasonable to ask what such a partial state of self-consciousness would be like. Despite the difficulty of addressing this question (for example, Davidson [1999, p11] remarks that we have no vocabulary for describing such intermediate states) it is intriguing and worthy of some analysis. The question itself may be interpreted in several ways, for example, structurally speaking; behaviourally speaking; and phenomenologically speaking. Of these, the latter seems most pertinent in the sense of *what it is like* to be in any state, however I will briefly address the other two questions first.

Structurally speaking, an example of an intermediate state might be what is commonly referred to as a 'cognitive map' (I am grateful to Kim Sterelny for suggesting cognitive maps

[20] In chapter 4 I briefly discuss the issue of 'proto-capacities' in general; I do not deny their existence outright but I do caution that their usage threatens infinite regress. In any case, the fact that the *transition* from level 2 to level 3 remains unexplained does not interfere with the central thesis that a correlation exists between concept possession and self-consciousness.

as a possible example). In section 4.2, I suggest that map-like representations could be the basis for conceptual thought but that non-conceptual thought might also have a map-like structure. The point being made in that section is that language *might not* be necessary for conceptual thought but what also emerges is that map-like content might just be the *bridge* between non-conceptual and conceptual thoughts. Inherent in simple maps are spatial relational information from which inferences *could* be made (e.g., the forest is east of the lake, therefore the lake is west of the forest). However, an organism with a cognitive map representing a locale might have no concepts of what a map is or of any element in the map and yet be able to navigate the optimum route between points (for example, see the discussion on spider navigation in section 7.7). Conceivably, then, the answer lies in between these two extremes; map-like content might be a model for minds that are *partially* concept possessing (and hence partially self-conscious).

Behaviourally speaking, we can consider the human model and look to infants that are on the brink of attaining self-consciousness. I argue in chapters 4 and 5 that this occurs in parallel with the onset of symbol-mindedness and self-recognition. In chapter 5 I discuss the range of behaviours that reflect developmental progress around self-recognition (i.e. searching behind the mirror then mirror self-recognition then self-naming, etc.). In the case of symbol-mindedness, we see that the infants on the brink of attaining this capacity can pass a test of symbol-mindedness in one configuration yet not another (i.e. in the case of viewing the model room with or without the benefit of a frame). Behavioural cases like this may indicate transitional states.

Phenomenologically speaking, we could again turn to the human model and ask what it is like to be a human infant at around the age of 18 months or so. I choose this age because this is around the age at about which they become symbol-minded (as discussed in chapter 4). Of course, asking an infant of this age about their experience in regard to understanding something would most likely be futile in respect of our gaining any purchase on what a state of partial concept possession is like. We were all that age once so we have all been in that state, but human adults are mostly considered to not have any memory from before the age of 3 and even if that were possible (e.g. see Reese [2009]) it is doubtful that any such memory could be useful in describing the required phenomenology.

Perhaps the closest we can get to the phenomenology of intermediate states are certain experiences as when one is struggling to grasp a concept or get traction on an idea in its

formative stage and then the ‘penny drops’. A state of partial concept possession possibly exists in the brief transition between ‘not getting it’ and then suddenly ‘getting it’. Now, we have to be cautious in this approach since the example just given describes *concept formation* or *concept acquisition*. What we really want are not partial states between *not grasping a particular concept* and *grasping that concept*, but rather between *not having the capacity to possess concepts* and *having the capacity to possess concepts.* Nevertheless, the phenomenology may not be too dissimilar.

Another possible approach to get a grip on the phenomenology of partial concept possession is in situations where one is taking an action ‘instinctively’ without knowing exactly why, but having some partial ‘back of the mind’ understanding. In some cases of human instinctive or automatic responses to situations the subject might never question her own actions (e.g. say in ‘automatically’ driving a car) but there are some situations in which an action that is performed is simultaneously questioned by the subject. The subject may just know that it is the right action to take and might have some vague idea of why but can’t precisely ‘put a finger on it’. This might be the case, for example, in making choices under time pressure. In such a situation, the accompanying questioning is also likely to be fleeting, since there is no time for any deep introspection (at least not until later). This ‘questioning’ of course verges on an act of metacognition, so along with the partial understanding of why an action is undertaken we perhaps have a case of partial metacognition.

In both the above speculations we can at best have only a proxy for the phenomenology of partial concept possession but they at least provide some level of familiar experience from which to draw. Now imagine an organism that is only ever capable of being in those partial states. That is, capable of being on the brink of ‘getting it’ but never being able to see the penny drop, or never being able to grasp exactly why they behaved beyond some vague idea. This might be what it is like to be an organism that is in the transitional state of partial concept possession (and hence partial self-consciousness).

### 3.3 Concepts and Non-Conceptual Content

Earlier I talked broadly about ‘intelligence’ and its connection with self-consciousness. Distilling these capacities to their essence is what leads to CPH as I have previously

expressed it, in terms of concepts. The core element of intelligent behaviour is concept possession. Concept possession is necessary for inferential reasoning and abstract thought and it permits rational behaviour (I explore rationality further in chapter 4). On the other side, self-consciousness as I have been using the term can be thought of as the possession of a self-concept, so concepts enter into both sides. Now, I take it to be uncontroversial that concept possession is necessary for my notion of self-consciousness as, obviously, an organism must be concept possessing in order to have a self-concept. My further claim is that concept possession is not only necessary but *sufficient* for self-consciousness, so I now present a brief discussion on concept possession.

Smith and Medin [1981] speak of the various views of concepts as *all* acknowledging that "...concepts have the twin functions of categorization and inference" (p9). I gloss this as the ability for *understanding*, that is, the ability to ascribe *meanings*[21] to mental representations. 'Meaning' here can be taken in the Gricean sense of 'non-natural meaning' [Grice 1957], which Grice describes as that which has a tendency to produce a belief in the subject. Of course we could attempt to define 'understanding' and 'belief' but that process would just continue to throw up other words that themselves demand definition. However, a formal definition of 'concept' is not critical for my purposes. It is not important as to (for example) which of the successors to the classical view of concepts is superior[22]; it is more important in the context of CPH to understand what it means for organisms to *possess* concepts. I discuss this issue in chapter 4 as it is necessary to get a grip on what it will take to identify concept possession in animals and human infants. For the ensuing argument, as a working definition of 'concept', I rely on the sketch above, which I take to be essentially our common, everyday usage of 'concept'.

[21] Let me clarify the distinction I make between *meaning* and another term often used in similar contexts: *content.* All representations have content – that is, *informational content*. They can all be analysed and described in different ways. For instance a visual image might be described in terms of curves and angles, or by specifying the colour and brightness (etc.) of individual 'pixels' at a given level of granularity. But having informational content does not imply *understanding*. Ascribing a *meaning* to a representation *does* imply understanding – in other words, the grasping of the concept represented. A more detailed and rigorous explication of much the same point of view can be found in Dretske [1981], particularly in chapter 7: *Coding and Content*.

[22] Whereas the classical view held a definitional view of concepts (i.e. membership requirements for particular categories can be explicitly defined), its successors (such as the *prototype; exemplar* and *knowledge* approaches) relied on more flexible conceptions. In these, concepts are characterised by prototypes or best examples (*prototype view*); or by one's total remembered experiences of category members (*exemplar view*); or by a more fuzzy approach in which categories and general knowledge covary as more information about the world is learned [Murphy 2002]. The focus in these discussions is more in regard to how new concepts are formed and modified and less on what it takes to be able to possess concepts in the first place, which is of greater importance to my project.

In the previous section I described a category of organisms (level 2) that are capable of mental states but are not self-conscious, and therefore (by CPH) incapable of possessing concepts. According to CPH these organisms do not *understand* things as such; they do not have *knowledge* (although they clearly do have informational states capable of motivating behaviours). Thus, an underlying assumption in CPH is the possibility of mental states existing in the absence of concept possession, a position that has been hotly debated in the literature. I analyse this debate below to validate this assumption.

**Non-Conceptual Mental Content**

The view defended here is the theory of *non-conceptual content*: the notion that "some mental states can represent the world even though the bearer of those mental states need not possess the concepts required to specify their content" [Bermũdez & Cahen 2012]. According to Gareth Evans [1982]: "The informational states which a subject acquires through perception are *non-conceptual*, or *non-conceptualized*. Judgements based upon such states necessarily involve conceptualization…" (p227). The first sentence is pretty much what I am asserting is the case with level 2 organisms, which do not have the capacity for abstract thought – that is, they are not concept possessing. An organism cannot be said to be concept possessing unless it is able to *ascribe meanings to mental representations*. In other words, they lack the capacity for metarepresentation of their first-order (non-conceptual) mental representations. A mental representation that has been ascribed a meaning represents a concept grasped by the ascriber. I return to my example of a predator perceiving prey. The cognition involved in this perception, if it may properly be called cognition, is on the level of *affordance*. The predator has no need for a concept of prey as such, it just needs to detect that it affords the action of feeding. Again, no concept of feeding is necessary on the part of the predator in order for it to engage in the activity. The leopard might perceive some prey and will try to capture and eat it. But it need not have the concepts of 'prey' or 'food'. It is behaving according to a stimulus/response paradigm, albeit of a more complex form than that displayed by (say) cockroaches. Level 2 organisms are able to perceive their environment but not to conceive it. I like to say that level 2 organisms only have a *perceptual field*, while level 3 organisms also have a *conceptual field*. I return to these ideas in section 3.4.

Conceptualism is the opposing view to that espoused above and is (roughly) the thesis that thought or mental content is constrained by the concepts possessed by the subject (e.g., Gennaro [1996]; Noë [1999]; McDowell [2009]). A consequence of this view is that an organism capable of *any* mental states must possess at least some concepts. If this view is correct then by CPH I would have to commit to the further consequence that any animal capable of mental states must be self-conscious. However, I have already argued that many animals (say, mammals) are most likely capable of mental representations (such as visual images) but it is implausible that they are all self-conscious. Below I present the case against conceptualism.

**The Conceptual Constraint**

To motivate the discussion on the conceptual constraint as given above, I focus on the depiction provided by Bermũdez & Cahen [2012]. According to Bermũdez & Cahen the plausibility of the conceptual constraint stems primarily from the following idea: "How a thinker, perceiver or speaker apprehends the world in *having beliefs* about it, *perceiving* it or *speaking* about it is a function of the concepts he possesses" (sec 2; italics added). This formulation of the conceptual constraint specifies three capacities, which, it is claimed, require concept possession: *believing*; *perceiving* and *speaking*. Thus, if a subject is capable of *any* of these capacities, then that would mean the subject possesses concepts (and hence, by CPH, must be self-conscious). I argue below that one of these three (*perceiving*) does not require concept possession and thus the conceptual constraint is false.

Before beginning the argument I want to highlight a general failing that I see in many of the positions taken by conceptualists: it is that the starting point for their analyses is usually the human being, which we already take for granted as concept possessing organisms. Indeed, much of the literature on the nature of concepts generally centres on *human adults* (e.g., Jackendoff [1992]; Murphy [2002]; McDowell [2009]). For example, McDowell [2009] asserts that "...we need to conceive *our* perceptual experience as an actualization, in sensory consciousness, of conceptual capacities" (p127, emphasis added)[23]. Right away it is assumed

[23] McDowell [2009] considers the type of perceptual awareness possessed by non-rational animals to be different to that possessed by rational animals (see his footnote 7 on page 134). Thus, we should perhaps classify McDowell as a conceptualist only as regards rational animals. McDowell makes several claims in regard to rational animals that are not inconsistent with (at least my form of) non-conceptualism. Among these

that concepts are at least available to the subject in question. In such cases the argument turns on whether any particular mental state (of a normal human adult) is necessarily constrained by the concepts that subject possesses. However, my concerns lie at the threshold of concept possession and questions whether non-human animals and human infants are even capable of possessing concepts. Thus the scope of the argument – at least for my purposes – must encompass organisms about which there is some doubt as to their conceptual abilities. The arguments that follow are presented in that spirit.

The capacities identified as potential threats to CPH in the passage quoted earlier are: (i) *speech*; (ii) *beliefs*; and (iii) *perception*. Two of these, *speech* and *beliefs*, can be quickly eliminated from our enquiries. All linguistic organisms possess concepts, and as such they must be categorised as level 3 organisms. This obviously applies to humans, but possibly also applies to sign language trained species (provided we are convinced that the subjects are truly using the signs as a language rather than simply as learned responses due to, say, associative learning). So, I would feel comfortable in categorising (for example) chimpanzees (which are thought capable of using sign language) as possibly level 3 organisms. Thus, I accept that the conceptual constraint *does* apply as regards the capacity for speech: one must possess concepts to be linguistic (though, as I argue in chapter 4, not necessarily the other way around). In regard to *beliefs*, it is uncontroversial (though not universal – see Stalnaker [1998]) that belief (or, indeed, any propositional attitude) requires concepts. Of course, as I explore further in chapter 4, in animals it is not easy to identify the existence of true beliefs as opposed to (non-conceptual) encoded information as both may motivate similar behaviour. Nevertheless, if an organism truly does possess beliefs then it must possess concepts and therefore must be categorised at level 3.

That leaves only the final capacity, *perception*, to deal with. Most people (including myself) are willing to ascribe the capacity for perception to practically the entire animal kingdom. But I am not willing to elevate the entire animal kingdom to the status of level 3 (i.e., concept possessing and hence, by CPH, self-conscious organisms). Thus, CPH is incompatible with the view that how an organism *perceives* the world is a function of the *concepts* the organism

---

are (i) that experience has *content over and above* what is invoked in explaining the belief and (ii) that experiential content *need not* lead to belief formation (p131). Both (i) and (ii) acknowledge that there must be some form of experiential content that is not involved in belief formation. In addition, (ii) also covers Brewer's [2005] assertion that "...sense experiential states provide reasons for empirical beliefs" from which standpoint he claims that sense experiential states *must* have conceptual content. But even though sense experiential states *can* provide reasons for empirical beliefs that does not mean that they always *must* do so.

possesses. In section 3.4 I argue that there is an *analogy* between perception and conception, which is illuminative in regards to the nature of conception, but next I present the case that the contents of perceptual states are (or can be) non-conceptual. The main existing arguments in this regard can be named as: *fineness of grain; domain-specificity; perceptual inconsistency;* and *non-conscious perception.*

**Arguments against the Conceptual Constraint for Perceptual States**

The 'fineness of grain' argument centres on the claim that perceptual experiences are far richer than can be captured by individual concepts. The usual examples are discrimination of hues in the visual mode (e.g., Evans [1982]; Tye [1995]) and musical intervals in the aural mode (e.g., Peacocke [1992]; Tye [1995]). We have concepts such as RED, BLUE, GREEN (etc.) and more finely defined concepts such as SKY BLUE, ROYAL BLUE, NAVY BLUE, (etc.) but humans can differentiate some 10 million or so different colours [Christie 2001] and we do not have concepts for each of them. Therefore, we must have some perceptual states that are non-conceptual in nature. McDowell [1994] argues that we do not have or need all these concepts in advance; we can capture the finest detail of all our colour experience by uttering demonstratives such as 'that shade', which gains its meaning on identity of a sample shade. However, Wright [2003] rightly points out that true concepts should be available for transfer and re-use, while the capacities McDowell relies on are not likely to last long enough to facilitate identification of other samples. Thus the fineness of grain argument remains a strong challenge to conceptualism.

The application of concepts can be thought of in terms of an ability to generalise (read this as a simple expression of Evans' [1982] famous *generality constraint*[24]). However, animals tend to interact with their environment using habitat-specific and probably innate behaviours not indicative of an ability to generalise. For example, Cheng [1986] showed that rat navigation primarily relies on geometric aspects of their environment despite a range of alternative cues such as brightness, texture and smell. Similar behaviours are observed in human infants [Spelke 1994]. Spelke concludes that as such behaviours have not been shaped by the infants'

[24] "if a subject can be credited with the thought that *a is F*, then he must have the conceptual resources for entertaining the thought that *a is G*, for every property of being *G* of which he has a conception" [Evans 1982, p104]

perceptual and motor experience they must be innate[25]. This 'domain-specificity' can perhaps be thought of as analogous in chess-playing computers as they are unable to perform other functions (such as play checkers). In the case of level 2 organisms, it is plausible that their perceptions allow them to flourish by triggering either innate behaviours suited to their environments or by stimulating advantageous responses learned by experience (or laboratory training). Concept possession in these cases need not be invoked to explain these behaviours.

It is generally thought that *conceptual contents must be consistent*; for example, one cannot simultaneously believe a proposition and its negation. An apparent counter example in which contradictory propositional attitudes seem to exist is the simultaneous desire 'to eat the cake and to *not* eat the cake'. However these can always be resolved as two distinct desires rather than one self-contradictory one (for example, a desire to experience the pleasure of eating the cake and a separate conflicting desire to avoid incurring the health detriment). But, unlike conceptual content, with perceptual content it seems inconsistency is possible, as with certain illusions such as the 'waterfall illusion' whereby a static image has the appearance of motion [Crane 1988]. Here the perception is self-contradictory, which violates the requirement of conceptual consistency. In response to Crane, Mellor [1988] points out that to even be subject to the waterfall illusion one must have concepts in the first place. I take Mellor's assertion to be correct but *not* to invalidate Crane's argument. To me, the important point of Crane's argument is that the content of the perception does not map isomorphically onto any propositional attitude; it matters not whether the subject is concept possessing. In the waterfall illusion the image appears to be both in motion and simultaneously not to be in motion. The former perception *inclines* one to believe the image is in motion while the latter *inclines* one to believe the opposite. But those are just the inclinations; the reality is that only one or the other belief will obtain.

Siegel [2011] describes an example of non-conscious perception that she attributes to MGF Martin. In this example someone searches for cufflinks in a drawer but does not notice them even though they are visible; later he is able to visually recall that they were there. If this is a possible occurrence, it implies that perception can occur without the accompanying formation of a concept (i.e. the concepts involved in the belief that cufflinks are in the drawer). Dretske

[25] Carey [2009] refers to these innate capacities as 'perceptual input analysers', produced by evolution rather than experience of the world. To Carey, concepts (or, 'mental symbols') are merely a *subset* of all mental representations – hence she implicitly endorses the need for an account of non-conceptual mental representations (p487).

[2006a] (amongst others) offers the phenomenon of blindsight as another example of this type: subjects with blindsight are able to unconsciously perceive objects that they claim they cannot see. Finally, similar effects can be observed in subjects undergoing the experimental technique of 'masked priming'. In masked priming experiments a subject is flashed an image which is too fast to be consciously perceived yet is nevertheless registered by the subject, as evidenced by its subsequent influence on the subject's behaviour. Since a conscious experience is lacking for the (unconscious) perception the subject cannot be applying a concept to it. Therefore, perceptions of this type must be considered non-conceptual.

**State vs. Content Non-Conceptualism**

The foregoing discussions provide reason to accept that perception is indeed unconstrained by the subject's possession of concepts and hence I stake my claim as a non-conceptualist. However, it is useful now to discuss a distinction between *state* non-conceptualism and *content* non-conceptualism. The following definition of state non-conceptualism is provided by Byrne [2004]: "state M with content p is a non-conceptual state iff it is possible to be in M without possessing all the concepts that characterize p." According to Heck [2000] there is nothing "unusual" in the idea of non-conceptual perceptual *states*, but there is still a burden of defence for the idea of non-conceptual *content*. What I argue below is that even in cases where perceptions align with propositional content (and hence are characterised by concepts) this does not necessarily imply concept possession by the subject. Organisms *incapable* of concept possession may still be in perceptual states the content of which is 'conceptual' (I use the scare quotes to indicate that I mean this in a very particular sense as explained below).

Another useful distinction to make is between *perception* and *experience* as these are sometimes conflated. Noë [1999] talks of *perceptual experience* in the following way:

> ...when I have a visual experience as of geese flying overhead, I exercise my knowledge of (among other things) what geese are, and what flying is. The experience is concept dependent because I could not have had just that experience as of geese and flying if I did not have those concepts. This is not to say that one needs the concept of a goose to see a goose. The point is that one could not see a goose *as a goose* or *as flying* if one lacked these concepts. (p257)

It is true that the *experience* described above is concept-dependent. But the crucial point, acknowledged by Noë, is that one does *not* need any concepts to *see* a goose, only to see a

goose *as a goose*. To have an *experience* 'as of a goose' one needs a concept of GOOSE, but one does not need this concept in order to *perceive* an object that happens to be a goose. It is likely (or at least plausible) that my *perception* of a goose is very similar to that had by (say) a fox, even if our respective *experiences* are very different. That is, an objective observer that is hypothetically able to see the visual image presented to me via my eyes and to the fox via its eyes would describe the image exactly the same way (using terms such as lines, angles, shapes, colours, etc). Thus, our *perceptions* (here construed in physical terms – say as the common retinal image) are more-or-less the same (constrained only by differences in human and fox eyes). But while I am having an experience 'as of a goose' (because I have a concept of GOOSE) there is no knowing what the fox is actually experiencing; at most we might be warranted to say of its visual image that the fox detects its affordance to be eaten (if it behaves accordingly). Let's assume for argument that foxes are *not* concept possessing organisms. Then, it could be said that the *content* of our common perception is what we humans can describe as the concept GOOSE, but only I am in that *conceptual state*; the fox shares the *content* but is in a *non-conceptual state*.

The point being made here is that even though the content in the example given can be described as 'conceptual' in that in contains the concept GOOSE, this does not mean that *all* subjects with that content can so describe it. Concept possessing organisms can describe the content in terms of the concepts involved, but some subjects can have the same content without possessing the concepts that describe it. This is essentially how I have characterised level 2 organisms: they can perceive and appropriately interact with objects in their environment but, unlike level 3 organisms, they have no concept of what they perceive. Concept possession is what separates level 3 organisms from level 2 organisms, just as it separates the relationalist side of the Fundamental Dichotomy from the intrinsicist side, as discussed in chapter 2. In the next section I argue that organisms that do possess concepts must possess the self-concept – which is to say (in the sense explained in chapter 1) they must be self-conscious organisms.

### 3.4 Self, the Fundamental Concept

I return now to the concept possession hypothesis. If my claim is correct, that concept possession is sufficient and necessary for self-consciousness, then this gives us our yardstick

for gauging the existence of self-consciousness in experimental subjects: if their behaviour conclusively indicates conceptual ability (such as demonstrations of abstract thought or inferential reasoning power) then they must be self-conscious. When evaluating the results of self-consciousness research in part 2, we can apply this principle to infer the existence of self-consciousness. To justify my claim, it remains to show that being concept possessing implies having a self-concept. This is the aim of the current section. I present three discussions to that end: an analogy with perception; an argument regarding the web of concepts; and an interpretation of Bermũdez [1998]. As mentioned earlier, the ideas expressed here are born of a common philosophical intuition that there is a correlation between 'intelligence' and self-consciousness, and it is this intuition that drives the three discussions below.

**An Analogy with Perception**

One facet of CPH is the impossibility of concept possession without possession of the self-concept: to be concept possessing at all necessarily means having a concept of the self. One way to view this idea is to think of the act of *conception* itself as always self-specifying. Below I illustrate what I mean by using an analogy with the act of *perception*. Self-perception, as I have previously stressed, does not count as a case of self-consciousness proper. However, perception provides an illustrative model for conception.

Anything that is perceived is perceived relative to the self; all perception is in this way self-specifying. For example, consider the visual mode of perception: whatever is seen bears a spatial relation to the perceiver. The position of the visual percept is specified in relation to the perceiver and hence the position of the perceiver is itself co-specified. Nothing can be perceived without the perceiver also co-perceiving itself. This applies not just to sight but to all of the other traditional senses as well. In taste, smell and touch the physical body is immediately presented to the perceiver through direct contact with the percept (i.e. the tongue for taste, the nose for smell and some other part of the body's surface for touch). In hearing, like sight, a distal object is discerned as being in a location relative to the perceiver. Again, self-perception does *not* mean self-consciousness as I have defined it. The self perceived is not the subjective self but rather the objective self ('objective' here meaning just in the sense of a physical object in the world). The perceiver does not necessarily perceive itself as a perceiver as such. An organism can perceive its own physical self in just the same (non-

conceptual) way as it perceives other objects in its immediate environment. The point is just that any perceived object must bear a spatial relation to the perceiver and in that way specifies the location of the perceiver relative to the other perceived objects. Thus, the bat echo-locates its prey and is able to close the gap between the prey and itself. This is so even if the perceiver has no concept of itself as a perceiver. J J Gibson describes this idea as "two sides of the same coin" in that information about the self accompanies information about the environment: "One perceives the environment and coperceives oneself" [Gibson 1979, p126].

I emphasise again that perception has nothing to do with conception and self-perception does not mean self-conception, but the notion of perception can act as a *model* for the notion of conception. Just as the perceptual field always includes the perception of the perceiver, the 'conceptual field' always includes the concept of the conceiver – that is, the self-concept. All other percepts in the perceptual field are defined in relation to the perceived self, and all other concepts in the conceptual field are defined in relation to the conceived self. Self-perception provides the *frame of reference* for the perceptual field. Analogously, the self-concept provides the *context* for the conceptual field. What a conceiver conceives is conceived relative to its own self-concept: the fact that a conceiver ascribes a meaning to something implies there is a conceiver for which it has that meaning. This is one way in which I view the self-concept as integral to any cluster of concepts. As discussed next, concepts are interrelated and always include a relation to the self-concept.

**The Web of Concepts**

Concepts do not exist in isolation: '...the grasping of a single concept requires the grasping of an entire body of concepts' [Brown 1986]. In other words all concepts are relational – they bear relations to other concepts. The much stronger claim I argue for in this chapter is that *all concepts bear a relation to the self-concept*. I assert that if an organism is capable of holding *any* concept it is must already have (also) grasped the *self*-concept. So, level 2 organisms have no self-concept (that is, they are not self-conscious) because they are not concept possessing. If a subject is not concept possessing then it cannot hold any concept including the self-concept. But my claim is that an organism cannot be concept possessing without having the self-concept. The self-concept is the *primary* concept, if you will. I do not mean 'primary' in the sense of 'epistemically foundational' for it is possible that (as Sellars

[1956/1997] argues) there are no foundational concepts[26]. I mean primary in the sense that, in whatever way the first web of concept arises in an organism[27], that initial web must include the self-concept.

My aim in this section is to convince the reader that all concepts are related to the self-concept. In the previous section I showed one way to look at it – as a necessary duality between any concept and the conceiver as a 'co-concept'. I now present some examples of how this connection between concepts and the self-concept might be seen as inevitable. By 'self-concept' I mean understanding one's own existence as an intentional agent as previously discussed in chapter 1. In this discussion we will assume that we are dealing with organisms that are at least conscious *of* their environment; they can perceive affordances and can interact with their environment accordingly. Now I need to show that if such an organism can grasp any concept, it must have already also grasped the self-concept. In essence, I am starting with at least level 2 organisms and showing how we can recognise that they are level 3 organisms. I proceed by way of an example.

Let's pick a simple concept and see how it might necessarily imply the self-concept. Consider the concept BLADE (imagine a caveman perceiving a flake of flint). If an organism has grasped the concept BLADE, that is, it understands what a blade is, then it must also have grasped the concept CUT[28]. Cutting is what a blade does – it is impossible to grasp the concept BLADE without simultaneously grasping the concept CUT. Cutting is an action; a blade does not cut of its own accord, it must be used for that purpose by an organism. So, along with the concepts BLADE and CUT necessarily comes the concept of ACTION.

[26] I remain neutral on whether foundational concepts exist. I began this section with the claim that no concepts exist in isolation, a claim that forms part of Sellars' [1956/1997] argument against epistemic foundationalism. Nevertheless, it is conceivable that there is one exception to the rule and that one isolated concept is the foundational concept on which all others are built. It would be consistent with my framework if that foundational concept is the self-concept.

[27] I do not tackle in this work the question of how the initial web of concepts (or, potentially, the foundational concept SELF) is able to form. Leaving this question unresolved does not affect my arguments.

[28] One might still ask at this point "but what is it to grasp a concept beyond being able to use a blade effectively?" In answer to this question I refer back to the previous section: use yourself as a model for concept possession – for example, consider your own metacognitive understanding of your actions in eating a festive goose and compare this to a fox that has no such understanding (but still eats the goose). Or, if you suspect that foxes are concept possessing, substitute a cockroach feasting on leftovers (or any 'lower order' organism that you consider non-concept possessing). A key point here is that although there may be an internal difference between a concept-possessor and a non-concept-possessor, externally (in terms of their behaviour) there might be no discernible difference. This issue is discussed in chapter 4 and is critical for part 2 of this work, in which we will need to rely on observed behaviour in order to determine concept possession in non-linguistic organisms.

Actions must be performed; there has to be a 'doer' or, more correctly, an agent[29], so the organism must also have grasped the concepts AGENCY and AGENT. Thus are we led from the simple concept BLADE through a chain of necessarily related concepts to the concept AGENT. If an organism has grasped the concept AGENT, then it must also have grasped the concept of the *self* as agent (as argued below). In other words it must hold the self-concept, in the sense previously defined, that is, recognition of itself as an Agentive Self. Note that I am not implying that an agent that is able to *use* a blade must therefore be self-conscious – a caveman raised in a blade-wielding tribe will have blades as part of his environment and may well be able to learn how to *use* a blade without having any *concept* of a blade. And I propose no theory here of how an organism might come to acquire the concept BLADE or indeed how it might come to be concept possessing. My argument starts with the assumption that an organism is concept possessing and has a particular concept (BLADE) and concludes that it must also have the concept AGENT and therefore also the self-concept.

If you are not completely comfortable with the final link, from the concept AGENT to the self-concept, consider the alternative. In that case, we would have an organism that has grasped all the previously mentioned concepts including the concept AGENT but without having grasped the concept of itself as an agent. Now, being an organism, it is a perceiver and an agent. It must therefore perceive its own agency. Of course, just perceiving its own agency does not mean it has a self-concept or any concept – an organism may be a perceiver and an agent and yet interact with its environment in a purely stimulus-response manner. But we have already assumed that the organism in our example is concept possessing and has grasped the concept of agency. To suggest that it perceives its own agency and has grasped the concept of agency but is yet unaware of itself as an agent is implausible. Presumably our organism is aware of other agents (let's again visualise a caveman and name him 'Homer'). Perhaps Homer has seen other cavemen using a blade in the action of cutting skins off animals. Homer, we have assumed, is concept possessing and *understands* what he sees; he *knows* that these others are agents performing an action; he has *grasped the concept* of agents with causal powers in the world. As he has the same agency himself and can perceive his own agency in the same way as he perceives the agency of others, he must be able to apply the same concept to himself as easily as he applied it to others. After all, as discussed in section 3.3 the ability to generalise in this way is a characteristic of concept usage. Further, as I

[29] As mentioned in chapter 1, 'agent' here can include both intentional and non-intentional agents. Later in the text I explain how *intentional* agency enters the picture.

discuss later, it is most plausible that Homer acquired the concept of AGENCY through his perception of himself as an agent.

In chapter 1 I defined self-consciousness as an understanding of one's own existence as a psychological subject with *intentional* agency. We have established that Homer is an intentional agent (because he is concept possessing and hence is able to deliberate and decide on his actions) and that he understands that he is an agent, but does he understand himself to be an *intentional* agent? The answer is yes. As discussed in chapter 1, intentional agency and metacognition are inseparable. If Homer is an intentional agent then he must know that he is – that is, he must know that he has some control over his actions else he would not be able to exercise that control (in which case he could not be said to be an *intentional* agent). So, Homer might not necessarily know if *other* agents he perceives are acting intentionally but he must know that he himself is an intentional agent (and, no doubt, will at least *assume* that his conspecifics are, too).

In the BLADE example, the description seemed like the link from BLADE to SELF was via a chain or a network where each concept is a discrete node, much as Deacon [1997] describes symbol-symbol relationships as a "tangled hierarchic network of nodes and connections that defines a vast and constantly changing semantic space" (p100). This is a simplification for the purpose of illustration, for concepts are clearly not so discrete and well-defined but rather overlapping and fuzzy. Rather than a 'web of concepts' probably a 'cloud of concepts' is a more accurate description, with the self-concept at the 'centre' or origin of the cloud. Then as the cloud grows and thickens with the addition of more concepts (and more abstract concepts), it grows outward from the central concept of self-as-agent. The linking of concepts like discrete nodes is intended as a useful simplification to demonstrate the interrelatedness of concepts and that this interrelatedness always integrates the self-concept, but I really want to suggest that the connection between other concepts and the self-concept is more direct. The following examples will bring this out.

BLADE is an example of what Bolton [1977] calls a physical concept. Being a physical concept allows a physical interaction, which implies a link with the key concept of AGENCY. However, I maintain we could start with any concept and will always find a link with the concept of agency. Bolton identifies at least three different types of concepts, physical, logico-mathematical and philosophical, and gives BEAUTY as an example of a philosophical concept. So how does holding this concept necessarily imply holding the self-concept?

Beauty, as the saying goes, is in the eye of the beholder – an organism has to apprehend beauty. There is no way to hold a concept of beauty without understanding that the object in question is beautiful to someone. Or, at least, to some*thing* – in any case, to an intentional agent capable of the action of apprehension. So, having the concept BEAUTY implies having the concepts AGENCY and AGENT. To add BEAUTY to its cloud of concepts the organism must already know that there is an agent that can apprehend beauty and the model for that agent is the self. An organism will acquire the idea of beauty based on its own reaction to apprehending beauty and will understand that any agent apprehending beauty will undergo that reaction. That does not mean that any organism that undergoes such a reaction – in a purely physiological sense – must have a concept of beauty. For example, an animal need not have a concept of FEAR for its fur to stand on end in the face of danger. But an understanding of what FEAR is – or what BEAUTY is – means an understanding that an agent is undergoing this reaction. And as argued earlier, having a concept of AGENT must mean a concept of self-as-agent.

Let's try another of Bolton's examples: COLOUR. Many animals have colour receptors in their eyes and so are able to perceive different colours and yet are not necessarily those we would consider self-conscious. Just being able to perceive different colours does not imply holding the concept COLOUR. It only allows the organism to interact with its environment and discover affordances where colour is important – say, in the discrimination between ripe and non-ripe fruit. The organism need not understand what it is doing when performing this discrimination (where here 'discrimination' means being stimulated to respond in different ways to different colours). But having the *concept* COLOUR, by contrast, implies understanding that an organism is *able* to discriminate between different colours. And once again discrimination of colours is an action that has to be performed by an *agent*. So, understanding that there are colours (or 'visually discriminable features of a perceived object') implies the corresponding grasp of the concept of AGENCY. The same may be said of any discrimination task: an organism may be able to perform it without knowing what it is doing, but if it *does* know what it is doing, if it has the relevant concept, then it must have the related concept of the doer, the agent. The claim, in effect, is that concept possession involves metacognitive awareness of being a concept possessor, as I now highlight with one final example.

In the case of abstract concepts like BEAUTY and COLOUR or any of the logico-mathematical concepts, it is easy to see the direct connection between the concept in question and the concept of self-as-agent, for these concepts require relatively sophisticated cognitive capabilities, and it is difficult to imagine an organism applying these concepts without knowing that that was what it was doing (and therefore perceiving its own agency in the process). I assert, however, that the principle applies to more primitive concepts also. Take a bland and perhaps primitive concept such as ROCK. A non-conceptual being might treat a rock of a certain size and (say) a tree stump the same way – for example, they both afford resting on, or climbing upon to get a better view (and so on). But an organism that has a concept of ROCK has this concept by dint of its ability to discriminate ROCKs from non-ROCKs. The act of discrimination requires an agent to perform (whether concept possessing or not). What I suggest is that when a *concept possessing* agent performs this act of discrimination (by identifying it *as a rock)* it knows that that is what it is doing. This is an aspect of the inseparability of conception and metacognition. Note that this is so whether the application of the concept is voluntary or involuntary. As Fodor [1983] remarks, "You can't hear speech as noise *even if you would prefer to*" (p53). But even so, the point is that when a concept is manifest, the agent *knows* that it is. A non-concept possessing organism might be able to perform an act of discrimination (in the sense mentioned in the earlier example of detecting the affordance of ripe fruit) but it cannot know that that is what it is doing. A concept possessing organism, on the other hand, *will* know and thus will be aware of its own agency in the process. This might be seen as a 'reflexivity requirement' on concept possession, reminiscent of Sellars' [1956/1997] reflexivity requirement on knowledge. According to deVries [2011], Sellars' idea of a 'battery of concepts' means that "any cognitive state...can be cognitive only as one element in a complex, reflexively structured system of such states responsive to epistemic norms and goals" (sec 4). In other words, Jones not only possesses knowledge (i.e. is concept possessing) but he knows when he is applying that knowledge.

**Objections to the Web of Concepts**

It will be seen that the web of concepts argument is a type of holism about concepts. Holism suffers from a well-known objection that it makes mutual understanding mysterious (e.g. Schwitzgebel [2011]). Furthermore, holism would seem to require a 'halting criterion' to

avoid indefinitely many concepts and justify the position that grasping some linked concept is not part of what it is to understand the focal concept (SELF)[30]. Holism about concepts comes in many forms and I aim to show in this subsection that my version of it avoids these objections.

The 'halting criterion' objection to holism gets its strength from the apparent lack of a boundary for concept possession. If grasping any one concept relies on grasping many others, which themselves depend on the grasping of yet many more, one could legitimately ask what the criterion would be to halt this process. The answer depends on the model of holism one subscribes to. The objection has greater validity in a model that sees the web of concepts quite literally, with the core concept (SELF) sitting in the middle like a spider and each concept transitively connected with each of its neighbours. But the picture is not so clear cut. In the first place, as mentioned earlier, concepts are not well defined and a perhaps better metaphor of concept possession would be a cloud rather than a web. However, sticking with the web metaphor for the sake of argument, there are alternative valid models for the web-of-concepts. It is likely that concept connectivity is not always transitive, so grasping some linked concept need not be part of what it is to understand the focal concept (the focal concept being SELF in my view). The connection between two concepts might be either bidirectional or unidirectional. Thus, it could be true that having concept Y necessarily entails having concept X, but not vice versa. For example, having the concept 'animal' does not entail having the concept 'coypu', but having the concept of 'coypu' does entail having the concept 'animal' (provided both concepts are reasonably accurate). Similarly, as I argue in chapter 7, while having the concept of self-extended-in-time *does* entail having the self-concept, having the self-concept does not necessarily entail having the concept of self-extended-in-time. As discussed in chapter 7 the case study of the amnesic 'KC' supports this view: KC lacks a sense of past and future but does not appear to lack a sense of his own existence as a psychological subject (which is how I have defined my usage of 'self-consciousness' since chapter 1).

The foregoing discussion is consistent with alternative models of the web of concepts. It could be that the model is like that previously described, as a single web with the SELF at its core. However, concepts could exist as independent, well-bounded clusters connected at only one point (the core concept SELF), so that the model resembles the petals of a daisy. Or,

[30] I thank Kim Sterelny for bringing these issues to my attention.

clusters could overlap like a Venn diagram. In any case, the main point is that a valid model of concept possession is of concepts existing as independent or partially independent, well-bounded clusters, thus avoiding the 'halting criterion' objection.

Schwitzgebel [2011] illustrates a holism worry with the example of Ani who thinks salmon are fish but so are whales while Sanjay thinks salmon are fish but whales are not. An *atomist* would say that Ani's and Sanjay's concepts of salmon are identical but a *holist* would say they are not. In an extreme case Ani's concept of fish might be outlandish and include all manner of non-fish creatures, in which case it would be fair to say that mutual understanding (in regard to fishy matters) between Ani and Sanjay would be unlikely indeed. But this is an extreme case and it would be wrong to say that mutual understanding is mysterious under usual circumstances. Conceptual understanding differs between subjects to different degrees and the level of understanding in communication between them will be a function of the degree of overlap in the concepts involved. When complex concepts are involved in conversations where there is little overlap between the participants' understanding of them, then mutual understanding will be low. In such cases we generally say that the conversants are 'talking past each other'. However, in communications involving everyday concepts one can expect a high degree of mutual understanding – which is not mysterious at all. For example, Ani probably has a reasonably good grasp of the concept FISH but simply has not learned that whales do not have gills and are in fact mammals that just look like fish and live in water. Despite this, Ani and Sanjay are likely to be capable of a coherent conversation about salmon and this high degree of mutual understanding is explicable simply based on their largely overlapping concepts of FISH.

**Bermũdez and the Paradox of Self-Consciousness**

Many of the ideas articulated by Bermũdez [1998] in *The Paradox of Self-Consciousness* are similar to those I have expressed here and, despite some differing views, his argument can be interpreted as broad support for CPH. Bermũdez conceives self-consciousness in terms of mastery of the first-person pronoun: "...self-conscious thought in the absence of linguistic mastery of the first-person pronoun is a contradiction in terms" (p. 28). The paradox for Bermũdez is based on a perceived circularity: what comes first, mastery of the first-person pronoun or self-conscious thought? Both seem to rely on the existence of the other. At the

heart of the paradox is an element of the classical theory of content that he calls the Conceptual-Requirements Principle: ascriptions of content to an individual are constrained by the concepts the individual possesses. This is an alternative formulation of the 'conceptual constraint' discussed in section 3.3. To break the paradox, Bermũdez rejects this principle and replaces it with a theory of representational non-conceptual content. He posits a 'primitive' form of (non-conceptual) self-consciousness, to be distinguished from what he calls 'full-fledged' self-consciousness. Bermũdez then devotes considerable effort arguing (convincingly, I think) for the existence of non-conceptual first-person content ascribable to creatures without conceptual or linguistic abilities.

Of course the foregoing is a highly condensed version of Bermũdez's very elaborate and fine-grained argument. Nevertheless, I believe I have captured the crucial elements. Bermũdez makes a distinction between 'full-fledged' self-consciousness and more primitive forms. It is only the full-fledged variety that equates to my conception of self-consciousness, for in both cases a self-concept is inherent. I acknowledge the existence of the non-conceptual, 'primitive' forms as developmentally intermediate forms of self-consciousness, but as discussed since chapter 1, these forms do not fit within the notion of self-consciousness I defined as my target for examination.

Bermũdez describes full-fledged self-consciousness in terms of linguistic mastery of the first-person pronoun, while for me it is concept possession that is the assumed necessary condition. But to Bermũdez these are more-or-less equivalent: "...there is a constitutive link between language mastery and concept mastery" (p. 70). Here I must admit to a point of departure with Bermũdez, for I do not wish to commit to the idea that concept possession depends on linguistic abilities (though I remain open to this possibility - I say more on this in chapter 4). However, either way, Bermũdez's conception is consistent with mine in that he sees concept possession as a necessary condition for full-fledged self-consciousness and demotes what he terms *non-conceptual self-consciousness* to a 'primitive' form (which, as previously mentioned, I am loathe to describe as 'self-consciousness' at all).

This non-conceptual primitive self-consciousness described by Bermũdez is equivalent to the cognitive capacity I ascribed to my 'level 2' organisms; those that are not concept possessing and so not self-conscious but which do have access to some of their mental states. As Bermũdez puts it, "…states with representational content can be properly ascribed to

individuals without those individuals necessarily possessing the concepts required to specify how those states represent the world" (p268).

So far I have shown a correspondence between Bermũdez's ideas and my own including the notion that concept possession is a necessary condition for (full-fledged) self-consciousness. As previously stated, I view this as relatively uncontroversial as it is obvious that an organism cannot have a self-concept if it is unable to hold any concepts. The further claim I make in CPH is that concept possession is *sufficient* for self-consciousness, which, although not explicit in Bermũdez, can be inferred from his line of argument. Compare creatures with full-fledged self-consciousness and those with only the primitive non-conceptual kind. The only thing that separates them is concept possession. Both have access to internal representational states with self-specifying content, but only the full-fledged self-conscious creatures are concept possessing. It follows from this that the addition of concepts to the 'primitively' self-conscious creatures is sufficient to confer on them full-fledged self-consciousness. On this reading of Bermũdez his analysis provides clear support for CPH and is reminiscent of the distinction made in chapter 2 regarding the Fundamental Dichotomy and later again in this chapter as the top two tiers in the 'three levels' taxonomy of consciousness.

Of course the same issue arises here as for the 'three levels' taxonomy: that perhaps there is an intervening transitional stage, in which organisms attain concepts but not full-fledged self-consciousness. But this state does not exist in Bermũdez's scheme: concept possession plus self-specifying representational content is sufficient for an ascription of full-fledged self-consciousness. Again, if a transitional stage exists whereby the organism acquires only *proto*-concepts, then I would say this organism has correspondingly attained only *proto*-self-consciousness, and not the (full-fledged) self-consciousness I am interested in.

### 3.5 Conclusion

It is common to see self-consciousness spoken of in terms of a self-concept. Of course it is clear that in order to have a self-concept (or any concept) one must first be concept possessing. What distinguishes my position is the further claim that concept possession alone is sufficient and necessary for self-consciousness. I argue that if an organism has any concept at all, it must also have (at least) the self-concept. To illustrate how to look at this idea I

provided an analogy between perception and conception. Just as in perception the self is co-perceived, so it is that in conception the self is co-conceived. The network of concepts possessed by an organism necessarily contains the self-concept.

Thus, the defining property of self-conscious organisms is concept possession. As such, evidence of concept possession in animals or human infants should be regarded as an indication of the existence of self-consciousness. Of course, demonstrating concept possession in non-linguistic organisms is itself very difficult so it might seem we are no better off with CPH. However, I maintain that there are ways to determine concept possession and some experimental paradigms will be suitable in this regard. In chapter 4 I conclude part 1 of the thesis with a discussion on what should count as evidence of concept possession. In part 2 I put these ideas to work in analysing empirical research on self-consciousness in animals and human infants.

# Chapter 4: Evidence of Concept Possession

## 4.1 Introduction

In the previous chapter I presented the Concept Possession Hypothesis (CPH), which claims that concept possession alone is sufficient evidence for the existence of self-consciousness. In this final chapter of part 1 I expand on some of the points raised in chapter 3 in regard to how we might determine the existence of concept possession in a subject as well as what should *not* count as sufficient evidence of concept possession. These ideas will be put to work in part 2 where I analyse experiments on animals and human infants aimed at discovering whether self-consciousness is present. I also here briefly discuss the special case of language. To some authors concept possession depends on language ability (e.g., Sellars [1956/1997]; Deacon [1997]; Bermũdez [1998, 2003, 2006]). If this is true then by CPH this would imply that self-consciousness likewise relies on language and consequently animals are precluded from the capacity for self-consciousness. I have already declared myself neutral in earlier chapters on whether there is such a dependence between concept possession and language. Nevertheless, in section 4.2 I argue that it is at least plausible that concept possession does *not* rely on language so we should not discount the possibility of self-consciousness in non-linguistic organisms out of hand.

Regarding the task of determining the existence of concept possession, I suggest four alternative approaches: (i) identify a specific individual concept grasped by the subject; (ii) show evidence of propositional attitudes; (iii) show evidence of rationality (e.g., via demonstrations of inferential thinking); or (iv) demonstrate symbol-mindedness. I examine each of these in section 4.3 and I argue that it is this type of evidence we should be looking for when analysing the experimental results presented in part 2. In section 4.4 I argue that we need to maintain a high standard of evidence. For non-linguistic and pre-linguistic organisms we only have observations of behaviour from which to make a judgement on the possibility of concept possession. Animal behaviour is subject to multiple interpretations and where behaviour can be plausibly explained using non-cognitive accounts we should discount it as evidence of concept possession. I present several examples of behaviour that I argue can be explained in non-cognitive terms: (i) programmatic; (ii) occurring as a result of associative learning; and (iii) explicable as ‘hard-wired’ or ‘innate’ (the latter two being specific

examples of the former). Finally, I present a simple 'test case' (tool usage by animals) to illustrate the process.

### 4.2 The Language Question

If, as some claim (e.g., Sellars [1956/1997]; Deacon [1997]; Bermũdez [1998, 2003, 2006]), conceptual thought requires language, then non-linguistic organisms are incapable of conceptual thought and as such, by CPH, are precluded from the possibility of self-awareness. Earlier I declared myself agnostic on the issue of language as I do not believe that the question is settled. Doubtlessly, language plays an important role in human cognition (e.g., Carruthers [2011]) but at issue is whether *lack* of language necessarily implies lack of conceptual abilities. In this section, while remaining neutral on the language requirement thesis, I present some arguments as to why it might be incorrect or, even if correct, why that should not prevent us investigating the possibility of self-awareness in animals.

The proposition being examined here is the possibility of conceptual thought in the absence of linguistic ability. All writers (and indeed speakers), I am sure, are like myself in having suffered a situation of knowing 'deep down' what they want to say but being unable to find the right words or right expression to properly capture the thought. Carruthers [2009b] describes some empirical evidence in his example of 'introspection sampling'[31] studies in which subjects report the presence of 'propositional yet unsymbolised' thoughts. For example, subjects reported "...thinking something highly determinate – such as wondering whether or not to buy a given box of breakfast cereal – in the absence of any visual imagery, inner speech, or other sensory accompaniments" (p134). Thus there seem be ways of knowing, believing, etc., without the occurrence of a natural language expression of the thought – even, perhaps, without an equivalent expression in Fodorian 'Mentalese'[32]. For example, Camp [2007] considers the possibility of conceptual thought with map-like structure rather than language-like structure (i.e., in which geometric relations between map elements are conceptually understood)[33]. Even Bermũdez [2003], in some sense, concedes

[31] Subjects wearing a paging device that beeps at random intervals are asked to 'freeze' the content of their consciousness at the very moment of the beep and report on it.
[32] That is, a 'language of thought' structured more-or-less on the lines of natural language [Fodor 1975; Aydede 2004].
[33] Although worth considering this as a possibility, this should not imply that *all* map-like thought need be conceptual. It might be possible for there to be both conceptual thinking as well as non-conceptual thinking

this: "The proposal that a thought be analyzed through the sentence that expresses it does not entail that any thinker capable of thinking that thought should be able to express it" (p19). Think of 'tip of the tongue' experiences of single words that perfectly capture a concept in a way that no other word can but is not recalled. This can also happen with entire propositional thoughts – sometimes one just cannot express a thought, even though it is in principle expressible as a linguistic proposition. If that were not the case could we ever come up with new concepts? After all, it is unlikely that all new concepts arise because of the coincidental advent of a sentential expression that encapsulates it (though I'm sure that can happen also); presumably a new concept can somehow form in someone's mind prior to the formulation of its linguistic expression. Given this, it seems at least reasonable that conceptual thought might be possible in the absence of language.

There is some evidence based on studies of deaf children that certain concepts are harder to form in the absence of language [Mayberry 2002]. This somewhat supports the view that conceptual thought requires language. But even so, with these subjects we can accept that their conceptualisation capacities may be impaired without concluding that they are *totally* incapable of concept possession due to lack of natural language. In fact, deaf children, as a group, appear to follow the same Piagetian stages of early conceptual development as their hearing peers despite pervasive language delay [Mayberry 2002]. Rakison [2007], citing infant studies on labelling, suggests that the emergence of accessible declarative knowledge at around 18 months of age depends on already established perceptually-based concepts rather than the faculty to think using language. Thus, the empirical evidence does not conclusively support the view that conceptualisation is impossible without language.

Even if it were the case that conceptualisation is impossible without language, it has been argued that so-called non-linguistic animals do in fact possess some form of language unfamiliar to us but still sufficiently structured as to satisfy the requirement for conceptualisation. In other words, some animals may communicate in something closely akin to natural language, for example, via primate vocalisations or cetacean auditory signals. Marler, Evans & Hauser [1992] have suggested that some animal alarm calls might be referential in that the same species will use different alarm calls according to the type of (or even *specific*) threat. For example, Gunnison's prairie dogs give different alarm calls for

based on map-like mental content. Indeed, as mentioned in section 3.2, map-like thoughts might be the bridge between non-conceptual and conceptual thoughts.

humans, hawks, and canines [Andrews 2011]. Parrots in the wild have demonstrated vertical transmission of socially acquired signature calls, which look enticingly like the act of offspring naming [Berg et al. 2011]. Additionally, of course, there is the possibility that animals (e.g. chimpanzees) trained to employ sign language might indicate inherent linguistic ability. Although I am not fully convinced that these examples should be considered language-like, it is at least questionable whether all non-human animals are indeed non-linguistic. This therefore provides another reason to stay open to the possibility that so-called non-linguistic organisms do indeed possess concepts.

Given the foregoing discussion, while remaining agnostic on the 'language question' (i.e. as to whether language is pre-requisite for concept possession) I submit that there is still good reason to investigate the possibility of concept possession in non-human animals.

## 4.3 Positive Indications of Concept Possession

In chapter 3 I discussed the differences between concept possessing and non-concept possessing organisms. In this section I expand that discussion with the aim of providing definitive tests of concept possession in animals and human infants.

### Identification of a Specific Individual Concept Grasped by the Subject

This method takes a direct and rather obvious approach as compared with the other three, which are indirect ways to demonstrate concept possession. Simply put, to show the capacity for concept possession in a subject we should find direct evidence of at least one actual concept possessed. This is of course very difficult to do since (for example) animals might behave as if they possess a concept even when they do not. In the first place we must discount 'concrete' concepts such as FOOD or any other physical objects in the subject's environment since all animals naturally interact with these. As briefly discussed in chapter 3, an animal does not need a concept of FOOD in order to interact appropriately with it. Instead, we must find evidence of the subject grasping an abstract concept. Even so, we will still be hard pressed to prove our case, as in the example of the abstract concept BEAUTY also discussed in chapter 3. In that example I pointed out that a non-conceptual subject might react

(physiologically) in exactly the same way to the apprehension of beauty as would a conceptual subject and we might not be able to tell the difference simply by observation.

Despite this difficulty and even when maintaining a very high standard for evidence, I argue that there are cases where concept possession is the best explanation for observed behaviour. I argue in chapter 7 for one such case. In chapter 7 I review the connection between self-consciousness and episodic memory and I also critically evaluate many experimental paradigms purporting to demonstrate a sense of subjective time in animals. Although I discount almost all of them, I argue that in one particular experiment the observed behaviour of scrub jays provides convincing evidence that they have a concept of their own future existence.

**Thoughts in Propositional Form**

In chapter 3 it was accepted as relatively uncontroversial that beliefs and other propositional attitudes imply concept possession. As such, propositional attitudes could be used as a way to demonstrate self-awareness. Indeed, any thoughts in propositional form should be taken as evidence of concept possession. Humans, of course, routinely demonstrate this capacity by expressing their propositional thoughts using language. Demonstrating the existence of propositional thoughts in non-linguistic organisms, however, seems impossible. In chapter 3 I made the point that even if the content of a mental state could be described as 'conceptual' in the sense that an expression of the content contains concepts, that does not imply that the subject in question is concept possessing. In other words, even though we humans can express the content of a subject's mental state as a proposition, that does not mean the subject in that state is likewise able to do so. For example, let's say a pigeon in a Skinner box has been trained to recognise a specific shade (say, royal blue), such that the pigeon can obtain a food reward by pecking a button of that particular shade. It makes sense to describe the content of the pigeon's mental representation using the concept ROYAL BLUE; for example, in the form of the proposition that 'pecking the royal blue button produces a food pellet.' But that is how *we intelligent and self-conscious organisms* would describe the pigeon's mental content; not necessarily how the pigeon itself would. We have the requisite concepts involved in that proposition, and we surely are at liberty to describe the content in those terms between us. But that does not mean that the pigeon itself has those concepts or could in any way

entertain (linguistically or otherwise) that *proposition*. We can only be sure that the pigeon's brain has encoded the information that motivates the observed behaviour (i.e. the affordance of the royal blue button to produce food); this says nothing about whether the pigeon is able to *understand* what it is doing. There is no evidence in this example that the pigeon had a propositional thought. This is not to say that for me pigeons will forever remain categorised as 'level 2 organisms' as described in chapter 3. My point here is that the behaviour described above is not sufficient evidence of propositional thinking and to thus categorise pigeons as (concept possessing) level 3 organisms.

As yet, in the absence of language, I am unable to see how we could definitively demonstrate that an organism has thoughts in propositional form. Despite this, some researchers do in fact make that claim. In chapter 8 I examine a claim that rats have propositional-like thoughts [Dickinson 1985] and reach the same negative conclusion.

**Rationality**

Concepts allow inference, or reasoning power [Smith & Medin 1981]. That is, the ability to solve problems or make decisions through reasoning. Therefore, by CPH, all truly rational entities should be considered self-conscious. However, there are multiple conceptions of rationality, so if rationality is to be a criterion for determining the existence of self-consciousness we need to ensure a common understanding of what it means to be rational. I have made it clear that what I believe to be the correct usage is one which indicates that a process of conscious reasoning takes place, implying concept possession. Examples are: deductive reasoning ("All men are liars; King David is a man; therefore King David is a liar"), inductive reasoning ("the street is wet, which usually happens after rain, therefore it must have rained this morning") and abductive reasoning ("the best explanation for why the couple are kissing in public is that they are in an intimate relationship"). The examples are intended to illustrate the processes involved, but once again I emphasise that in the case of animals I do not assume that the reasoning process necessarily must occur in those explicit linguistic forms.

Taking my lead from Kacelnik [2006], I focus here on this view of rationality and compare it with some alternative notions. This will be a guide to evaluating some others' usages of

'rationality'. Note the distinction between *non-rational* and *irrational*. Non-rational means no process of reasoning takes place while irrational means a *faulty* process of reasoning takes place. Thus both rational and irrational behaviour fall within the 'space of reasons' because, faulty or otherwise, a conscious process of reasoning is involved. Thus, even irrational behaviour counts as evidence of concept possession.

The various views on rationality can be fairly well categorised into two broad categories: those that focus on the processes involved and those that focus on the outcomes. The former category has been named 'PP-rationality' (for 'philosophical/psychological' usage) by Kacelnik [2006], to be understood in terms of processes such as thoughts and beliefs. According to Kacelnik, "To judge whether behaviour is PP-rational one needs to establish if it is caused by beliefs that have emerged from a reasoning process" (p89). This is like my usage of rationality, since having beliefs implies concept possession. As such, I endorse Kacelnik's 'PP-rationality' as the proper conception of rationality, at least in regard to applying CPH to determine the existence of self-consciousness.

Kacelnik contrasts PP-rationality with E-rationality (economic rationality) and B-rationality (evolutionary biological rationality). E-rationality focuses on whether behaviour is consistent with the 'maximisation of utility' (interpretable as maximisation of energy efficiency in the context of ecology) and B-rationality focuses on behaviour driven by genetic pre-dispositions (itself determined by E-rational natural selection). Both E-rationality and B-rationality are outcome-based and fall into the second broad categorisation mentioned above. There is no requirement for concept possession in E-rationality or B-rationality. Neither function has any dependency on a process of *reasoning*, so they hardly deserve to be characterised as any type of *rationality* at all. With E-rationality and B-rationality we can say that there are *reasons* for behaviours, but this is to use 'reason' in a different sense, much as we ought to use 'cause'. A non-conceptual animal is incapable of *reasoning*, although its perceptions may be a *cause* of its actions. The distinction is obviously crucial for my purposes since without concept possession in the subject there is no case for self-consciousness based on CPH. There are other variations on what counts as rationality that do not all fall neatly into Kacelnik's taxonomy. I briefly review these below and explain why none of them meets the criterion required by CPH.

Millikan [2006] has argued that reasoning is just "trial and error in thought" (p118) at the perceptual level. I do not think this characterisation properly captures rationality. Millikan's

example is a squirrel in her yard that surveyed a bird feeder from several angles apparently in order to determine a way to reach it. Millikan herself doubts that the squirrel engaged in any propositional thinking and thinks it was "trying to *see* an affordance" (p119). As no concept possession is required to explain this type of behaviour, it does not count as PP-rationality. Perhaps it would fit under B-rationality, as the animal used its inherent motor skills in its multiple attempts to reach the feeder.

Dretske [2006b] introduces a notion of 'minimal rationality' which he claims is more demanding than B-rationality but less demanding than PP-rationality: "Minimal rationality requires that what is done be done for reasons, but it doesn't require it be done for *good* reasons. Nor does it require reason*ing*" (p108). According to Dretske minimal rationality is under the control of thought and so is available to animals but not available to plants and machines. This conception is plausible enough but not useful to our cause. No doubt there are gradations in cognitive capacities, but in keeping with our principle of maintaining a high bar for concept possession, we need to apply a more stringent conception of rationality for a demonstration of self-consciousness. For our purposes, reason*ing* must be involved.

Bermũdez [2003] also grades rationality into levels (not to be confused with my 'levels' taxonomy of organisms as described in chapter 3). For Bermũdez, level 0 rationality represents 'hard-wired' behaviour where no decision-making is involved. Level 1 represents the ability to select from alternative possible actions, but still not real decision making; level 1 seems to be akin to Millikan's conception, in which the subject just 'sees' the appropriate action as an affordance. Finally, level 2 rationality involves decision-making. The problem Bermũdez has is his commitment to the view that inferential reasoning requires natural language. In order to allow non-linguistic creatures the capacity for level 2 rationality, Bermũdez is forced to postulate developmentally intermediate forms of (for example) beliefs and logic, which he refers to as 'proto-beliefs' [Bermũdez 1998] and 'proto-logic' [Bermũdez 2006]. I do not deny these intermediate forms, as it is likely that capacities like belief and logic are not 'all-or-nothing' and must admit of a gradation. As mentioned in chapter 3, to the extent that there is such a thing as proto-belief or proto-concept, there comes with it 'proto-self-consciousness'. But the danger with this way of thinking is that it threatens infinite regress. Do we also need to postulate proto-proto-belief? Despite the likelihood of a gradation of cognitive capacities, it behoves us to keep a sharp distinction between levels to retain a sense of what it means to achieve any particular milestone. Of course this position will leave

unexplained the question of how the transition between levels occurs, but this is not our primary concern here. For our purposes, it is better to eschew proto-capacities and instead raise the burden of proof to a high platform. Unlike Bermũdez, and as previously discussed, I am not committed to the notion that inferential reasoning requires natural language so I see no impediment to the possibility of inferential reasoning by non-linguistic creatures. Accordingly, I characterise Bermũdez's 'level 2 rationality' as *decision-making through inferential reasoning* (i.e., matching the notion of PP-rationality).

Hurley [2006] talks of non-human animals as occupying 'islands of practical rationality', which she describes as domain-specific reasons for action despite a lack of conceptual abilities. According to Hurley, flexibility and generality (hallmarks of concept possession) come in degrees and can be present in intentional agency even when they are domain-specific. This view seems at odds with CPH, which holds, as argued in chapter 3, that intentional agency implies concept possession. However, Hurley concedes that her stated position is a 'notational preference' and perhaps concepts themselves come in degrees, so that she is content to recharacterise her position as "...creatures without *full* conceptual abilities can have reasons for action" (p151; emphasis added). Thus, she allows some level of conceptual ability for intentional agency. This view seems reminiscent of Bermũdez's 'proto-capacities' and I have already acknowledged the likelihood that concept possession is not an all-or-nothing capacity. Nevertheless, as just argued, invoking proto-concepts can lead to a slippery slope down the cognitive scale. The way to avoid a slide down is to maintain a high standard of evidence that avoids reliance on proto-concepts.

To summarise, it is best to keep the bar high with respect to a demonstration of rationality and Kacelnik's [2006] conception of PP-rationality is the appropriate one. This requires we only accept clear cases of inferential reasoning. The examples of other views of rationality presented above by Millikan, Dretske, and Hurley do not seem sufficiently stringent for the reasons given. Of course, I am even less willing to accept 'rationality' conceived as 'B-rationality' or 'E-rationality'. In cases where (for example) experimental psychologists apply the term 'rationality' to animals in either of the latter senses there is no implication of self-consciousness. However, where the sense of *PP-rationality* is clearly intended, by CPH self-consciousness is implied.

Once again, it would appear to be very difficult to demonstrate rationality to this standard, but once again such claims have been made by researchers. In chapter 8 I analyse several

claims by experimental psychologists that rats are rational (in the sense of PP-rationality) and argue that the observed behaviour does *not* warrant this conclusion.

**Symbol-Mindedness**

As discussed in chapter 3, concept possession can be thought of as the ability to ascribe meanings to representations, which is one way to describe symbol usage. Thus, proper usage of symbols is a demonstration of concept possession. Judy Deloache [2004], studying symbol recognition in young human infants, offers the following working definition of 'symbol': *a symbol is something that someone intends to represent something other than itself*. What denotes something as a symbol in DeLoache's view is the symbol's *intentional* nature. That is, what makes a symbol a symbol is the specific intention of the symbol's creator that it is taken as a symbol. Of course this means it is a symbol in the eye of the symbol creator. One can make an equivalent definition to represent the viewpoint of a symbol perceiver. I offer the following definition to capture this perspective: *a symbol is something that someone has taken to represent something other than itself*[34]. The key element here is the understanding that the perceived object is not an *instance* of the represented object. For example, a road sign depicting a predator might startle one of that predators' prey if it mistakenly took it as a silhouette of an actual animal. This would of course not be a case of symbol usage by the prey in question, for it has taken the image as an instance of the predator. For human adults driving by and recognising it as only a road sign depicting an animal to be avoided on the road, it would be a case of symbol usage. This definition of symbols is by no means the only one possible[35], but it is this specific aspect that is relevant to concept possession.

As discussed in chapter 3, concept possession can be characterised as the ability to *ascribe meanings* to mental particulars, such as images of perceived objects in the visual field. Concept possession is necessary for symbol usage, since recognising something as a symbol

[34] DeLoache [personal communication] has indicated that her definition is meant to cover both the creator and perceiver perspectives. However, I see them as quite different. From the perceiver perspective there is usually no *intention* as such – the object just is or is not taken as a symbol.

[35] For example, Deacon [1997] describes three types of 'symbol' (icon – the object's form resembles the referent; index – the object correlates with the referent; and symbol – the object represents the referent by convention). Each of these meets our fundamental criterion here in that the object stands in for something else. Sterelny [2012] notes that there is a distinction to be made between public symbols important in the social/cultural context and the sense in which an *individual* is able to understand symbols (which he agrees is a 'signature of cognitive sophistication'). For our purposes in determining concept possession, the latter is the relevant sense.

implies ascribing a meaning to it. For example, an organism might see a pie and detect its affordance to be eaten – possibly without ever having a concept of food or eating. The organism does not need to ascribe the meaning 'food' to the visual image of the pie in order to appropriately interact with it. The pie is not taken a symbol; it does not represent something other than itself. But another organism might see π and recognise it as standing for a number with particular geometric properties – that is, standing for something other than itself (a squiggle[36]). In this case, the organism is symbol-minded and has treated the visual image as a symbol. In order to treat it as a symbol – that is, to ascribe a *meaning* to the squiggle – the organism must be concept possessing[37].

Symbol-mindedness provides a practicable paradigm for determining concept possession. In chapter 5 I explain why subjects that show mirror self-recognition (MSR) are symbol-minded and therefore concept possessing, and from this why MSR should be taken as conclusive evidence of self-awareness. But here it will be instructive to consider DeLoache's symbol-mindedness experiments on human infants, for two reasons. Firstly, this will elucidate the nature of symbol-mindedness, which will be useful when applied to chimpanzees in chapter 5. Secondly, it will provide an age-based comparison with the onset of MSR in children.

DeLoache [2004] studied symbol-mindedness in infants up to 3 years of age. When the infants treated symbol-objects as if they were 'real' objects, they were considered to be not yet symbol-minded. For example, a 9-month-old infant placed his lips on a photograph of a baby bottle. There are other behaviours that also correspond to a lack of symbol-mindedness, such as manually exploring the symbol-object or even attempting to grasp the depicted object in a picture. DeLoache describes one experiment in which a scale model of a room is used as the symbolic object. The experiment was designed to test whether the subject would take the scale model as a symbol representing the actual room that it replicated in miniature. The infants observed experimenters hiding a miniature toy within the scale model (for example, behind a miniature piece of furniture) and were then asked to locate the actual toy in the real room. Children who have achieved symbol-mindedness recognise the miniature toy and scale

[36] This example highlights the point made earlier that we need to consider symbol usage from the perspective of the *perceiver* and not just the symbol creator, for the squiggle may have been unintentional. For instance, a squiggle resembling the Greek letter π may have been nothing more than an infant's doodle, yet later taken as π by an adult.

[37] I make a distinction between 'representation' and 'symbol'. The content of a mental state may be a representation (for example, a visual image corresponding to a seen object) but this need not imply concept possession. This is the reason that I emphasise 'symbol' rather than 'representation': in order to argue for concept possession, the representation must be ascribed a meaning by the possessor (i.e. taken as a symbol).

model as symbols[38] representing the life-size toy and life-size room and are able to locate the hidden toy within the life-size room. The symbol-minded child is able to act upon the meanings of the symbols in this experiment: the miniature toy hidden behind a miniature chair in the scale model of the room means that the actual toy is hidden behind the actual chair in the actual room.

As one might well expect, there is no definite cut-off age at which an infant suddenly becomes symbol-minded. Apart from the obvious variation in developmental progress between different children, the same child will perform differently on symbol-mindedness tests depending on a variety of factors. For example, in the hidden toy experiment children who did not fully appreciate the symbol-referent relationship between the scale model and the actual room obviously performed poorly at the test. However, their performance could be improved by decreasing the salience of the model as an object by placing it behind a window (and hence presumably increasing its symbolic nature). Despite the inability to narrowly define the age range at which children become symbol-minded, the evidence suggests an age range centred more-or-less around 18 months. DeLoache says that by this age children cease manual exploration of photographs and "…point to and talk about pictured objects instead" [DeLoache 2004, p68]. I do not think it a coincidence that this age range is about the same as when infants generally achieve mirror self-recognition, as I explain in chapter 5.

## 4.4 The Standard of Evidence

It comes as no surprise to either the ethologist or the layman that an organism flees in the face of danger or fights when cornered. Neither should it be a surprise that foraging activity is usually optimised for energy efficiency – we would expect that as a consequence of natural selection. What may be a surprise is to discover that a non-human animal is capable of conceptual thought in the way we humans are. What we should look for is evidence of this type of thought, that is, thought that involves conceptualisation as discussed in section 4.3. The task is to detect behaviour that implies the existence of *knowledge* rather than mere encoded raw *information*. I argue below that while only the former implies concept possession the latter is still available to motivate behaviour. Being able to tell the difference

[38] Perner [1991], perhaps, would rather call the scale models 'analogues'. Even so, these remain examples of symbols according to the way DeLoache and I have characterised them. That is, the subjects treat them *as* symbols by regarding what they see as standing in for something else.

between knowledge and encoded information in non-linguistic organisms is difficult. As a general principle, we should keep the bar high with respect to ascriptions of higher cognitive capacities to animal and human infant subjects. This is of course the principle of parsimony, or the oft-quoted 'Morgan's canon' that exhorts us to interpret animal behaviour according to the lowest psychological faculty feasible. In this section I suggest specific principles to guide the analysis of animal behaviour. In summary, we should discount behaviour that can be explained as programmatic, specific examples of which are those occurring as a result of associative learning or those explicable as 'hard-wired'. I finish this section with a short case study to illustrate the process I advocate.

**Programmatic Behaviour**

Programmatic behaviour is that which occurs as the result of a set of embedded 'rules' or 'instructions' to respond in certain ways to given stimuli. Programmatic behaviour occurs automatically and does not rely on concept possession. Thus, if an organism's behaviour can be plausibly explained in this way we cannot use this behaviour as sufficient evidence of concept possession.

The paradigmatic example of programmatic behaviour is, of course, that displayed by domain-specific computer programs such as chess computers[39]. Chess computers are not concept possessing, unlike (say) humans. We can characterise this difference as that between *knowledge* possessed by a human being and *raw information* encoded into a chess computer. With knowledge, a human can make inferences and draw conclusions. For example, if I see several persons walking on the street dressed in colourful uniforms and carrying instruments, I might infer that they are part of a marching band. Then, if it is still early, I might conclude that a parade is yet to begin, and this might motivate me to remain at my cafe table for longer and enjoy the entertainment. I have used knowledge about marching bands and time of day (etc.) to undergo a process of reasoning and allowed the results to motivate my behaviour. I have *conceptual understanding* of the information input, which allows me to reason

[39] I am of course aware of alternative computer architectures such as those based on neural networks. These are not included as examples here as they are usually designed with the intention that they *not* behave according to a set of pre-programmed instructions like chess computers and so are not paradigm examples of programmatic behaviour. I also do not discuss the wider implications opened up by the advent of artificial neural networks as the topic extends beyond the scope of this work. Nevertheless, I do touch upon the subject briefly in chapter 9 in the section on Future Research.

inferentially to draw conclusions. This is how I prefer to use 'knowledge' and why I distinguish it from the *raw information* that can be encoded in a computer. A chess computer, for example, will have encoded into its memory banks representations for the chessboard and the pieces, and it will also have an encoded set of rules. But it has no concept of ROOK or PAWN or for that matter CHESS or even GAME. In the marching band example I underwent a process of reasoning using something like *modus ponens* to reach the conclusion. By contrast, a chess computer has no understanding of any datum of information encoded; no ability to perform inferential reasoning; and the operations it does perform must be specifically programmed in (by rational human beings). For example, whereas a human might think 'if the street is wet then it must have just rained', this is markedly different to encoded statements in computer languages along the lines of 'IF ... THEN ... ELSE' despite the beguiling syntactic similarity. The human example is a case of inferential reasoning (i.e. 'if X then that *means* Y') but the computer example is just a set of rules; that is, a set of conditional operational instructions to be executed on given input (i.e. 'if X then *do* Z').

A consequence of the claim that domain-specific computers are incapable of concept possession is that their programmatic behaviour can be used as a model for *non*-concept possessing organisms. If it can be shown that an organism's behaviour can be explained as programmatic (i.e. in terms of an inflexible set of pre-programmed rules), then that behaviour cannot be held up as an example of concept possession. Consider a very simple example: the construction of nests by paper wasps is elaborate and might even look like intelligent behaviour. But paper wasps construct nests according to a pre-programmed sequence of steps whereby each step is cued by the configuration of the nest following the previous step [Downing & Jeanne 1988]. Programmatic behaviours of this sort cannot be provided as evidence of concept possession. In particular, those behaviours that can be simulated on computers with mathematical models such as Bayesian nets [40] must be discounted as sufficient evidence of concept possession. This reasoning will be employed on occasion in part 2 during the analysis of experiments on animals[41].

[40] Bayesian nets are discussed in more detail in chapter 8.

[41] It might appear from the discussion in this section that I am challenging the Computational Theory of Mind. In fact, I do challenge *certain* forms of CTM, but I will not tackle this complex issue here due to considerations of space (I will, however, address this issue in a future project). It is sufficient for my purposes here to make the point that computer systems which we can uncontroversially accept as being non concept possessing can therefore (obviously) not serve as models of concept possession in organisms.

To emphasise: I do not mean to imply that an organism whose actions can be explained programmatically is therefore necessarily non-concept possessing. After all, humans, too, can sometimes behave in programmatic ways (e.g., through reflexes, habits, conditioning, etc.). The point is that behaviour of this sort can be accounted for by non-cognitive explanations and so on its own cannot be accepted as sufficient evidence of concept possession and therefore of self-consciousness.

**Associatively Learned Responses to Stimuli**

Associative learning refers to the process of training an association between two stimuli (operant conditioning) or between a stimuli and a response (classical conditioning). Associative learning is therefore one way to establish programmatic behaviour in an organism. It has been accepted for many decades that animals – including those we would not normally think of as concept possessing – can be trained to respond automatically to a stimulus through conditioning. Even the microscopic worm *Caenorhabditis elegans*, a 1mm nematode of only 302 neurons [White, Southgate, Thompson & Brenner 1986] is capable of associative learning [Nuttley, Atkinson-Leadbeater, & van der Kooy 2002]. Thus, when analysing animal behaviours those that can be plausibly explained in terms of associative learning should be discounted as sufficient evidence of concept possession. Several such cases will be encountered during the analyses conducted in part 2 of this thesis.

**'Innate' (Hard-Wired Species-Specific) Behaviour**

Another type of programmatic behaviour is commonly referred to as 'hard-wired' or 'innate' and is usually domain-specific. In contrast to associatively learned behaviours, this type is not learned during an individual organism's lifetime but inherited as species-specific traits through natural selection. These 'instinctive' behaviours are elicited by eventual or cyclical stimuli in the animal's natural environment and do not rely on concept possession. For example, caching food for winter cannot be taken as evidence of (say) a concept of the future self if it is part of that species' behavioural repertoire. Some of these behaviours will also emerge in artificial environments so we need to take care in laboratory settings when observing seemingly novel behaviour. For example, rats apparently 'solving' mazes need

imply nothing more than an expression of natural foraging behaviour. Thus, when analysing animal behaviours those that can be plausibly explained as expressions of hard-wired, species-specific behaviour should be discounted as sufficient evidence of concept possession. Once again, several such cases will be encountered during the analyses conducted in part 2 of this thesis.

**A Simple Test Case: Tool Use**

As a test case of the stringency I advocate for analysing animal behaviour, consider tool use, which has been suggested as evidence of rational thought (e.g., Andrews [2011]). If true then observing an animal interacting with a tool might seem like good evidence of rationality and hence (by CPH) self-consciousness. However, the picture is not so straightforward. In the first place, an animal may be genetically predisposed to use tools (for example, a beaver's dam might be considered a type of tool). So any species-specific tool use of this type must be discounted. On the other hand, an animal might learn to use a tool through imitation (i.e. by directly replicating the observed movements) without the application of rational thought (see chapter 6 for more on non-conceptual imitation). So, *prima facie*, tool use does not appear to be sufficient evidence of concept possession. Alternatively, we could rather specify tool *invention* rather than just tool *use*. Behaviour that implies tool invention might seem to require inferential reasoning of the form 'in order to solve such-and-such a problem I need to construct a device of such-and-such a shape' (of course, as I have often stressed, I do not mean to imply that the animal itself must be able to express such a thought in that propositional form in order to have that type of thought). Even so, in analysing behaviour we must be careful to distinguish tool invention from tool *discovery*, as the latter might occur in the absence of rational thought. For example, an animal may have chanced upon an object that happened to be shaped perfectly well to extract ants from a hole and so accidentally discovered its affordance for that use. Subsequent replication of the tool (i.e. by manufacture) also need not be considered evidence of concept possession: it might rather be considered a form of imitation (i.e. replicating the discovered tool by 'imitating' its shape using other material). Indeed, even tool 'invention' by a process of *trial and error* need not imply concept possession. An animal might struggle with many different attempts to reach a goal using more-or-less random actions with no rational thought behind any particular attempt, but might this way discover a successful operation, including the novel utilisation of an object (in

effect, accidentally 'inventing' a new tool). Thus, observations of an animal's interactions with a tool may have several plausible interpretations not implicative of concept possession. In summary, I advocate a critical eye in the interpretation of animal behaviour and maintaining a high standard of evidence for concept possession.

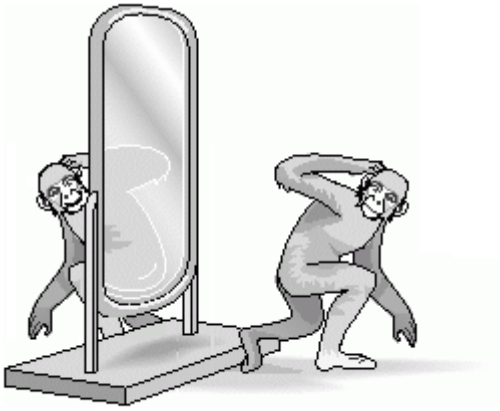
## 4.5 Conclusion

In chapter 3 I suggested an approach to the analysis of research on self-consciousness in animals and human infants based on the Concept Possession Hypothesis (CPH), which essentially claims that a demonstration of concept possession alone is sufficient evidence of self-consciousness. In this chapter I discussed several ways to detect concept possession. Properly linguistic organisms or indeed any organisms capable of propositional thought must be concept possessing (although non-linguistic organisms are not to be precluded on the basis of their lack of language). Rational creatures must possess concepts in order to engage in inferential reasoning. Organisms that are symbol-minded are concept possessing. Despite the seemingly many indications of concept possession, actually determining concept possession in animals and infants is not an easy task. With non-linguistic and pre-linguistic organisms we have only observations of behaviour to go on but we should remain sceptical and eliminate all alternative non-conceptual explanations first. To that end I suggested some simple principles to apply. To reiterate, we should discount behaviour that can be explained as (i) programmatic; (ii) occurring as a result of associative learning; or (iii) explicable as 'hard-wired' or 'innate'.

In part 2 many experimental paradigms are analysed and the above principles are applied. In the vast majority of cases the standard set above is not met, but in a few cases a conclusion can be drawn that certain animal subjects do indeed possess concepts and, by CPH, are therefore self-conscious. In chapter 5 I argue that mirror self-recognition as displayed by, for example, chimpanzees, implies symbol-mindedness. In chapter 6 I show that selective imitation is evidence of inferential thought. In chapter 7 I argue that scrub jays have a concept of their own future existence. In chapter 8, after scrutinising the results of key research paradigms, I conclude that there is insufficient evidence of rationality in rats.

# Part 2: Empirical Studies of Self-Consciousness

# Chapter 5: Mirror Self Recognition



"I knew it! It's just another chimp doing a Marx Brothers routine!

## 5.1 Introduction[42]

In a landmark 1970 paper Gordon Gallup Jr. describes a test he had devised to demonstrate that chimpanzees are able to recognise themselves in a mirror (the *mark test*). Since then Gallup has consistently maintained that mirror self-recognition (henceforth MSR) is a sufficient demonstration of self-awareness (e.g., Gallup [1970, 1975, 1997, 1998a, 1998b]; Gallup & Suarez [1986]; Gallup & Povinelli [1993]). The arguments rest on the idea that one needs a self-concept in order to recognise oneself in the mirror. However, MSR is not universally accepted as an indicator of self-awareness. Furthermore there are objections to both the validity of the mark test itself as a demonstration of MSR and to the experimental methodologies applied. In this chapter I address these objections and present a positive case for MSR as a valid demonstration of self-awareness using a new approach that focuses on the nature of the mirror image itself. The result is an argument founded on solid theoretical grounds and consequently more robust than the arguments presented by Gallup.

The first part of this chapter (sections 5.2-5.5) analyses the 'mark test' as a demonstration of MSR and reviews its application in animal and human infant studies. The main points emerging from the analysis on animal MSR are as follows. First, I conclude that the mark test

[42] A version of chapter 5 has been accepted for publication as a paper (Savanah [2012b]). The differences in the versions are inconsequential with regard to the line of argument. The published version excludes the detailed analysis of experimental data on animals and human infants, and includes a section on symbol-mindedness which here appears in chapter 4.

protocol, especially as applied to primates, meets the highest levels of experimental rigour and is a definitive demonstration of MSR. Second, chimpanzees have conclusively passed the mark test and there are good indications that so have dolphins, elephants and magpies. Since my thesis is that MSR implies self-awareness, this commits me to the view that chimpanzees are self-aware and most likely so are dolphins, elephants and magpies.

With regard to the experiments on human infants, there are three significant findings, which subsequently feature in my argument that mirror self-recognition implies self-awareness. Firstly, there are a variety of self-related cognitive skills that develop over time, with MSR being only one such skill (e.g., self-naming, self-recognition in photographs; self-recognition in videos; etc.). This will be recalled later when evaluating Gallup's arguments equating self-recognition with 'all-or-nothing' self-awareness. Secondly, the ability to match simultaneously moving mirror self-images with one's own movements greatly enhances self-recognition (compared with still images or delayed moving images). This supports the 'kinaesthetic visual matching' view of MSR explained later. Thirdly, MSR appears to develop in infants at around 18-24 months. This provides a comparison with the onset of symbol-mindedness in infants and supports the argument I present later that MSR implies symbol-mindedness in the subject.

In section 5.6 I argue from a new perspective that MSR does imply self-awareness. I first analyse the main inflationary and deflationary views on self-recognition to show that neither provides a satisfactory account. Although I agree with Gallup's ultimate conclusion that MSR implies self-awareness, his argument is inadequate in that it is not founded on theoretical grounds beyond simply assuming that self-recognition implies some level of self-awareness and therefore full-fledged self-awareness. On the other hand, accounts based on associative theories alone (those not involving a self-concept) do not stand up to scrutiny. On the contrary, I conclude that one core argument on the deflationary side (Mitchell's [1997a, 1997b] 'kinaesthetic-visual matching' account) actually favours the conclusion that MSR implies self-awareness. My argument is composed of two independent and individually significant parts. Firstly, based on the studies of 'symbol-mindedness' by DeLoache [2004] described in chapter 4, I show that MSR subjects are symbol-minded. I argue that a mirror image recognised as the self cannot be taken as an extension of the physical body or as an actual object behind the mirror and so must be taken as a *representation* of the self. Thus, it functions as a symbol. As such, species that have passed the mark test, such as chimpanzees,

must be considered ‘symbol-minded’, a capacity that develops in human infants at around 18 months of age [DeLoache 2004]. Secondly, as discussed in chapter 4, symbol-mindedness implies concept possession and hence, according to the Concept Possession Hypothesis of Self-Consciousness as explicated in chapter 3, self-consciousness.

## 5.2 The Mark Test

To determine mirror self-recognition subjects are exposed to a mirror and observed for any mirror-aided self-directed behaviour. In Gallup’s original experiment on chimpanzees, two female and two male wild-born subjects with little or no presumed prior exposure to mirrors were tested. For several days after initial exposure to mirrors the subjects reacted to their own reflections as if they were seeing conspecifics; that is, they displayed social responses such as bobbing, threatening and vocalizing. After a couple of days of habituation to mirrors, the chimpanzees were observed engaging in mirror-guided self-examination and experimentation with facial gestures.

> Such self-directed responding took the form of grooming parts of the body which would otherwise be visually inaccessible without the mirror, picking bits of food from between the teeth while watching the mirror image, visually guided manipulation of anal-genital areas by means of the mirror, picking extraneous material from the nose by inspecting the reflected image, making faces at the mirror, blowing bubbles, and manipulating food wads with the lips by watching the reflection. [Gallup 1970]

Gallup took these observations of mirror-aided self-directed behaviour as evidence of self-recognition by the chimpanzees. Nevertheless, Gallup wished to create a more objective measure of MSR and so devised the ‘mark test’.

In Gallup’s mark test protocol subjects are marked on parts of the face visible only in a mirror. Subjects who see these marks on their *reflection* in a mirror and react as though the marks are on their *own* face (i.e. by touching them) are deemed to have passed the mark test and shown evidence of self-recognition. The number of touches within a set period of time provides a quantifiable measure. Gallup marked his chimps on the uppermost portion of an eyebrow ridge and the top half of the opposite ear. Gallup applied rigorous methodology to avoid any possibility of his subjects knowing about the marks before seeing them in the mirror. His subjects were marked under anaesthesia with an odourless, non-irritating dye and

so had no information about the marks due to olfactory or tactile cues. Furthermore, after the subjects recovered from the anaesthesia, they were observed for a suitable period of time before further mirror exposure to ensure no behaviour indicating subject knowledge of the mark application [Gallup & Suarez 1986].

Control subjects were also tested in Gallup's original experiments. Another two wild-born chimpanzees of the same approximate age as the marked chimpanzees were selected (though in his 1970 paper Gallup makes no mention of the actual age of the chimpanzees other than to describe them as preadolescent). Unlike the test subjects, the controls were not first habituated to mirrors before applying the mark test. They were anaesthetised and marked in the same way and confronted with a mirror. No mark-directed responses were observed. Gallup concluded from this result that the other chimpanzees had learned the capacity for self-recognition [Gallup 1970].

**Objections to the Mark Test**

Heyes [1994] claimed that the hypothesis of MSR as an explanation of chimpanzee mark-touching is no more plausible than the hypothesis that mirror introduction elevates arousal and thereby produces an increase in the frequency of a range of behaviour patterns. But this ignores the evidence that significantly increased mark touching occurs well after chimpanzee *habituation* to mirrors [Gallup 1970], not after newly introduced mirrors. Heyes also objects to claims that the mark test shows chimpanzees have the capacity for MSR whereas monkeys do not on the grounds that apes spontaneously touch their faces more often than do monkeys. But the key factor is the *differential* in mark touching between mirror-absent and mirror-present conditions *within the same species*: in chimpanzees the mark touching increases, in monkeys it does not.

Some researchers failed to replicate Gallup's findings with chimpanzees [Swartz and Evans 1994]. They claim that it was theoretically interesting and important that in their study, only one out of eleven chimpanzees passed the mark test. Gallup [1994] countered that the chimpanzees involved in this study had an atypical medical history – they were all maintained at a medical research facility. Furthermore, many of the subjects failed to show

much interest in mirrors from the start, thereby failing a pre-requisite for the mark test in Gallup's view.

Since Gallup's original experiment hundreds of chimpanzees have been tested for MSR, more than all other animals tested put together [de Veer & van den Bos 1999, p464]. Not only are the results predominantly positive, but there is further evidence that MSR is an acquired cognitive skill and not simply learned associations. Morin & DeBlois cite an experiment by Thompson & Calhoun that demonstrates chimpanzees are able to retain MSR capability even after one year without any intervening mirror experience [Morin & DeBlois 1989, p290]. Given the abundance of positive results it is now generally accepted amongst MSR commentators that chimpanzees have successfully demonstrated the capacity for MSR (for example, see Anderson [1984]; Morin & DeBlois [1989]; de Veer & van den Bos [1999]; Povinelli et al. [1997]; Schilhab [2004]; for dissenting opinions see Heyes [1998]; Mitchell [1993, 1994, 1997a, 1997b]).

**Summary**

Gallup's methodology as applied to primates appears to set a very high standard for experimental rigour. Experimental conditions were carefully designed to minimise the possibility of alternative interpretations of behaviour and included the use of control subjects for comparison. Furthermore, Gallup's mark test protocol allows an objective and quantifiable measure for mirror self-recognition. It is now widely accepted that chimpanzees have demonstrated the capacity for MSR.

## 5.3 Mark Test Studies of Primates

**Great Apes**

Given that self-awareness is high on the scale of cognitive capacities there is an obvious temptation to link this ability with the most 'advanced' primate species, to wit, humans and other great apes. If MSR is indeed sufficient evidence for self-awareness, we would expect to

see a correlation between the capacity for MSR and phylogenetic development. In fact there is tantalising evidence that this may be the case. Figure 5.1 is a condensed taxonomy of primates indicating tested species that have putatively passed the MSR test, those that have failed and those for which results are uncertain. As can be seen, all the great apes have passed with the possible exception of gorillas. As discussed below there is insufficient evidence to conclude that old world monkeys and new world monkeys have passed.

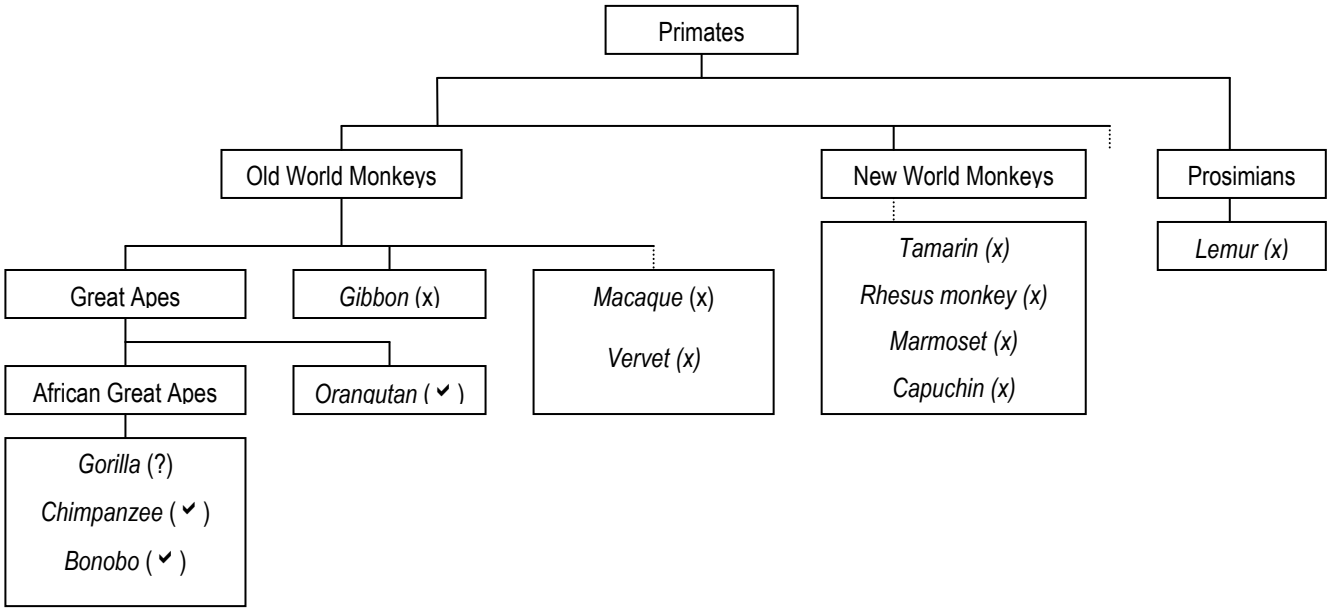



Figure 5.1: Condensed Primate Taxonomy (based on Schilhab [2004])

Gallup [1970] was the first to suggest that self-recognition (and by implication self-awareness) might not extend below man and the great apes. Such a neat dividing line would be consistent with the view that the development of self-awareness is an evolutionary process. However, there is a question mark over gorillas, which could potentially overthrow this idea.

**Gorillas**

No wild-born gorillas have passed the mark test [Shillito, Gallup & Beck 1999], though two home-reared gorillas have passed [Patterson & Cohn, 1994; Swartz & Evans 1994]. Several reasons have been proposed for this apparent anomaly, founded on the supposition that

gorillas are self-aware despite their inability to pass the test. Shillito, Gallup & Beck [1999] suggest that as gorillas show pronounced gaze aversion they might be avoiding eye contact with their mirror reflections and thus never receive enough exposure to their reflections to recognise themselves. Patterson and Cohn [1994, p286] put forward as a 'likely explanation' that gorillas' behaviour is inhibited by the presence of unfamiliar experimenters. According to Shillito et al. [1999], Ristau [1983] proposed that the reason for gorillas' failure to pass the mark test might be due to a lack of interest in foreign marks on their bodies[43]. Shillito et al. [1999] conducted mark tests on two adult lowland gorillas at the Washington D.C Zoo to specifically test each of these hypotheses. In the first they used angled mirrors such that the gorillas could see their own reflections without making eye contact (the experiment was then repeated using regular flat mirrors). The results were negative: neither gorilla touched the facial marks during the mirror-absent period or the mirror-present period, indicating that gaze aversion probably was not the explanation for lack of MSR. Next they performed the mark test using video cameras to observe the gorillas with no 'unfamiliar experimenters' present, again with a negative result. Finally they marked the gorillas' wrists instead of their faces and observed that they did indeed extensively attend to the mark, ruling out a possible lack of interest in foreign marks on their bodies.

There are other explanations offered for why gorillas fail to pass the MSR test, such as that gorillas show little interest in mirrors [Swartz & Evans 1994, p203] or that they lack motivation [Patterson & Cohn 1994:286]. Still others [Gallup 1994; Povinelli 1994; Povinelli & Cant 1995] propose that gorillas might fail the MSR test as they do in fact lack the capacity for self-recognition; that this capacity, inherited from the common ancestor of the great apes, has been lost in the gorilla lineage. The reason that wild-born gorillas fail the mark test is yet to be explained, but two gorillas bred in captivity, Koko and King, have apparently passed the mark test. This suggests that the reason wild-born gorillas have not passed might indeed be due to environmental or behavioural factors as yet undetermined.

Koko is a language-trained gorilla reared in an environment similar to that of a human child, replete with mirrors [Patterson & Cohn 1994, p273]. She was given a variation of the mark

[43] To be precise, Ristau was less specific on this point, suggesting only that the deficit involved 'motivational factors'. Ristau also suggested the possibility of motor coordination difficulties with mirror-image reversal (p504), although Shillito et al. did not test for this.

test[44] at the age of 19 and observed by two independent observers. On four days during exposure to the test mirror while unmarked, Koko touched the target area on average once per session. On the fifth day after being marked for the first time she touched the target area 47 times and in addition her viewing time increased to 88% versus an average of 48% during the previous four days. King underwent the mark test at the age of 22 and was given mirror exposure for ten minutes per day for 2 months prior to the mark test [Swartz & Evans 1994, pp201-202]. On the day of the test King touched the target area only twice within the ten minutes exposure to the mirror, but after the first touch he apparently smelled and tasted his fingers. He also demonstrated other mirror-guided mark-directed behaviour: he played 'peek-a-boo' with the mark by hiding it behind cage bars for a few moments and he also rubbed the marked brow against the cage bars.

Thus, perhaps captive-bred gorillas such as Koko and King can pass the mark test because they are unencumbered with the species-specific impediments of their wild-born cousins. If this is the case then perhaps the gorilla as a species does indeed have the capacity for MSR but (except for Koko and King) have not yet been able to demonstrate it in experiments. As such, the theory that self-recognition is correlated with phylogenetic development remains viable. On the other hand, this theory would be somewhat threatened if (less developed) monkey species were to show the capacity for self-recognition. Claims of this sort are discussed next.

**Monkeys**

Hauser, Kralik, Botto-Mahan, Garrett & Oser [1995] devised a variation of the mark test intended to demonstrate that it is probably due to methodological reasons why monkeys (and possibly other species) have not previously passed the mark test[45]. The distinctive white hair of cotton-top tamarins was dyed during anaesthesia. A change to such a salient and species-unique feature was thought more likely to be noticed and cared about by the subjects, especially if the feature is associated with specific behaviour such as threat displays. Based

[44] The only significant difference in the experimental set up was that Koko was not anaesthetised for the marking step. Instead, she was 'sham-marked' with a swipe across the brow using a damp cloth for each of several days before the mark test. On the day of the test she was marked with an identical cloth that had been dipped in clown paint. A similar procedure was used on King [Swartz & Evans 1994, p201]

[45] Hauser et al. also performed the 'classic' mark test on the tamarins by dying the right eyebrow and left ear (which they termed 'Gallup marks'). The tamarins did not pass this version of the mark test.

on their observations, Hauser et al. concluded that their tamarins passed the mark test and note that these data contradict the 'inferred phylogenetic gap.' These results, however, subsequently failed to replicate [Hauser, Miller, Liu & Gupta 2001] and must be considered questionable. A separate claim has been made that rhesus monkeys possess MSR despite failing the mark test [Rajala, Reininger, Lancaster & Populin 2010]. The claim is based solely on observed apparently mirror-guided behaviour; however a critical examination of the results reveals them to be at best inconclusive [Anderson & Gallup 2011]. In summary, there is as yet insufficient evidence of MSR in monkeys.

## 5.4 MSR Investigations on Non-Primates

If passing the mark test provides sufficient evidence of MSR and hence self-awareness (as I argue later) then I will be committed to attributing this cognitive capacity to all organisms that pass the test. Thus it is worth briefly examining the results of MSR studies on other animals. MSR experiments on non-primates necessarily vary significantly from the protocol established by Gallup, as the subjects have very different morphologies (most pertinently, they do not have hands with which to touch the marks). As such it may be questionable as to whether any non-primate subject can really be said to have passed the mark test – that is, the *classic* mark test. There are variations of the mark test that researchers claim to demonstrate MSR despite deviating from Gallup's protocol. But these tests, if not directly comparable to those used on primates, make it all the more difficult to interpret exactly what the results actually show. Nevertheless, some cleverly designed experiments provide quite convincing evidence of MSR in some non-primates. The non-primate species tested include the pigeon, parrot, magpie, elephant and dolphin. The non-primates said to have passed a version of the mark test are the bottle-nosed dolphin, elephant and magpie.

### Dolphins

Early MSR experiments on dolphins were inconclusive (Marten & Psarakos [1994]; Marino et al. [1994]). Apart from the inability to observe mark touching, there are several other salient differences between the early dolphin mark test and the classic mark test. The marking fluids were tactile and the dolphins were not anaesthetised during the marking procedure

(also most dolphins were not even sham-marked), so the subjects may have been aware of having their bodies marked. Also the dolphins were not tested in isolation so any effects due to the presence of conspecifics were not controlled for, such as that their behaviours may have been influenced by communication with their pool mates. These departures from the classic mark test as used on primates made the dolphin results incomparable to primate results and highlighted the need to design a species-specific conceptual replication of the mark test for the dolphin.

In later experiments, Reiss & Marino [2001] first made three predictions regarding behaviour indicative of self-recognition and then tested for each of these, taking the position that all three must be confirmed for a convincing demonstration of self-recognition. These predictions were as follows: (1) the dolphin should display no social behaviour at the mirror (non-self-recognising species treat their reflections as conspecifics; self-recognising species should therefore not do so); (2) the dolphin should spend more time at the mirror when marked than under any other conditions; (3) there should be a shorter latency in the interval between being marked and a mirror visit than in sessions when non-marked. Statistical analysis of the test results confirmed each prediction, leading Reiss & Marino to claim they had convincing evidence of MSR in the bottlenose dolphin.

**Pigeons**

The pigeon data are important because it has been used to argue against the conclusion that the mark test indicates MSR. Epstein, Lanza and Skinner [1981] conducted an experiment over a ten-day period whereby three previously mirror-naïve adult male White Carneaux pigeons were *trained* to pass the mark test. Using food as reinforcement, the pigeons were first trained to peck at various dots on the body and around the cage. A mirror was used in the case of dots around the cage, in which the dots were flashed when they could only be visible in the mirror – the pigeons were rewarded for pecking at the positions where the dot had been. Then a test was conducted whereby a dot was placed on the pigeon's breast and a bib was used to render the dot invisible to the pigeon except in a mirror. If the pigeon bent its head forward even slightly the bib slid down the breast and covered the dot. The pigeon did not peck at the dot unless a mirror was present to allow the pigeon to see it, in which case the

pigeon pecked at the position on the bib that corresponded to the position of the covered dot. This result was taken by the experimenters as a passing of the mark test.

This pigeon version of the mark test differs from the classic mark test even more than does the dolphin test. The pigeons were neither anaesthetised nor sham-marked, and the marks were actually 1cm wide blue stick-on dots. There is no element of training or reinforcement in the classic mark test. Moreover there was no mention by Epstein et al. of any observations of previous self-directed behaviour by the pigeons in front of the mirror, or whether the pigeons ever expressed social responses toward their mirror images, as was the case for chimpanzees. Given these differences, the results for pigeons are incomparable with those for primates and it cannot be justifiably claimed that pigeons are capable of MSR.

In any case, the researchers did not conclude from their results that pigeons are self-aware; rather, they concluded that there must be a non-mentalistic explanation for the ability to pass the mark test. They therefore suggested that such a non-mentalistic explanation might also account for primates' ability to pass the mark test. This issue is further discussed in section 5.6.

**Parrots**

Pepperberg, Garcia, Jackson & Marconi [1995] performed tests on the mirror use of two African Grey Parrots. They specifically tested for (i) mirror image stimulation (reactions of the subjects to their own reflections), (ii) mirror-mediated object discrimination (whether the birds would react appropriately to positive/aversive stimuli visible only in a mirror) and (iii) mirror-mediated spatial locating. In test (i) no clearly self-exploratory behaviour was observed; the parrots always treated their own image as conspecifics. The parrots were under two years old and had had limited mirror exposure. Given that parrots have a lifespan similar to humans and that even in children self-recognition does not appear until about 18 months, it may have been optimistic to expect these subjects to demonstrate self-recognition. Still, Pepperberg et al. went on to study the parrots for other forms of mirror use that would compare their cognitive abilities to those of mammals. Test (ii) was used as a control for test (iii) by first proving that the parrots did not simply use mirrors as a cue to begin an object search. In test (iii) the parrots were found to be able to locate hidden objects using a mirror.

Although parrots cannot be said to have passed the mark test, Pepperberg et al. concluded that parrots are able to process mirror information: they can apparently differentiate reflective versus non-reflective information and use a representation to locate hidden objects. However, I argue against the conclusion that these subjects used (symbolic) representations. If they had used representations for hidden objects in this way, I would be bound (by CPH) to ascribe self-awareness to the subject, as I explained in chapter 4. However, as I argue later, being able to use mirrors to locate hidden objects is insufficient evidence of representational abilities. The reasoning for this is spelled out in section 5.6.

**Magpies**

Mark tests on magpies (*pica pica,* members of the corvid family) were conducted by Prior et al. [2008]. Prior to the mark tests the magpies were habituated to mirrors. Subjects initially reacted to the mirror images as conspecifics; they exhibited social behaviours such as aggressive and submissive displays. Two types of controls were applied. In the first, the test birds were marked with a brightly coloured (yellow or red) mark while the controls were marked with or a black (sham) mark. The black mark was not visible against the black feather background. In the second, the control half of the trials replaced the mirror with a non-reflective plate of the same size and position. Each bird was tested twice in each of the conditions. Passing the mark test was determined by a statistically significant increase in mark-directed behaviour.

Two out of five magpies tested convincingly passed the mark test. Prior et al. note that corvids belong to a phylogenetic group characterised by large brain sizes relative to body weight and that among this group corvids have particularly large relative brain sizes, with the obvious implication that brain size correlates with self-awareness.

**Elephants**

Elephants display sophisticated mirror-directed behaviour such as monitoring their reaching efforts in the mirror and demonstrated an understanding of the left-right reversal property of mirrors, yet failed to show self-recognition in early experiments [Povinelli 1989]. However,

later researchers, suspecting this failure was due to the relatively small sized mirrors used, successfully performed mark tests on Asian elephants using elephant-sized mirrors [Plotnik, de Waal & Reiss 2006]. Plotnik et al., noting that humans, apes, dolphins and elephants are all known for empathy and altruistic behaviour, speculate that there is a correlation between these attributes and the capacity for MSR. They suggest a convergent cognitive evolution related to complex sociality and cooperation. This theory does not account, however, for the apparent presence of MSR in magpies.

**Summary**

Of the primates only and all the great apes have demonstrated MSR, if home-reared gorillas are included. The fact that wild-born gorillas have not done so is as yet unexplained but may be due to species-specific impediments unrelated to their cognitive capacity. Among non-primates there is as yet insufficient data to be absolutely decisive but there are promising results for dolphins, elephants and magpies. Gallup, Anderson & Platek [2011], though particularly impressed with the magpie research, caution against accepting any of the other non-primate results too readily, as in the dolphin and elephant cases only one individual convincingly passed and the results are yet to be replicated. Nevertheless, the results so far are promising and with further positive evidence we should be ready to accept MSR in these species. As I argue later, MSR is a demonstration of self-awareness and so, given further confirmation, I am ready to attribute self-awareness to the aforementioned species.

## 5.5 MSR Investigations on Children

In human infants one knows from the outset that self-awareness *will* emerge eventually, so the question focuses not on ‘if’ but rather on ‘when’ and ‘how’. Naturally it is tempting to draw parallels between the ontogenetic and phylogenetic development of self-recognition. There are three significant findings relating to MSR in infants that will feature in my arguments claiming that MSR implies self-awareness and that therefore animals that show MSR are self-aware. These findings emerge during the review on infant MSR experiments presented below. In brief they are, firstly, the age at which MSR arises in infants, which provides a comparison with the onset of symbol-mindedness as discussed in the previous

chapter and supports the view that there is a link between symbol-mindedness and MSR. Next, the fact that MSR is but one of a suite of self-related cognitive capacities that generally develops in a given order tends to refute the stance taken by Gallup. Lastly, the finding that bodily motion enhances recognition of the self lends support to the 'kinaesthetic visual matching' view, which although was originally raised to deflate MSR ends up supporting the view that MSR implies self-awareness.

**Mark Test Methodologies for Children**

In Amsterdam's [1972] version of the mark test (sometimes referred to as the *rouge test*), the child's mother marked the side of the child's nose with rouge according to standard instructions from an experimenter and then exposed the child to a mirror. As with Gallup's chimpanzees, self-recognition was assumed by Amsterdam if the child touched the mark on his nose or used the mirror to examine his nose. Other subsequent investigators added their own twists to the methodology; Schulman & Kaplowitz [1977] used a distorting mirror as well as a normal (flat) one; Lewis & Brooks-Gunn [1979] and Bigelow [1981] in addition to mirrors used videotape recordings of their subjects to reflect their images back to them; Bertenthal & Fischer [1978] added hats and toys as devices in their experiments.

The controlled conditions Gallup used for his studies on chimpanzees were absent in the rouge test: the child was conscious at the time of the marking and the substance used was not chosen for its lack of odour or tactility. Furthermore, the area chosen for the mark was more readily accessible to the child's direct vision than in Gallup's methodology. This led to criticisms from Gallup [1975] that the rouge test results were questionable: mark-directed responses may not have indicated self-recognition as the child may have been responding to tactual and olfactory cues associated with the marking procedure or indeed to direct visual access. To overcome these objections Bigelow eschewed the mark test methodology altogether and instead employed a visual stimulus that was located behind the subject. In Bigelow's method a clown face suspended from above was silently lowered behind the subject and could therefore only be seen by the subject in a mirror behind his own self-image. If upon seeing the clown face in the mirror the subject turned to look behind himself the subject was deemed to have demonstrated MSR, as he would have realised from what he saw in the mirror that it was his own image that the clown face was behind. Bertenthal & Fischer

used a similar method to Bigelow's as well as the rouge test, claiming that these represented different stages of self-recognition. All the investigators mentioned herein also tested for verbal self-naming but whereas some (e.g., Amsterdam; Bigelow) used it as a conjunctive test for self-recognition (i.e. indistinct from MSR), others (e.g., Schulman & Kaplowitz; Bertenthal & Fischer) considered this a distinct cognitive capacity. Despite the variations in approach from different investigators, a broad consensus emerges from these experiments, as shown below.

**Developmental Stages toward Self-Recognition**

As might be expected the presumed stages or phases in the development of self-recognition tend to be age-related. Amsterdam saw three distinct phases in the child's reaction to his mirror image in her study of 88 subjects. In the 6-12 month age group 85% of the children were socially directed ('playmate in the mirror'). In the 13-24 month age group 90% of the children withdrew from the mirror. Other behaviour in this age group included searching for the image, with some subjects displaying signs of embarrassment and self-admiration: 75% of the subjects engaged in this behaviour after 20 months. Self-recognition was observed in the 20-24 month range. The transition is graphically depicted in Table 5.1, which is a schematised presentation of Amsterdam's data. Amsterdam used a 'mirror behaviour checklist' containing 11 categories broken down into a total of 34 distinct actions (e.g., 'reaching into mirror' and 'searches behind mirror' were categorised as 'Searching Behaviour'). These were compressed into the six most frequent behaviours as displayed in Table 5.1. In this table the percentage figures have been replaced by grey-scale shading rounded to the nearest 5% to highlight the progression through various stages toward self-recognition: the darker the shade, the higher the percentage. There is a clear transition through various phases of behaviour, starting with sociable behaviour in the younger ages through to self-directed behaviours in the older ages.

Table 5.1: Percentage of children in each age group displaying indicated behaviours

| Age Group | Sample Size | Sociable Playmate | Observes Movement | Search for image | Withdraws | Self-Admire/ Embarrassed | Recognition |
|---|---|---|---|---|---|---|---|
| 3-5 mths | 12 | | | | | | |
| 6-8 mths | 12 | | | | | | |
| 9-11 mths | 12 | | | | | | |
| 12-14 mths | 12 | | | | | | |
| 15-17 mths | 12 | | | | | | |
| 18-20 mths | 12 | | | | | | |
| 21-24 mths | 16 | | | | | | |

Adapted from Amsterdam (1972). Grey-scales rounded to the nearest 5%

Schulman & Kaplowitz's study of 72 children was similar to Amsterdam's, except that they also observed children's reactions to blurred and distorted mirrors. They separated their subjects into four 6-month age groups; 1-6 months; 7-12 months, 13-18 months and 19-24 months. They used a 'mirror behaviour checklist' much the same as Amsterdam and their results are broadly in agreement. For the sake of simplicity the data on blurred and distorted mirrors has been excluded from my presentation of their results in Table 5.2. Also, for clarity's sake, I have excluded their data on 'Observes Image' and 'No Interest' as virtually all subjects studied displayed these behaviours at some point. Once again there is a clear progression toward self-recognising behaviour.

Table 5.2: Percentage of children in each age group displaying indicated behaviours

| Age Group | Sample Size | Social Behaviour | Avoidance | Admires Image | Observes Nose | Names Self |
|---|---|---|---|---|---|---|
| 1-6 mths | 12 | | | | | |
| 7-12 mths | 20 | | | | | |
| 13-18 mths | 19 | | | | | |
| 19-24 mths | 21 | | | | | |

Adapted from Schulman and Kaplowitz [1976]. Grey-scales rounded to the nearest 5%

Whereas Amsterdam lumped together mark-directed behaviour and self-naming into the same category of 'Recognition', Schulman & Kaplowitz separated these out. In Schulman & Kaplowitz's study no child under the age of 19 months named himself on seeing his image in the mirror. Both studies agreed that the 2nd year of life might be a transitional stage for children who are beginning to experience self-recognition.

**Matching Self-Movements and Reflected Movements**

Bigelow [1981] conducted a longitudinal study over 8-10 months of 11 children who were aged 18 months at the beginning of the study. The purpose of Bigelow's study was to test the hypothesis that early self-recognition is achieved through sensorimotor means via the exact matching of self-movements and reflected movements. Thus two predictions were made: (i) children recognise themselves in self-images that move simultaneously with their own movements *before* they recognise themselves in self-images without this simultaneity and (ii) prior to self-recognition children engage in movement *testing* while attending to their self-images. In order to test these predictions Bigelow had to find several different ways to reflect the child's self-image back to him, some of which moved simultaneously and some that did not. The methods chosen were:

*Simultaneously moving self-images*

a) 'Mirror Condition' (exposing the child to his own reflection in a mirror)
b) 'Simultaneous Condition' (immediate videotape feedback)

*Non-simultaneously moving self-images*

c) 'Discordant Condition' (showing the child a videotape of himself made earlier) and
d) 'Photograph Condition' (showing the child photographs of himself and other children and asking him to point to the picture of himself).

By far, the greatest amount of time spent by the subjects on movement testing (waving, bouncing, handshaking, etc.) was for the Simultaneous Condition. This was followed by the Mirror Condition and then to a significantly smaller extent by the Discordant and Other Child conditions. The self-recognition results are graphically represented in figure 5.2, showing another view of the progressive stages involved. None of the individual children recognised themselves in the Discordant or Photograph conditions prior to self-recognition in the Mirror and Simultaneous conditions, and all the children recognised themselves in the Mirror Condition before they recognised themselves in the Simultaneous Condition. These results therefore support the hypothesis that early self-recognition is achieved through sensorimotor means via the exact matching of self-movements and reflected movements.

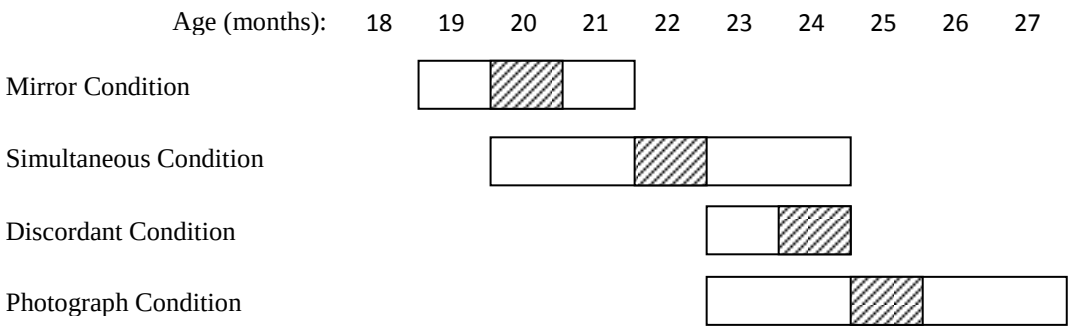



Figure 5.2: Age range for self-recognition in each test condition (shaded = mean age)

(Adapted from Bigelow [1981])

**Objections to the MSR studies in Infants**

Schilhab [2004, p117] casts some doubt over the validity of the mark test as applied to human infants. She referred to results obtained by Lewis & Brooks-Gunn [1979, p39] in which some infants, on seeing their mothers with a rouge marked nose, would touch their own noses even without the benefit of a mirror being present. This makes the mark test seem inconclusive: the child might attend to the mark on his nose even if interpreting the mirror image as another child. The objection does not hold against MSR tests that used an alternative behaviour to indicate MSR [Bigelow 1981]. Bigelow used the previously described 'clown face' technique to determine MSR as it is impervious to the aforementioned objection. Her results broadly agreed with others' mark test results as to the age range for the onset of MSR in infants. Interestingly, though, Bertenthal & Fischer [1978] applied 'Hat' and 'Toy' tests that were essentially equivalent to the clown face test to 48 infants aged 6-24 months. They also applied the mark test. In their sample, passing the Hat/Toy tests always occurred *before* passing the mark test rather than at the *same* developmental stage as assumed by Bigelow. Unfortunately, their reported results were not indexed by age so it is difficult to assess the impact of this discrepancy with Bigelow's results.

Schilhab also raised another objection to the validity of the mark test on human infants. Citing Bigelow, she asserted that children who recognised their own mirror images as measured by the mark test did not respond adequately verbally; they named all child images self-images. However, this is not quite an accurate interpretation of Bigelow's results. In this longitudinal study, it was in the period *prior to MSR* when some children identified all child

images as self-images [Bigelow, 1981, p24]. There were also cases of subjects verbally identifying other children as self after passing the MSR test but before passing the Discordant Condition (in which a video of the subjects' image is played back from an earlier recording). Bigelow puts this down to the instability of the subjects' recognition of self-images when the cues of correspondence between self and image *movement* were absent, further strengthening the supposed connection between self-recognition and self-movement matching.

**Summary**

The methodology of the mark test as applied to human infants necessarily lacks some of the rigour imposed by Gallup for primate investigations; it would not have been ethical, for example, to anaesthetise human subjects. Nevertheless, despite the variations of the methodology applied, the results are broadly in alignment. The results presented above indicate that there is a variety of self-related cognitive skills that develop over time, with MSR being only one such skill. Taking all results together indicates that MSR develops in infants at around 18-24 months. The results from Bigelow's experiments demonstrate that self-recognition skills are enhanced by the ability to match simultaneously moving mirror self-images with one's own movements. These results feature prominently in the next section in which I present a new argument that MSR does indeed imply self-consciousness as defined in part 1 of this thesis.

## 5.6 MSR and Self-Awareness

In this section I discuss the ramifications of MSR for self-awareness. The extreme views here can be called the 'Gallup approach' and (after Schilhab [2004]) the 'non-mentalistic interpretation'[46]. The inflationary Gallup approach is to simply equate MSR with full-fledged self-awareness including relatively sophisticated cognitive abilities such as theory of mind and personal identity. At the other extreme is the deflationary non-mentalistic interpretation (e.g., Heyes [1998]), in which associative theories are used to explain MSR without the need

[46] Schilhab [2004] uses this expression to describe an organism's ability to distinguish its own body sensations from externally sourced sensory input, which does not relate to a mental category. In other words, this ability, common throughout the animal kingdom, need not imply the capacity for conceptual thought, but potentially could explain MSR (as discussed later in the text).

to invoke the existence of a self-concept. I do not review the full range of positions but I briefly analyse these extremes as an introduction to a third approach that I name the 'Indirect Interpretation'. In this approach I evaluate MSR based on the role of the mirror image itself. Although I conclude that MSR does provide sufficient evidence for self-consciousness, I find the Gallup argument inadequate. The Indirect Interpretation, explicated below, provides the missing theoretical grounding for a more robust argument.

**The Gallup Approach**

The approach taken by Gallup obviates the need for reasoned argument. It relies on the view that one either does or does not possess self-awareness. Gallup insists, unjustifiably in my view, that self-awareness is 'all-or-nothing' in the sense that one cannot be partially self-aware. As Mitchell [1997a] says of Gallup: "…he identifies the self-awareness necessary for mirror-self-recognition with any and all other forms of self-awareness" (p23). These other forms include Theory of Mind (ToM) [Gallup & Suarez 1986; Gallup 1998a] and even personal identity: "…the ability to infer correctly the identity of the reflection requires an identity on the part of the organism making that inference" [Gallup & Suarez 1986, p4]. By taking this position, Gallup then needs only to show *any* indication of some kind of self-awareness to close his case. Thus, solely on the basis that self-recognition must imply some level of awareness of the self, Gallup concludes the subject must therefore have an "integrated concept of self" [Gallup 1975, p331].

The Gallup approach is implausible in that it does not allow for levels of self-awareness or for a progressive development of cognitive capacities of which MSR is only one stage. The results presented earlier on infant MSR studies argues against this position; there are clearly developmental stages of cognition involved in self-recognition. Furthermore, some of the other examples of self-awareness used by Gallup appear to develop at much later stages than self-recognition. For example, whereas a reasonable argument may be made that a demonstration of ToM implies the existence of self-awareness (on the grounds that ToM relies on modelling others based on knowledge of the self[47]), the reverse is not necessarily so.

[47] This point holds irrespective of which of the most popular accounts of ToM is accepted. The three leading variants ('rationality theory', 'simulation theory' and 'theory theory' [Goldman 2005]) are discussed in more detail in chapter 6. The main conclusion for my purpose here is that in each case the organism must somehow infer the mental states of conspecifics by reference to its own.

It is conceivable that an organism has developed self-awareness without achieving ToM. Furthermore, in at least some experiments used to gauge ToM abilities (such as 'false belief') this cognitive capacity appears to develop in children at a generally later age (3-4 years) [Wellman, Cross & Watson 2001] than does MSR (18-24 months). Also by this age infants have not yet developed other significant cognitive capacities such as episodic memory, which some see as necessary for a sense of personal identity (as discussed in more detail in chapter 7).

Given the lack of both substantial argument and supporting evidence, I cannot endorse the Gallup Approach. MSR itself does not indicate the full range of cognitive capacities associated with self-consciousness and indeed there is no reason to simply assume that self-consciousness is all-or-nothing. I do believe, however, that MSR provides sufficient evidence for a level of cognitive capacities that we can comfortably characterise as self-consciousness, as I explain later.

**The Non-Mentalistic Interpretation**

If we are to attribute any level of self-awareness to non-human organisms on the basis of their passing any type of test we must take care to rule out the possibility of non-mentalistic explanations of the observed behaviour. Heyes [1998] claims that this condition has not been met with respect to MSR studies on non-human primates. For example, Heyes cites the pigeon studies of Epstein et al. [1981] to argue for the non-mentalistic interpretation: the finding that pigeons can apparently be *trained* to pass the mark test. Chimpanzees' passing of the mark test might be essentially no different to the trained pigeons'. Because in the case of chimpanzees grooming plays a crucial role, self-grooming might count as positive reinforcement, and hence substitute for the food reward used to train the pigeons.

However there are several problems with this argument. Firstly, even if we were to accept that self-grooming (i.e., mark-touching) could act as the reward in a reinforcement paradigm, this was not the only behaviour chimpanzees displayed in front of the mirror; they also engaged in facial gesturing and examination of body parts visible only in the mirror. Although Schilhab [2004] considers it 'likely' that self-grooming parts of the body induced by mirrors might reinforce these other mirror-guided behaviours, I find this speculation

unconvincing. Grooming is a social behaviour so self-grooming is unlikely to have the same reward value. Secondly, the whole point of Gallup's mark test was to provide an objective method to quantify MSR in organisms given that these organisms had already been observed to display mirror-guided self-directed behaviour. According to Gallup & Suarez [1986] as the pigeons did not engage in any such behaviour, the mark test in this case is uninterpretable. Others also show similar scepticism toward the pigeon data, especially as the results have not been replicated [Thompson & Contie 1994]. These objections taken together give me cause to reject the pigeon data as evidence to support the non-mentalistic interpretation of the primate data.

Mitchell [1997a, 1997b] claims that MSR could be better explained by kinaesthetic-visual matching rather than the ascription of a self-concept on the part of chimpanzees. In kinaesthetic-visual matching the organism facing the mirror is mapping his own proprioceptive sensations onto the mirror image of his own body. In this way, the organism demonstrates what Heyes loosely describes as a 'body concept'[48], but this is not what we refer to when using terms like 'self-concept' and 'self-awareness'. As Heyes puts it:

> …a 'body concept' does not relate to a mental category, and since it is equally necessary for mirror-guided body inspection and for collision-free locomotion, the former no more implies possession of such a concept than does the latter. [Heyes 1998]

Thus, Heyes and Mitchell both give a type of behaviouristic account of MSR, whereby the organism correlates sensory input from the operations of its body with sensory inputs from elsewhere (visual data from the mirror image). The general idea about the organism's mirror experience is that it learns to associate its kinaesthetic sensations with the simultaneous movement of the mirror image and from this associates the mirror image with its own body. The data on human infants bears this out; in Bigelow's [1981] experiments self-recognition occurred earlier in conditions where the subjects were able to match their own movements with that of the mirror image.

What is not made clear in the kinaesthetic-visual matching argument is whether the subject takes the mirror image as its *actual* (physically extended) body, or as a *representation* of the body. It seems unlikely that the proponents of this line of argument are proposing that the

[48] I think the term 'body concept' (albeit given in scare quotes) is an unfortunate choice by Heyes as no concept is involved. What she refers to is better described as 'body schema' [Gallagher 1995].

subject is taking its mirror image as a physical extension of its own actual body, given that the image is distal and unconnected. However, in my view it is *only* if the mirror image were taken as a physical extension of the actual body that the argument could support a non-mentalistic interpretation, as it is the proprioception of the *actual* body that corresponds to a body schema, not the image. I argue that kinaesthetic-visual matching implies the subject cannot be taking its mirror image as a physical extension of its own body and, consequently, that the subject sees its image as a *representation*.

The kinaesthetic-visual matching argument seems to depend largely on the assumed lack of visual accessibility to one's own body. For instance, consider this line from Gopnik & Meltzoff [1994]: "…our bodies, after all, are only peripherally part of our field of vision". Perhaps so, for humans accustomed to mirrors in everyday life, but even for humans a quick glance downward will confirm that we can actually see most of our bodies without mirrors; only the head and back are visually inaccessible. For our more flexible primate cousins an even greater portion of the body is directly visually accessible. Given that a primate will see much of its own body directly as well as simultaneously reflected in the mirror, we should assume that it not only does kinaesthetic-visual matching but also 'visual-visual' matching. The movements of (say) an arm and the arm's identical mirror image will be perfectly synchronised and both simultaneously visible. Any visual matching of kinaesthesia going on is more likely to correspond to the sight of the primate's actual arm moving in the (non-mirror) visual field rather than its mirror image, as it is the actual arm's position in space that is accompanied by proprioception, not the mirror image. That the arm and the arm's mirror image are both simultaneously visible might then lead to an alternative interpretation to the one defended by Mitchell and Heyes. Certainly I agree that the kinaesthetic-visual matching argument implies the organism recognises the mirror image as *its own body*, but the point I want to emphasise is that it does not recognise the image as part of its *physical* body. The organism learns to associate his kinaesthetic sensations with *both* the sight of his actual body and its identical visual image in the mirror. However, as only his *actual* body parts matches the sensations of proprioception, and as the mirror image is distal and unconnected, he acquires the knowledge that the mirror image is only a *representation* of his actual body. The import of this fact for self-consciousness is explained next.

**The Indirect Interpretation**

I think of this approach as 'indirect' because, unlike Gallup, I do not make a direct link between the *self* that is recognised in MSR and the possession of a *self*-concept. It is significant that the *self* is recognised, but not in the way Gallup and his supporters say; the route from MSR to self-awareness is not direct. I propose to lay out this pathway based on the hypothesis developed in chapter 3, the Concept Possession Hypothesis of Self-Consciousness (CPH). On this hypothesis, concept possession is sufficient for self-consciousness. If so, a strong case can be made for the existence of self-consciousness if a subject is demonstrably concept possessing. In chapter 4 I discussed several possibilities for demonstrating the existence of concept possession, including symbol-mindedness. As I explain below, MSR provides an excellent paradigm as a positive demonstration of self-awareness because as evidence of symbol-mindedness it is, I argue, incontrovertible.

Earlier when discussing kinaesthetic-visual matching theories of MSR, I proposed that organisms demonstrating MSR could take their mirror images in one of two ways. Either the mirror image is an extension of their own actual body, or it is a representation of their body. I then argued that the former alternative is implausible, given that proprioception matches body movements with visually accessible actual body parts but not with their mirror images (and, of course, because the images are distal and unconnected with the actual body). If so, then MSR subjects understand their mirror images as *representations* of their own bodies. In effect, they are treating the images as symbols – they are not their actual bodies but representations of them. As previously argued, symbol-mindedness requires concept possession and concept possession, by CPH, implies self-consciousness. However, an argument could be made that the mirror image is being used as a natural sign of positional and configurational information rather than as a representation, in which case concept possession need not be assumed. In fact, I would indeed strongly argue that for *most* mirror images the explanation of the MSR data in terms of natural signs is plausible and preferable. However, I argue that when the mirror image is taken specifically as the *self* this argument is not available. Rather, the most plausible interpretation in this case is that the subject takes the image as a representation of the self, and hence is symbol-minded.

A natural sign correlates with aspects of the environment but, unlike a symbol, is not ascribed any meaning by an apprehender. Take the example of footprints in the sand. These can be considered natural signs rather than symbols as they correlate with an aspect of the

environment (i.e., the path walked). An animal might just detect the affordance of paw prints to lead to (say) a waterhole. This could be just a learned association due to many exposures to the advent of paw prints connected to waterholes (or conceivably a genetically hard-wired pre-disposition). An organism need not possess concepts in this case. Nevertheless, when an organism interacts with a natural sign, this does *not* mean that they are not concept possessing. For example, a human can undergo reasoning along the lines of "those are my footprints from when I started the walk, therefore by following them backwards I can find the trailhead". To undergo such reasoning implies concept possession.

On the other hand, footprints can be actual symbols. For example, in some buildings such as hospitals or pre-schools, artificial footprints can be placed on the floor deliberately with the intention of leading people in particular directions. In such cases they act just like arrows on a signpost. These artificial footprints are then not natural signs but symbols. In some cases, conceivably, an apprehender might not even know whether the footprints are natural or artificial. Nevertheless, either way, the apprehender might react to the footprints in exactly the same way, i.e. to follow them (as in the earlier example of paw prints). Deacon [1997] emphasises that no objects are intrinsically symbols but are *interpreted* as such (p71). This point is crucial for my argument: it matters not whether the sign is intrinsically a natural sign or a symbol but how the apprehender interprets it.

In almost all cases involving non-linguistic organisms determining how the apprehender has taken the natural sign is extremely difficult as all we have to go on is behaviour. Animal behaviour can almost always be interpreted in multiple ways with varying levels of assumed cognition involved, just as the foot/paw prints example shows. Another simple example I have commonly used is when a lab rat presses a lever to gain a reward. One might wish to argue that the rat *knows* (that is, has a conceptual understanding) that pressing the lever will result in a food pellet. But, according to Morgan's canon (as discussed in chapter 4), we should explain animal behaviour according to the lowest psychological faculty feasible, and assume no more than that the rat has encoded an association between the lever press and a reward via training. Indeed, non-mentalistic accounts of MSR present just such associative types of explanation. However, in the case of MSR I argue that ascribing concept possession to the subject is indeed the more plausible explanation.

A mirror image correlates with aspects of the environment and so can be considered intrinsically a natural sign. But a mirror image, like footprints, could still be *interpreted* by an

apprehender as a symbol. Also like the footprints example, in most cases we could not tell from observation of a subject's interactions with the mirror whether it is using the image as a natural sign or interpreting it as a symbol. However, when the image is specifically of the *self*, we *can* tell. To start, let's examine the non-self case, say, the image of a banana.

Imagine a set-up whereby a mirror is angled in such a way that a primate can see the reflection of a banana but initially cannot see the actual banana due to a removable barrier. It is easy to predict that the primate will first seek the banana behind the mirror where it appears to be located. Most likely it will eventually locate the actual banana, and we can safely assume that after repeated trials the subject will learn to locate objects using a mirror. Now, it is tempting to say that the animal has learned about mirror properties and knows that the image in the mirror is not an actual physical object located behind the mirror (recall Pepperberg et al.'s conclusion regarding parrots described in section 5.4). If so then I would be committed to ascribing self-consciousness to it. The reasoning is as follows. If the mirror image is understood to not be a physical object located behind the mirror then it must be a *representation* of the actual object. In effect, the image is being treated as a symbol: it stands in for something else (the actual object). Thus, if a subject takes a mirror image – *any mirror image* – as not a physical object behind the mirror but standing in for one, then it is symbol-minded.

However, in the banana scenario a plausible explanation exists that does not assume such a high level of cognitive capacity. It could just be (for instance) that the primate has learned to associate the mirror image with the existence of a similar object in a different location. Even though it may eventually stop searching behind the mirror, we cannot assume that it realises that the image it sees is not a real banana – it has simply stopped looking because that activity has never been rewarded by success. Quite plausibly, the subject takes a mirror image as an actual, physical object behind the mirror, but just one that it cannot retrieve. None of this requires any conceptual understanding; the subject may simply be using the mirror image as a natural sign.

For most mirror images, as just argued, we simply cannot assume the subject takes them as symbols – even after they have stopped searching for them behind the mirror. But now, this is where MSR comes in. The one case for which we can be sure that the subject has taken the image as a representation rather than an actual object behind the mirror is when it has recognised the image as itself. Mirror images are not actual objects behind the mirror, but it is

likely that subjects that are *not* capable of MSR continue to take them to be so. Monkeys seeing their own mirror images continue to take them to be conspecifics. But a chimpanzee *cannot* be taking its mirror image as a physical object behind the mirror. This is because (as argued earlier) it has definitively taken the image as *itself*. All organisms can differentiate the physical self from non-self, perhaps mainly by proprioception. Since the image is not (part of) the physical self but is still understood to be the self, the most plausible interpretation is that it is representational of the self (in effect a symbol). A stark demonstration of this interpretation is that when a chimp looks at the distal image in a mirror, it reaches for its *own* forehead to touch the mark. It understands that what it sees in the mirror is not a physical object behind the mirror (a conspecific), but a representation of itself (its own body) onto which a mark has been applied.

At this point objectors might propose that this still unnecessarily ascribes too high a level of cognition to the chimpanzee; it might still just be (non-conceptually) interacting with a natural sign (or, put another way, detecting the mirror's affordance to locate marks on its forehead). The analysis provided in the first part of this chapter strongly argues in favour of the validity of the mark test as conclusive evidence of self-recognition, and I have also argued that an MSR subject cannot take its mirror image as a physical extension of its body. Nevertheless, the argument would go, the subject does not recognise the image as a representation but just learns a simple association between the mark on the image and the existence of a similar unseen mark on its own forehead. But this interpretation is implausible: one must question how such an association could be formed in the first place (that is, between the seen mirror image of the mark and the unseen actual mark on the forehead). Compare this with the association that *could* be formed with images of non-self objects as in the earlier banana example. Furthermore, on the associative account, one must also question why chimpanzees learn this association while the morphologically similar monkeys never do. It is more likely that the phylogenetically more advanced chimpanzees are exercising a more developed cognitive capacity, one that allows them to grasp mirror images as representations. The human example supports this view: symbol-mindedness is a capacity acquired as children become more cognitively developed.

It is interesting to compare the results of experiments on infant MSR with those of DeLoache's [2004] experiments on infant symbol-mindedness as described in the previous chapter. The age for the onset of infant symbol-mindedness (about 18 months) is similar – or

perhaps prior to – that for infant MSR (18-24 months). This is consistent with the claim that MSR is an indication of symbol-mindedness[49]. More specific experiments are needed to verify this. If my analysis is correct, infants cannot pass the mark test if they are not already symbol-minded. Therefore, the same infant subject should pass tests of symbol-mindedness at the same or earlier age than when they pass the mark test. I suggest longitudinal experiments be carried out to establish that symbol-mindedness arises prior to or concurrent with MSR in the same subjects.

## 5.7 Conclusion

The non-mentalistic interpretation of MSR discounts any need to invoke self-awareness as an explanation. However the argument that MSR is supported by a body-schema alone is implausible. Furthermore the non-mentalistic arguments cannot adequately explain why only *some* primate species demonstrate MSR. At the other extreme, the Gallup approach to MSR implies the full range of cognitive capacities associated with self-awareness, and is unjustifiable. The Gallup approach assumes self-awareness to be an all-or-nothing capacity but the evidence from infant studies shows that MSR is one of a progression of cognitive capacities. Furthermore, no reasoned argument is provided to sustain the Gallup approach.

I have used the 'Indirect Interpretation' to argue that MSR does provide sufficient evidence of self-consciousness. The argument is based on MSR being a demonstration that the subject has essentially taken its mirror image as a symbol representing its own body and as such is symbol-minded. Being symbol minded means being concept possessing, which, according to the Concept Possession Hypothesis, is sufficient evidence of self-consciousness.

Non-human animals that have so far convincingly demonstrated MSR are the great apes (save wild-born gorillas) while of the non-primates there are promising signs of MSR in dolphins, elephants, and magpies. When MSR is confirmed beyond reasonable doubt, these animals must be considered self-conscious organisms in some fundamental yet highly significant way.

[49] DeLoache [personal communication] has indicated that, to her, mirror images are not symbolic as they do not stand for anything as other symbolic objects do. However, I maintain that it depends on how the subject takes the image. Consider a picture hanging on a wall: DeLoache would agree that children above a certain age treat this as a symbol. But now substitute the picture for a mirror (at a certain angle) showing a similar image. Subjects might treat what they see just the same as for the picture – i.e., as a symbol. DeLoache herself used a similar trick as described in chapter 4: the 'window' view increased the symbolic nature of the scale model. A mirror should have much the same effect.

Exactly how significant remains for further discovery and debate, however, we can get some appreciation of it by comparing with a known standard: humans. Human infants demonstrate MSR at around the 18-24 month age range, at which age it is probably safe to say they have a sense of their own existence. It appears not unreasonable to equate the cognitive capacities of animals that have demonstrated MSR with those of (at least) 18-24 month old children.

# Chapter 6: Imitation

## 6.1 Introduction

In this chapter I evaluate imitation as a research paradigm for self-awareness. The majority of authors on this subject link imitative behaviour with self-awareness (e.g., Hart & Fegley [1994]; Parker & Milbraith [1994]; Asendorpf [2002]; Rochat [2002]; Gallese [2005]; Goldman [2005]; Gordon [2005]; Hurley [2005]; Meltzoff [2005]; Wolfgang Prinz [2005b]) though there are also pockets of resistance [Millikan 2005; Rawlins 2005]. But there are a variety of different types of behaviour that can be called imitation and no single, universally accepted definition of the concept. So rather than asking if imitation is a valid indicator of self-awareness, here I invert the question and ask which *types* of imitative behaviour can be taken as conclusive evidence of self-awareness. The answer is those types of imitative behaviour that occur as a result of conscious intent by the imitator. A particular case is Selective Imitation, in which the subject does not imitate the observed actions in total but selectively, choosing those actions it considers most beneficial.

### What Counts as Imitation?

Virtually all agree that, as a minimum, imitation implies the replication by an observer (or 'imitator') of an action performed by another (the 'model' or 'demonstrator'). However, while some authors are content to use this rather broad definition (or similar) as the sole criterion [Wolfgang Prinz 2005a; Jesse Prinz 2005; Meltzoff 2005], others impose more stringent requirements such as that learning is involved [Rizzolati 2005; Decety & Chaminade 2005] or that purposiveness must be demonstrated [Donald 2005; Zentall & Akins 2001] or that 'true imitation' implies the copying of novel actions (i.e., actions not already in the observer's motor repertoire) [Anisfeld 2005; Zentall & Akins 2001] or that it be both novel and complex [Byrne 2002]. For my purposes it is not important to define 'true imitation' precisely as my intention is to uncover that type of imitative behaviour that will demonstrate self-awareness. So, rather than quibble over what counts as 'true imitation', I prefer to view all these types of behaviour as imitation that can be characterised as either 'cognitively weak' or 'cognitively strong'. Cognitively weak imitation is of a type that

requires only very low or no conscious intent by the observer and will include, for example, reflex-like actions such as ‘catching yawns’. Cognitively strong imitative behaviour requires that there be clearly demonstrable conscious intent by the imitator. This is not to presume that imitative behaviour is dipolar, for there is probably a gradation of increasing levels of cognition involved in different types of imitative behaviour.

I begin, in section 6.2, with a discussion of the link between imitation and self-awareness, arguing that where the imitator is able to infer the mental state of the demonstrator this implies the capacity for a theory of mind and hence self-awareness. This conclusion is also consistent with the Concept Possession Hypothesis of Self-Consciousness (CPH) as described in chapter 3. As such, I examine which types of imitative behaviour meet this criterion. Before analysing the different types of imitative behaviour, however, I examine in sections 6.3 and 6.4 the discovery of ‘mirror neurons’. These have been used to argue both that imitation is ‘hard-wired’ (and as such not indicative of self-awareness) and also that it underpins ‘action understanding’ (which, according to CPH, would be indicative of self-awareness). I conclude that mirror neurons can indeed account for much imitative behaviour that is reflex-like. I also conclude that there is no need to attribute the capacity for action understanding on the basis of mirror neuron experimental results.

In section 6.5, I provide a taxonomy of imitative behaviour ordered in increasing cognitive dependency. From this I determine those that can be used as an indicator of self-awareness and their potential for use in animal experiments. In this regard I pick out Selective Imitation, which has been used successfully in human infant experiments. Selective Imitation has not yet been successful in animal experiments in my view, but may be so with suitable modification.

## 6.2 Imitation and Self-Awareness

The route from imitation to self-awareness is via Theory of Mind (ToM), which is also consistent with the Concept Possession Hypothesis (CPH) as discussed in chapter 3. An organism that has the capacity for ToM is able to infer the mental states of conspecifics. If, as I argue below, the capacity for ToM relies on *self*-knowledge, then a demonstration of ToM ability is a demonstration of self-awareness. If imitation is implicative of ToM then imitation,

too, is a demonstration of self-awareness. As with any observed animal behaviour, we should expect difficulty in interpreting animal imitation. For example, we will need to be able to distinguish imitative behaviour that is a true case of mind-reading from that which is merely 'behaviour-reading' [Sterelny 2003]. Nevertheless, I aim to show that in some cases this is possible. Imitation (of the right type) is linked to ToM if it means the imitator was able to infer the mental state of the demonstrator. That would be the case, for example, if the imitator understood the intention of the demonstrator and therefore performed the same action in order to produce the same result. As I discuss below, imitation comes in various guises, some of which require no (or very little) cognition at all. Therefore, in order to argue that at least some type(s) of imitation implies self-awareness it will be necessary to identify those types of imitation that definitively demonstrate ToM in something like the example just described.

According to CPH, concept possession alone is sufficient evidence of self-awareness. Therefore, any type of imitation that requires concept possession would be evidence of self-awareness. As discussed in chapter 4, one way to demonstrate concept possession is through rationality (i.e. inferential reasoning). Such a demonstration would, for example, be the ability to infer the intention of the demonstrator. This of course is why CPH is consistent with ToM as an indication of self-awareness: inferring the intention of the demonstrator was earlier given as a paradigm example of ToM. However, there are other reasons to consider ToM to be an indication of self-awareness independently of CPH, as discussed next.

**Theory of Mind**

A good *prima facie* case can be made that ToM requires self-awareness. According to Premack and Woodruff [1978], "An individual has a theory of mind if he imputes mental states to himself and others" (p515). Thus, implicit in this definition by the researchers who coined the phrase in the first place, is that in understanding the mental states of others one must also have the capacity to understand one's own mental states. My usage of 'understanding' in the previous sentence is deliberately intended to imply true *knowledge* of the mental states in question, rather than just an empathetic common experience (of, say, an emotion). In part 1 of this thesis I argued that an organism could have non-conceptual awareness (and hence no *self*-awareness) of some its own mental states (such as emotions). In the context of ToM, however, *conceptual* understanding of mental states is inherent.

Carruthers & Smith [1996] describe ToM as the ability to explain and predict the actions of oneself and other intelligent *agents*. Implicit once again in this conception is not only reference to the *self* alongside *others*, but also the reliance on *inference* (here construed as the ability to *explain and predict*) and to intelligent *agency*, the latter given as a hallmark of self-awareness in chapter 1. Despite this close association between ToM and self-awareness, ToM has been interpreted in multiple ways worthy of brief analysis.

The leading competitive theories of how ToM works are the 'simulation theory', the 'theory theory', and (to a lesser extent) the 'rationality theory' [Goldman 2005]. There are of course variations on each of these and they may not even be fully mutually exclusive (for example, Perner [1996] argues for a mixed simulation/theory account of ToM). These variations notwithstanding, these primary categories of ToM theories can be summarised as follows. In the *rationality theory*, an organism attributes rationality to their conspecifics and ascribes to them mental states that are rational for them to have under their particular circumstances. In the *simulation theory*, an organism creates pretend states intended to match those of the target and by applying its own internal mental state-generating mechanisms thereby replicates the mental state of the target. In the *theory theory*, an organism forms theories, or judgments, about the mental states of conspecifics and by extracting some psychological generalisations from an internal knowledge base infers the subsequent or prior mental states of the target. An essential element is common to each of them: that the organism must infer the mental states of conspecifics by reference to its *own* mental states in some way. If the rationality theory is correct, it is because the organism uses *itself* as a model of rationality; if the simulation theory is correct the organism appropriates its *own mental mechanisms* to simulate the target's mental state so as to generate a prediction of the output state; and if the theory theory is correct the organism directly matches the target's behaviour against *internally* recalled behaviours and their associated mental states. In each case the organism must have access to its *own mental states*. Thus, the self is the reference point for awareness of the other. On these views there is a reasonably strong case for considering any demonstration of ToM to imply the existence of self-awareness in the subject.

According to some authors (e.g., Gopnik & Meltzoff [1994]; Carruthers [2009b]; Carruthers, Fletcher & Ritchie [2012]) the capacity for self- and other-awareness develops concurrently. They suggest that the mechanisms involved in both self- and other-knowledge are the same and applied equally to both. Those mechanisms are a combination of perception and

inference. They dispute that we have direct (inner) *perception* of our own minds but *inference* of others' minds. They believe both perception *and* inference apply equally to access of our own and others' minds. So, we must infer our own mental states just as we infer the mental states of others[50]. Accordingly, Carruthers et al. [2012] predict that infants will become capable of self-consciousness as soon as they become capable of third-person mindreading. Some authors are sceptical of this view. Nichols [2001], for example, maintains that there are special introspection mechanisms for detecting one's own mental states but not for *reasoning* about one's own mental states. If so then an organism might have self-awareness without simultaneously a ToM. Indeed, in some versions of the ToM theory it is assumed that self-consciousness evolved first, allowing organisms introspective sense of their own behaviour, and was followed by the capacity to project this understanding onto others (see for example Povinelli & Prince [1998]). By analogy, self-awareness would also develop ontogenetically prior to the development of the capacity for ToM. In any case, these theories are common in the view that ToM does not exist *prior* to or in the absence of corresponding self-awareness. As such, for my purposes here it is unnecessary to adjudicate these views. Whether or not self-consciousness arises prior to ToM or self-consciousness and ToM arise concurrently, in either case the existence of ToM implies the existence of self-consciousness. Thus, if ToM can be demonstrated in a subject this provides evidence for the existence of self-consciousness in the subject.

### 6.3 The Mirror Neuron System

In the early 1990s a group of neuroscientists working at the University of Parma in Italy were researching the brain's motor cortex when they noticed a peculiar phenomenon. When one of the experimenters grasped a piece of food, certain neurons within their monkey subjects' motor cortex would fire in the same way as when the subjects themselves grasped the food. As these neurons seemed to directly reflect in the observer's brain acts performed by another, the scientists named them mirror neurons [Rizzolati, Fogassi & Gallese 2006]. This discovery had obvious and immediate implications for imitation researchers.

[50] In chapter 2 I also argued in favour of this view.

**Mirror Neurons and Self-Awareness**

The discovery of mirror neurons has fuelled the debate over the innateness or otherwise of imitative capabilities in organisms, and the level of cognitive capacity required to successfully imitate. This issue is important because if imitation is innate then (as argued in chapter 4) it cannot be used as a paradigm for self-awareness research in animals. We might observe similar behaviour in humans and in animals, but while in human cases we can usually be sure whether the actions of the imitator are under conscious control, this might not necessarily be the case for animals. If the advent of mirror neurons means primates are 'pre-wired' to imitate, then imitation (or at least many types of imitative behaviour observed in non-human primates) might be considered cognitively weak in not requiring much by way of conscious control. This would make it very difficult to attribute self-awareness to non-human subjects on the basis of observed imitative behaviour. For example, one such view is that imitation is subserved by a *common representational domain* for *perception* and *action*. Because the same neurons fire when an action is either observed or performed, the observation of an action may *trigger the motor activation* of the same action in the observer [Decety & Chaminade 2005]. In other words, the 'motor resonance' induced by observation of actions may subserve imitation[51]. This stimulus-induced imitation would of course not be indicative of any self-awareness.

There are several objections to motor resonance type theories, however. According to Susan Jones [2005], although mirror neurons respond to both sensory input and motor events, we cannot be sure of a causal connection; they do not necessarily respond to sensory inputs *with* motor events. Support for this view is the fact that monkeys have mirror neuron systems but monkeys are not particularly good imitators [Hurley & Chater 2005a, p3]. Humans also possess a mirror neuron system but, whereas the motor resonance theory might explain reflex-like imitation such as 'catching yawns', there must be an explanation for the fact that humans do not imitate habitually. Either motor resonance on its own is insufficient to produce actual motor activation in humans, or if it is then, as some believe, an inhibitory mechanism must exist to allow prevention of it [Kinsbourne 2005a; 2005b; Goldman 2005; Claxton 2005]. Some evidence supporting this view is that 'unwanted imitation' occasionally *does* 'break through' in some pathological cases such as dysfunction caused by brain damage

[51] There are several variations of this view. For example: 'active intermodal mapping' [Meltzoff 2002, 2005]; 'action modulation through perception' [Wolfgang Prinz 2005a]; and the 'shared circuits' hypothesis [Hurley 2005].

[Kinsbourne 2005a, 2005b]. Thus, at least in the case of normal human adults, some type of motor resonance acting on mirror neurons possibly predisposes us to spontaneously perform acts of imitation but we are (usually) able to consciously intervene to inhibit the action.

Some form of the motor resonance theory is probably correct, and as such there is a good deal of hard-wired predisposition to imitate in organisms with a mirror neuron system. As a consequence, we should be wary of using imitation per se as an indication of self-awareness, since the imitative behaviour might depend on nothing more than *perceptual* awareness. The motor resonance theory, I argue below, probably accounts to a large degree for neonatal imitation. Nevertheless, as I explore later, we can identify *some* imitative behaviours that will not be explainable simply by a motor resonance theory. These cases may be candidates for imitative behaviour indicative of self-awareness.

**The Correspondence Problem and Neonatal Imitation**

According to Hurley & Chater [2005] imitation requires a solution to the *correspondence problem*: how can the perceived action of another agent be translated into similar performance by the observer, especially when the imitated movement is perceptually opaque? For example, as an observer, one's own raised eyebrows cannot be *seen* and can only be *felt*. However, if mirror neurons are indeed transducers for motor activation as suggested by the motor resonance theory, then this problem is solved. As mentioned earlier, this account explains 'involuntary' imitative actions such as catching yawns. In addition, motor resonance can explain neonatal imitation, such as the well-known trick of sticking out a tongue at a newborn to make it do the same. Gopnik & Meltzoff [1994] claim that neonates perform such imitative acts from as young as 42 minutes old. According to Meltzoff [2005], 12-21 day old babies responded differentially to modelled tongue protrusions with tongue protrusions and not lip protrusions, indicating an ability to 'identify body parts'. They also responded differentially to lip protrusions rather than to lip openings, indicating the ability to imitate different movements of the same body part. That such young babies are able to do this can be explained if the mirror neuron system is a part of a hard-wired body schema and the motor resonance theory is correct.

Meltzoff [2005] argues at length that imitation begets ToM and, based on neonatal imitation, asserts that infants can infer the goals of adults. I argue that *inference* is not the mechanism involved, just some kind of reflex based on the motor resonance theory. There are a host of other sceptics about neonatal imitation generally (e.g., Kagan [1998]; Heyes [2005]; Gordon [2005]; Anisfeld [2005]; Jones [2005]; Elsner [2005]). Based on an exhaustive review of the literature, Anisfeld [2005] concluded that there was little evidence for neonatal imitation of invisible gestures, protesting methodological problems, atypical response rates, and inconsistent results. For example, although several studies showed an increase in tongue protrusion responses when tongue protrusion was modelled, there was no increase in mouth openings when mouth openings were modelled. Furthermore, both Jones [2005] and Kagan [1998] claim that tongue protrusions may be accounted for by an arousal interpretation, as it can be brought about by many other stimuli such as flashing lights or a pencil pointed at the mouth. Elsner claims that the apparent ability to imitate disappears after the first few weeks of life and reappears around 6-9 months later, a pattern which resembles that of neonatal *reflexes* (e.g., stepping reflex, grasping reflex). Thus, it is not necessary to invoke ToM to explain (apparent) neonatal imitation. The motor resonance account provides sufficient explanation for the observations.

The motor resonance theory provides a solution to the correspondence problem (especially in regard to neonatal imitation) and an explanation of involuntary imitative acts in adult humans. As such there appears to be good reason to accept that mirror neurons support at least some types of non-conscious imitation. However, it remains to explain why monkeys are not good imitators, since we can imagine that if the motor resonance theory is correct it is likely that monkeys would be highly predisposed to act imitatively often.

**Monkey Imitation**

It has been suggested that the existence of mirror neurons in monkeys made them ‘imitation ready’ and that the monkey mirror system could represent an evolutionary precursor of the mechanism for imitation in more developed primates [Iacoboni 2005; Goldman 2005]. For example, although monkeys do not imitate, chimpanzees do [Byrne 2005]. The difference between the more advanced human mirror system and the less advanced monkey version may account for the lack of imitative abilities in monkeys. Comparative experiments on the human

and monkey mirror systems confirm a functional difference [Rizzolatti 2005][52]. In some experiments several variations of demonstrator actions were tested. Not only were 'transitive' (meaningful) movements tested (the grasping of objects) but also 'intransitive' movements such as meaningless arm gestures. In both cases activation of the human mirror system was detected, but in the monkey mirror system only the transitive movements were activated. In another experiment, mirror neuron activation caused by observation of a grasping action was recorded at phased intervals. The results indicate that the human mirror system codes for the *temporal aspect* of observed actions, that is, coding of the precise chain of movements forming an action. Again, the monkey system does not do this. These experiments highlight two important differences between human and monkey mirror systems. In monkeys, the mirror system was not activated by intransitive movements, and the temporal aspect is not observed. Thus, according to Rizzolatti [2005], a lack of representation of intransitive actions plus a paucity of mirror neurons coding for precise copies of actions possibly present limits to monkey capacity for imitation. This would explain why monkeys are not *good* imitators (i.e. cannot replicate relatively complex movements). Monkeys have been shown, however, to imitate relatively *simple* actions such as lifting the lids off canisters [Voelkl & Huber 2000], and this is not inconsistent with a motor resonance theory.

**Summary**

Overall, it seems likely that the mirror neuron system subserves imitation. Plausibly, motor resonance of some type explains very simple imitative actions and involuntary imitative actions. Motor resonance combined with some kind of (voluntary or involuntary) inhibitory mechanism can account for the fact that more developed primates (i.e. humans) do not imitate habitually. At the same time, some type of motor resonance can solve the correspondence problem, explaining (for example) neonatal imitation of facial gestures. As some imitative acts may therefore turn out to be nothing more than reflex-like actions, we need to be stringent in ruling out this possibility before concluding that (a type of ) imitation indicates ToM and hence self-awareness.

[52] Unlike the monkey experiments, in which individual neurons were directly probed, monitoring of the human mirror neuron system was achieved through non-invasive techniques such as Motor-Evoked Potentials (MEPs) and Transcranial Magnetic Stimulation (TMS).

### 6.4 Action Understanding in Monkeys?

Before leaving the topic of mirror neurons we must address the implications of some results of mirror neuron experiments on monkeys that appear to indicate 'action understanding'. By CPH, if an organism is truly able to *understand* something – that is, if it has conceptual abilities – then it must be self-aware. In this section I examine the experimental results that might indicate this capacity in monkeys. I conclude that there is insufficient evidence of concept possession and hence no reason to ascribe self-awareness to monkeys on this basis.

The neurons in question reside in an area known as F5 in monkey brains; about 20% of which respond to visual stimuli. They are activated by actions in which the experimenter or the monkey interacts with an object; they discharge when a grasping act is either performed or when another agent is observed performing it. The target object's significance to the monkey has no influence on mirror neuron response – the same intensity of response is observed whether grasping food or a geometrical object. Interestingly, however, the mirror neurons will *not* fire at the sight of a pantomime of a grasp action in the *absence* of an object [Rizzolati 2005]. In one experiment, the final phase of grasping actions (the actual clasping of a target object) was blocked from the monkey's view by a screen: the monkeys only saw a hand reach behind the screen. The monkeys were aware of whether a target object was present or absent behind the screen, but in both cases the actual *visual* stimulus was identical since the target object area was visually blocked. Sure enough, the neurons encoding the grasping action fired when the target object was behind the screen but did not fire when it was absent. The discharge response could not have been dependent on the visual stimuli alone as they were identical in both cases. Rather, it appears that the monkey recognised the *goal-directedness* of the action.

Rizzolati [2005] suggests that these results indicate the subjects' "logical understanding" (p61) of the action. Furthermore, Rizzolati offers the following explanation of the observed effects: "When the motor templates represented by mirror neurons resonate, the *meaning* of the observed action becomes transparent..." (p60; emphasis added). These views imply concept possession by the subjects and if so then by CPH this means we must ascribe self-awareness to them. However, it is not necessary to assume concept possession to explain these results. Possibly, mirror neurons implement a simple non-inferential mechanism of *action recognition*, which, as one researcher put it, could be just a building block for imitative behaviour [Iacoboni 2005]. This non-conceptual action recognition can be thought of in the

same way as affordance detection as discussed in previous chapters. Adhering to Morgan's canon (that we should interpret animal behaviour according to the lowest psychological faculty feasible, as discussed in chapter 4) we should accept the simpler explanation. Plausibly, the firing patterns of the mirror neurons correlate not with conceptual understanding but rather with affordance detection. At best, the experiment indicates some capacity for object permanence in the monkey (it encodes information as to whether the object is behind the screen even though it can't see it), but we need not ascribe further cognitive abilities to the subjects based on these results alone.

In another set of experiments, monkey F5 neurons that were observed to discharge on presentation of actions *accompanied* by sounds were also observed to discharge in response to the sound *alone* [Rizzolati 2005]. Thus the stimuli leading to the 'action understanding' (visual or auditory) was not relevant; it was only the final result that mattered. One possible interpretation of this observation, favoured by Rizzolati, is that the neurons fire once the *meaning* of the action is specified. However, once again, it is not necessary to assume a conceptual understanding of the 'meaning' of the final result. In this experiment an association[53] is formed between the visual and auditory stimuli. Thus, either stimulus will produce the same response (in this case a specific neuronal firing pattern).

**Summary**

It is not necessary to assume concept possession in monkeys to explain the observations described in this section. As such there is correspondingly no need to ascribe self-awareness to monkeys based on these experimental results. In the next section I map out the full range of imitation-like behaviours and from this list I identify those that could be usefully applied in research on animals to determine the existence of ToM and hence self-awareness. I argue that Selective Imitation is such a case.

[53] A more detailed discussion of association theories generally is undertaken in chapter 8.

## 6.5 Types of Imitative Behaviour

There are a wide variety of behaviours describable as imitation in both humans and animals. In table 6.1 I outline these and attempt to order them from cognitively weak to cognitively strong. By cognitively weak I mean actions that are probably not under much conscious control of the organism, while cognitively strong implies the actions are deliberate. Now, at the extremes, it is quite easy to characterise certain imitative behaviours in this way; it is fairly obvious and uncontroversial to label (say) herding as cognitively weak and parody (a type of 'mimetic' behaviour) as cognitively strong. But in the middle the view is far less transparent. In any case the exact order is not important; the main concern is to identify a range of behaviours that are candidates for self-awareness indicators. So although I have included the full spectrum in my taxonomy of imitation for the sake of completeness; in the later analysis I concentrate on the cognitively strong end of the scale. I seek behaviours that we can and do observe in human infants and animals that we might consider good evidence for the existence of ToM and therefore self-awareness in the subjects.

Table 6.1: Proposed taxonomy of imitative behaviours

| Imitative Behaviour | Short Description | Comment |
|---|---|---|
| Animal Mimicry | The copying of the physical appearance of one species by another, so as to benefit from the model species' characteristics. For example, the markings on the apparently tasty viceroy butterfly mimic that of the unpalatable monarch butterfly and hence the viceroy gains a survival advantage by duping predators into avoiding it [Zentall & Akins 2001]. | Driven by natural selection of physiological characteristics; no cognition is involved. |
| Flocking/ Herding/ Schooling | This type of behaviour may be thought of as literally 'following the crowd' and as such counts as a very simple form of imitation in animals. | Perhaps a genetically predisposed form of social influence, but no cognition necessary. |
| Contagion | Two or more animals engaging in similar behaviour that is species typical. For example, the resumption of feeding by a sated animal simply by the introduction of a hungry animal that begins eating [Zentall & Akins 2001]. Contagion has also been used to describe human neonates that begin crying after hearing another baby cry. | A type of reflex-like imitation (see next category). |
| Reflex-like imitation | For example, 'catching yawns' or smiling when seeing another person smiling. Another type of example is when movie audiences repeat | Still cognitively weak and generated simply by the influence of a perceived |

| | | |
|---|---|---|
| | movements they see on the screen. Neonatal imitation, such as tongue protrusions, also probably belongs in this category. | action, though in some cases the imitator may be aware of the imitative action in progress. |
| Delayed imitation | Imitative acts that are performed after a substantial time delay rather than immediately | The advent of a delay indicates the possibility of cognition being involved. |
| Emulation | The reproduction of a goal, or the outcome of an action, without the reproduction of all the observed *actions* leading to the goal. | *Possibly* indicates 'action understanding'. |
| Synchronic imitation | Two subjects simultaneously play with the same type of objects in a similar way, presumably influenced by each other. | Possibly depends on the capacity for *spontaneous* perspective-taking. |
| Role play | Subjects imitate a *type* of behaviour, such as the typical actions of a bus driver. | Apparently depends on the capacity for *deliberate* perspective-taking. |
| Selective Imitation | Subjects discriminate between observed actions and imitate selectively rather than wholesale. | Appears to indicate some capacity for inferential reasoning and/or understanding of the demonstrator's intentions. |
| Mimesis | The motoric reduplication of an event for communicative purposes, requiring an audience to be taken into account [Donald 2005]. | Captures the urge to generate culture; it involves a high degree of social understanding and metacognition. Implies self-awareness. |

For imitative behaviour to be a demonstration of self-awareness it must be of a type that demonstrates a mind reading ability; an understanding of the model's mental state. Let us immediately discount without argument those types of imitation we placed at the lower end of the cognitive scale: animal mimicry, flocking, contagion and reflex-like imitation generally, as it is apparent that in none of these does mind reading play any part.

Moving to the other end of the cognitive scale, we find mimesis, a type of imitation pretty much defined by an assumption of self-awareness in its practitioners, human beings. Under the heading of mimesis falls such intentional imitative acts as parody, confidence tricks and group-specific activities such as custom and ritual. Mimesis involves a high degree of social understanding and metacognition and is uniquely human. Thus, although mimesis can be taken as definite evidence of self-awareness, it is not suitable as a research paradigm for animal studies. Instead we must turn our attention to the middle ground; to those imitative behaviours we placed around the centre of our cognitive scale. Each of these is now analysed in more detail.

## Delayed Imitation

As the name suggests, delayed imitation refers to imitative acts that are performed after a substantial time delay rather than immediately. According to Custance & Bard [1994] such acts are markedly different to imitation triggered immediately by a stimulus as they are presumably mediated by memory and are representational given that the memory is in the form of a stored mental image. Accordingly, delayed imitation must be considered cognitively stronger than reflex-like imitation. Meltzoff [2005] reports that 6-week-old human infants performed deferred imitation of mouth movements after a 24-hour delay. However, as previously discussed, neonatal imitation is controversial. Other evidence indicates deferred imitation of novel acts in infants only after about 11-12 months of age [Anisfeld 2005].

The fact that memory is involved in delayed imitation is not by itself sufficient evidence of self-awareness. Memory of an action does not imply propositional thoughts or inferential reasoning of the form 'I remember that to achieve this goal I must perform such-and-such an action'[54]. That a stimulus may trigger an associated response even after a substantial delay does not imply any form of concept possession nor does it imply there was any active ToM during the demonstration event. As such, we should not use delayed imitation as a research paradigm for the investigation of self-awareness.

## Emulation

One way to describe emulation is the reproduction of a goal, or the outcome of an action, without the reproduction of the sequence of actions leading to the goal. The fact that the actions are not reproduced leads some authors to discount emulation as a type of imitation [Gattis, Bekkering & Wohlschläger 2002], but clearly some replication has taken place, if only the endpoint of a series of actions rather than the whole series. Thus, emulation might perhaps be considered goal-directed imitation.

According to Whiten et al. [2005] chimpanzees are capable of emulation. They describe an experiment involving a clever invention they named the 'pin-apple'. This box-like device is

[54] The connection between self-awareness and different types of memory (especially *episodic memory)* is explored in depth in chapter 7.

so-called because it acts like artificial fruit: hungry subjects need to remove the device's defences in order to gain access to an edible core. The defences were two bolts, a pin and a handle. There were alternative ways to remove the defences. For example, the bolts could be poked through or could be pulled and twisted out, while the handle could be pulled or twisted out of the way. Subjects would be shown one of several methods of opening the box and then observed for their imitative ability. For one of the actions (handle removal), chimpanzees applied their *own* technique for removing the defence irrespective of the demonstrated action. Whiten et al. consider this to be relatively emulative behaviour.

In emulation some learning has taken place, so some reasonable amount of cognition is involved, but is mind reading involved? According to some authors (e.g., Gattis et al. [2002]; Tomasello & Carpenter [2005]) emulation requires some understanding of the demonstrator's goal or intention, which might be taken to imply mind reading is taking place. But 'understanding' a goal in this context need not involve inferring the demonstrator's mental state. An alternative conclusion is that what is 'understood' (or rather, *detected*) is the *affordance* of the object acted on. Thus emulation might be thought of as 'affordance learning by observation': the observer has learned the affordance rather than the imitated act (such as a bird learning one of several ways to remove a lid) [Zentall & Akins 2001; Byrne 2005]. In emulation learning, by watching the way a demonstrator interacts with objects, the observer may be learning quite complex new things about the objects and the environment, but not necessarily about the demonstrator's state of mind. An observer might not exactly replicate a demonstrator's acts upon an object, but might achieve the same end so long as the affordances of the object are activated. For example, an infant might discover that balls afford rolling, by observing a demonstrator acting upon one. But the infant might activate the rolling in any of several ways some of which may be discovered by himself. According to Tomasello & Carpenter [2005] infants can learn new things about objects and their affordances via emulation learning by the middle of the first year.

In the case of the pin-apple the chimpanzee learns that a bolt affords *poking*; a rod affords *twisting*; a handle affords *pulling*, and if combinations of these actions are enacted the box affords *opening* to reveal a prize. Having learned these affordances, the chimpanzee may employ any combination to open the box, that is, to replicate the observed goal. But learning that the box affords opening does not necessarily imply that the observer was able to infer this as the demonstrator's goal; it only reveals that the observer is able to learn affordances by

observation. It might be the case that the observer did indeed understand the model's intentions, but as we cannot definitively conclude this based on the evidence, we should not allow emulation as a test for self-awareness.

**Synchronic Imitation**

According to Asendorpf [2002], children in the latter part of their second year of life engage in synchronic imitation, in which two children simultaneously play with the same type of objects in a similar way. It is different from parallel play because it is 'real communication' as indicated by the usage of a common code (the shared activities). Asendorpf asserts that synchronic imitation requires the capacity for *spontaneous* perspective-taking, distinguishable from the later appearing *deliberate* perspective-taking (see *Role Play*). Spontaneous perspective-taking occurs as an immediate act of empathic identification with the play-partner, while deliberate perspective-taking occurs when the child is asked to take the view of others (such as in choosing a birthday gift a friend would like). If this is correct, then such behaviour looks to be reasonably cognitively strong, in that theory of mind appears to be involved.

However, Asendorpf's description of the activity does not rule out explanations that exclude perspective-taking. The distinctive features of synchronic imitation are given as the *visual regard* for the partner and the *reciprocity of behaviour*. It is conceivable that actions of this sort can occur without perspective-taking. Much species-specific behaviour in the animal world might be considered synchronic activities, for example courtship displays involving coordinated movements [Zentall & Akins 2001]. Perhaps some of these might even be called synchronic *play*, such as the 'play-fighting' of feline cubs that prepares them for future predatory behaviour. Such cases also involve visual regard and reciprocity of behaviour but might not involve ToM. Therefore, synchronic play does not appear to be a suitable paradigm for research into self-awareness in animals.

**Role Play (Pretend Play)**

By comparison with the apparently 'spontaneous perspective-taking' of synchronic play, *deliberate* perspective-taking is said to occur in children around the age of 2, when they begin to engage in role play (acting out the role of a person or animal) [Asendorpf 2002; Goldman 2005]. Role play tends to be creative; what is imitated is a *type* of behaviour rather than a match of actual observed behaviour. For example, a child acting out the role of a doctor will not necessarily duplicate recalled actions, but may engage in typical behaviours such as examining a 'patient'. Furthermore, the child may elaborate the event with novel actions not previously observed. Thus, role play includes mental imitation (or 'simulation'), in which the imitator is engaging to some extent in mind reading [Goldman 2005]. Role play, as described here, looks to be cognitively loaded. In effect, the imitator is mentally placing himself in the shoes of the imitated, in other words, demonstrating ToM. Observation of role play should therefore be taken as evidence of self-awareness.

Role play has not been reported in non-human primates. Monkeys do not engage in play except in play parenting (similar to human children), but this involves practising a single role that does not seem to involve pretence. Great ape children do engage in 'proto-pretend' play, which does not involve the relatively elaborate roles of pretend play in human children [Parker & Milbraith 1994] and indeed it may be somewhat difficult to induce such behaviour in controlled studies for non-human primates. Role play observation provides the opportunity for anecdotal evidence of self-awareness, but (at least for now) does not appear to provide a rigorous experimental paradigm for studies of animals.

**Selective Imitation**

Harris & Want [2005] have developed an interesting hypothesis to explain what they call the 'ratchet effect', the sudden (in evolutionary terms) cultural explosion that occurred at the Upper Palaeolithic era some 10,000-40,000 years ago. Briefly, they speculate that although hominins must have been capable of imitation for at least 1.4 million years (in order to account for the standardisation of tool manufacture), the ratchet effect may be explained by a shift from non-selective to selective imitation. In non-selective imitation the observer imitates all variants of observed actions indiscriminatively, whereas during selective imitation the observer will discriminate between them, favouring one over the other. The shift would

explain why, over the past 1.4 million years up until relatively recently, tool design remained virtually static and then suddenly became ever more sophisticated: hominins that were selective in their imitation would have favoured any improvements in tool design even if marginal and this process would have led to overall technical improvements over time. Of course, the true picture is likely to be more complex and dependent not only on such individual cognitive adaptations for cultural learning but also other variables (such as highly structured learning environments; conducive population structures; and so on [Sterelny 2012]). Nevertheless, this idea opened up a line of inquiry for Harris & Want.

Harris & Want [2005] inquired as to whether young children displayed any signs of selective imitation and designed experiments to find out. In one experiment they adopted a device called a trap-tube. In this device a toy is placed in the middle of a long tube that is open at both ends. The toy could be retrieved by pushing it out with a long stick, provided the stick was inserted into the correct end of the tube; if inserted into the other end the toy fell into a trap and was retrievable only by an adult (see figure 6.1). 2-year-olds and 3-year-olds were tested in the experiment. They were shown either the *correct* demonstration or the *incorrect + correct* demonstration. In the former, the experimenter inserted the stick into the correct end and thereby successfully pushed the toy out. In the *incorrect + correct* demonstration, the stick was first inserted into the wrong end, pushing the toy into the trap, at which point the experimenter said "oops!" and then later inserted the stick into the correct side to successfully retrieve the toy. Following the demonstration, the children were invited to retrieve the toy themselves. The results were telling. The 2-year-olds, irrespective of whether they had seen the *correct* demonstration or the *incorrect + correct* demonstration, performed at chance: they inserted the stick at random, trapping the toy on about half the trials. The 3-year-olds who had only seen the *correct* demonstration had the same result, also performing at chance. But the 3-year-olds who had seen the *incorrect + correct* demonstration were more selective: on about three-quarters of the trials they successfully retrieved the toy.

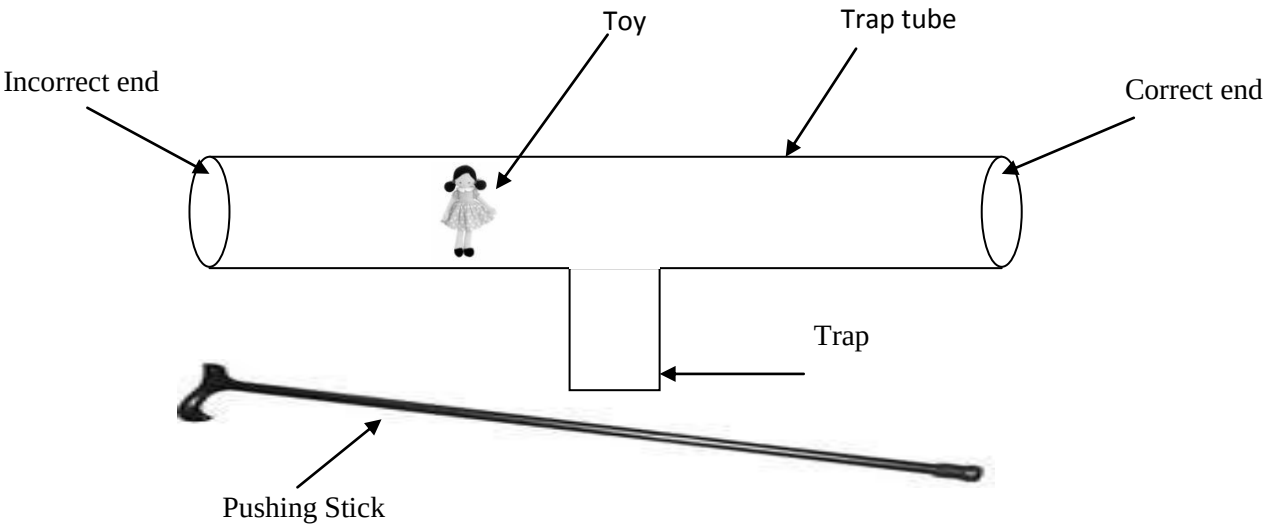



Figure 6.1. Schematic representation of the Harris & Want [2005] Selective Imitation experiment

These experiments are a clear demonstration of mind reading by some of the subjects. The 2 year olds were unable to infer the mental state of the experimenter modelling the toy-retrieval routine: statistically their behaviour did not change between the *correct* demonstration and the *incorrect* + *correct* demonstration. But the 3 year olds in the study did change their behaviour. Hearing 'oops!' exclaimed by the experimenter modelling the unsuccessful incorrect method and then observing him perform the successful correct method, the 3 year olds were able to infer the experimenter's mental state. The subjects could infer that the experimenter knew he had made a mistake in the first retrieval attempt and so deliberately changed methods to correctly retrieve the toy in the second attempt. Having thus read the mind of the experimenter, the 3 year olds were more likely to selectively imitate the correct retrieval method.

In the above analysis a theory of mind (ToM) argument is used to justify the view that selective imitation is a valid indicator for self-awareness. Although there is a good case to be made on ToM grounds, as explicated in the introduction to this chapter, there is also a further compatible interpretation based on CPH. In chapter 3 I argued that concept possession is evidence of self-awareness. If this is correct then we can apply the principles laid out in chapter 3 to defend the view that selective imitation is a valid indicator for self-awareness. Briefly, the idea is to show that a subject's behaviour provides conclusive evidence of

concept possession. As discussed in chapter 4, one such way is by demonstrating evidence of inferential reasoning. This approach appears to fit the observations in the selective imitation experiments. The 3 year old children who had seen the *incorrect + correct* demonstration were more likely to choose the correct method of toy retrieval, rather than performing at random. This provides good evidence they had *inferred* that there was an incorrect as well as a correct method. The trigger was the 'oops!' exclamation by the demonstrator. The 2 year olds did not make this inference but the more developed 3 year olds did. By the reasoning given in chapter 3, this means that the 3 year olds who were selective in their imitation have demonstrated self-awareness.

Selective Imitation appears to be a suitable paradigm for the investigation of ToM and hence self-awareness. However, much care will be needed in designing experiments for use with animals, as the following example demonstrates. Whiten, Horner & Marshall-Pescini [2005] performed experiments in both human infants as well as chimpanzees and concluded that chimpanzees also imitated selectively, in some cases more selectively than the infants. However, there were important differences between these experiments that might account for the different reactions of the chimpanzees and human infants. In the Whiten et al. experiment a series of actions on a box that led to the opening of the box was modelled. The same series of actions was modelled on an opaque box and a transparent box, the latter making it clear that some of the actions were not causally relevant to the opening of the box. 3 year old children largely imitated the whole series of actions to open the box even after they had witnessed the demonstration on the transparent box. That is, they did not elect to ignore the causally irrelevant actions. Chimpanzees, on the other hand, were selective in the actions they imitated after observing the actions modelled on the transparent box compared with the opaque box. They ignored the irrelevant actions and therefore more efficiently opened the box.

In this experiment the chimpanzees performed better than the 3 year old children. However, this might *not* be a case of selective imitation. Quite possibly this represents a case of *emulation* as described earlier. The chimpanzees may have been able to recognise all the affordances available in the transparent box and acted upon only those that resulted in the opening of the box. As earlier explained, detecting and acting upon affordances in this way does not depend on the subject inferring the mental state of the demonstrator. Contrast this with the Harris & Want experiment: in that case, the demonstrator signalled his mental state

by saying ‘oops!’ which enabled the 3 year olds to infer that the demonstrator realised he had made a mistake. The 3 year olds were thus able to infer which of the actions was the correct one and selectively imitate it. A comparable experiment usable on non-linguistic organisms would need to employ some way of similarly signalling the mental state of the demonstrator, which the transparent box experiment did not definitively do.

If the transparent box experiment did amount to a demonstration of emulation in the chimpanzees, a question still arises as to why the 3 year old children did not perform the same way. That is, why did the children not also detect the affordances of the transparent box and act only upon those necessary to open it? The answer is not immediately obvious and Whiten et al. [2005] speculate that perhaps children ‘overcopy’ in comparison to non-human primates. Their explanation is that “...we are such a thorough-going cultural species that it pays children, as a kind of default strategy, to copy willy-nilly much of the behavioural repertoire they see enacted before them” (p280). Whiten et al. also propose an equally plausible (or implausible) alternative explanation: the 3 year olds were mind reading on the experimenter and inferred that, in the absence of any indications to the contrary (no ‘oops!’), *all* the actions performed by the experimenter must have been necessary to the obtaining of the end goal. The upshot is, to re-emphasise the point made earlier, that in order to use a Selective Imitation paradigm the demonstrator must somehow signal his mental state as was done in the Harris & Want version. Furthermore, when applying this paradigm to animals, somehow this signal must be interpretable by the animal subjects since we cannot assume that an exclamation of ‘oops!’ would be intelligible to them. How this could be achieved is by no means clear but I do offer some suggestions in chapter 9 in the Future Research section.

**Summary**

The Harris & Want experiments on selective imitation provides a model for the positive testing of self-awareness in human infants, and with suitable modifications may be applied to animal species. Their results are good evidence that some 3 year old human infants are adept at theory of mind and are able to infer the mental states of others. As argued earlier, having ToM implies the existence of self-awareness, but further support is provided using the principles described in chapter 3 based on the Concept Possession Hypothesis of Self-Consciousness.

## 6.6 Conclusion

Imitation per se cannot be considered an indicator for self-awareness; there are too many different types of behaviours that can legitimately be considered a form of imitation and, despite the discovery of mirror neurons, it is not likely that the neural and/or psychological causes of the behaviours are the same in all cases. A more useful question to examine is *what type of imitation can be used as a positive indicator for self-awareness?* On the basis of the idea that ToM indicates the existence of self-awareness, both *role play* and *selective imitation* are candidates. In particular, selective imitation, using Harris & Want's protocol, provides a method that can be applied under controlled experimental conditions and may be adaptable to animal species. The technique provides good evidence of both ToM as well as inferential reasoning to support the claim that it demonstrates the existence of self-awareness in the subjects.

A negative result in the Selective Imitation task does not necessarily indicate the absence of self-awareness generally. We should not conclude, based on the Harris & Want experiments, that those 3 year olds who did not selectively imitate (or for that matter the 2 year olds) are therefore not self-aware. Having the capacity for inferential reasoning does not mean it will be used. Indeed, as discussed in depth in chapter 5, there is other evidence (mirror self-recognition) that infants as young as 18 months are self-aware. This means the technique may be used as another tool in the toolbox of self-awareness research as a positive test, if not a negative one. The key element in the human trials appears to be the signal for an incorrect demonstration (the 'oops!' exclamation), so equivalent signals recognisable by the species under study would need to be established.

# Chapter 7: Episodic Memory

## 7.1 Introduction

This chapter explores the link between memory and self-consciousness, with particular regard to whether tests for certain types of memory can be used as demonstrations for the existence of self-consciousness. The main focus will be on episodic memory, which has been linked to a form of self-consciousness (or 'self-knowing') known as *autonoesis* [Tulving 2005]. Other forms of memory will also be examined, however, in the context of the *Concept Possession Hypothesis of Self-Consciousness* (CPH) as espoused in Chapter 3. According to CPH, concept possession alone is sufficient evidence for the existence of self-consciousness. Therefore, any type of memory that requires conceptualisation would, by this hypothesis, indicate the existence of self-consciousness in the subject.

Different types of memory and their possible relation to self-consciousness are discussed in section 7.2. A particular class of memory, declarative memory, is found to be relevant to self-consciousness as this type of memory should imply concept possession. Declarative memory is subdivided into *semantic memory* (memory for facts) and *episodic memory* (memory of personally experienced past events). On analysis, semantic memory is found to be unsuitable as an experimental paradigm for self-consciousness because of the difficulty in distinguishing memory stored conceptually as opposed to non-conceptually (i.e. stored only as encoded raw information). Episodic memory, by contrast, has greater potential as an experimental paradigm for self-consciousness and as such the rest of the chapter is devoted to this topic.

Section 7.3 explores the defining characteristics of episodic memory to establish that it is indeed a distinct form of memory and to expose the aspects that are relevant to self-consciousness. In particular, episodic memory, unlike semantic memory, always involves the *self* and is phenomenologically dissimilar to semantic memory. Episodic memory is already closely associated with self-consciousness (or rather, 'autonoesis') by many authors. Section 7.4 examines this association and compares it with an analysis based on CPH. Autonoesis is found to be a more developed form of self-consciousness, in that it implies a concept of the self existing not only in the *present* (which is considered within CPH to be sufficient for self-consciousness) but also as existing in the *past*. As such, a conclusive demonstration of

episodic memory in a subject should be taken as evidence for the existence of self-consciousness in that subject.

To be useful as an experimental paradigm for self-consciousness, episodic memory needs to be clearly identifiable as such. With regards to stored memories, this would be very difficult to do in non-linguistic subjects, as observable behaviours of subjects could be the result of semantic memory. However, it has been proposed that the capacity for episodic memory implies a concept of time – that is, an understanding of the self's existence in past, present and *future*. I argue that experiments designed to demonstrate a concept of *future* rather than *past* should be less prone to misinterpretation and hence provide a suitable paradigm for experiments. In section 7.5 I analyse the claim that episodic memory implies a subjective sense of time and conclude that there is indeed sufficient reason to accept this claim. Thus, episodic memory experiments that exploit this feature (i.e. future orientation) may be superior to those that rely only on memory of past events. In section 7.6 I examine several experimental paradigms and conclude that those reliant on *past* memory cannot provide conclusive evidence of self-consciousness, while those based on future orientation (i.e. certain types of planning behaviour) can. Finally, in section 7.7, several experiments based on future orientation in various animals are reviewed. In most cases, there is reason to doubt that a sufficiently high standard of evidence has been met. However, one experiment on scrub jays does appear to have met these standards.

The key conclusions drawn in this chapter are as follows.

(i) Care should be exercised when using the term 'semantic memory'. Semantic memory is declarative, which means the content is factual (can be evaluated as true or false). But this does not mean that the possessor of the memory content has the conceptual capacity to perform the evaluation.

(ii) The term 'autonoesis', frequently associated with episodic memory, is to be distinguished from 'self-consciousness' in an important way. Autonoesis incorporates a sense of *subjective time,* which is not assumed to be necessary for self-consciousness, at least as I use the term. In my usage of 'self-consciousness' it is only necessary to have a concept of oneself existing in the present.

(iii) Episodic memory is a valid indicator for the existence of autonoesis (and hence also of self-consciousness).

(iv) Episodic memory has been convincingly demonstrated in scrub jays. Accordingly, we should be willing to accept that these creatures are not only self-conscious but autonoetic.

### 7.2 Which Types of Memory are Linked to Self-Consciousness?

Figure 7.1 maps out the different types of memory[55].

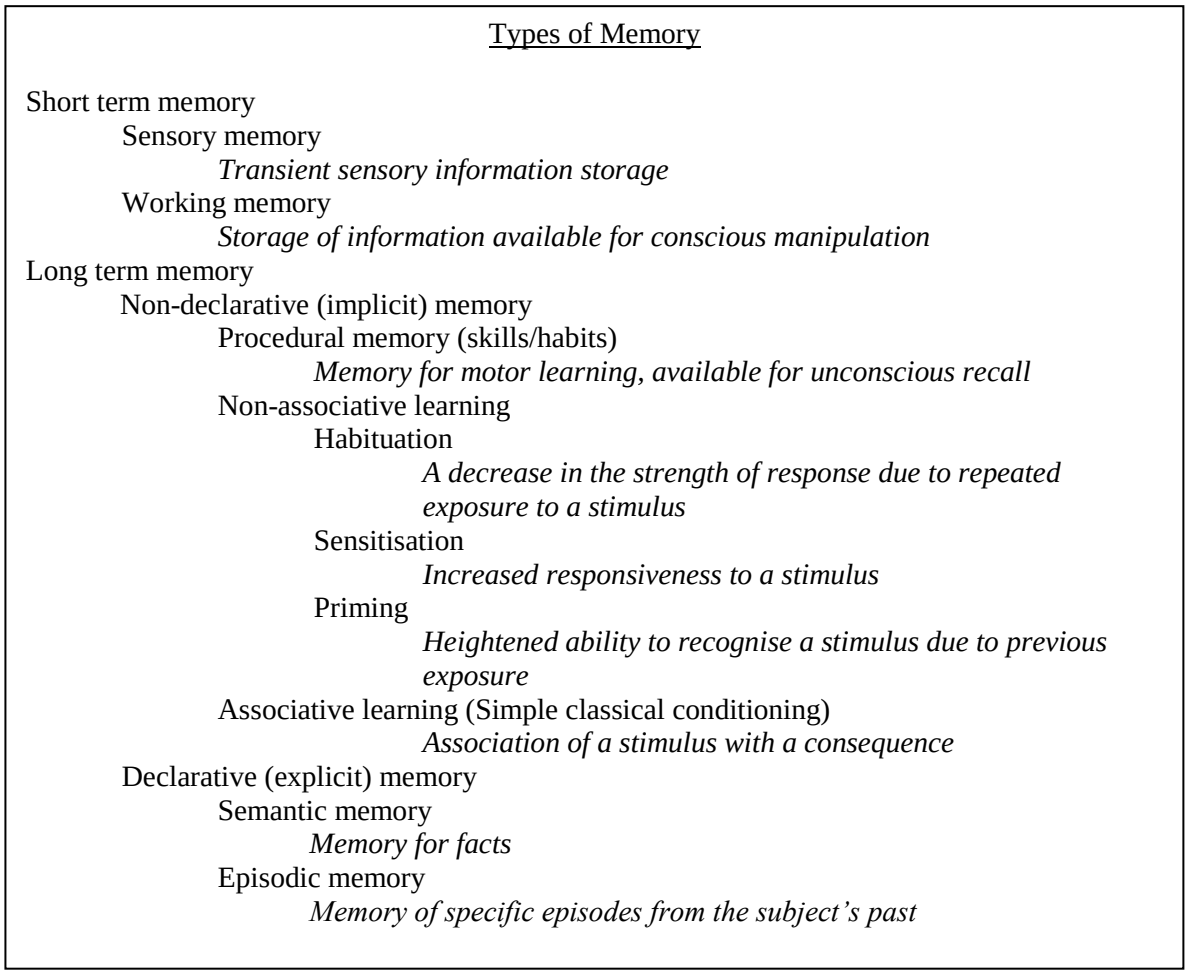



Figure 7.1: Showing the relationships between different types of memory with a brief definition or description. Based on Gluck, Mercado & Myers [2007] and Sweatts [2010]

[55] Some authors may wish to distinguish further varieties of memory (for example, Tulving [2007] tongue-in-cheekily lists 256 types he has seen in the literature!) However figure 7.1 covers the range generally accepted by the standard texts.

### Short-Term Memory

For simplicity I use the term ‘short term memory’ rather broadly here to include any kind of short-lived memory, such as sensory memory. There is no suggestion that sensory memory (or, sensory register) is linked to self-consciousness. Based on experiments on the visual mode of sensory memory (or, ‘iconic’ memory), sensory memory apparently acts as a buffer, encoding perceptual data for less than a second and unconsciously [Gluck et al. 2007]. On the other hand, working memory is associated with consciousness, since it is employed during active manipulation of the memory contents. Indeed, there is evidence of a correlation in humans between general intelligence and strength of working memory [Gluck et al. 2007, p182]. Since intelligence in this context requires concept possession, by CPH this indicates a connection between working memory and self-consciousness. The contents of working memory are available for conscious manipulation, such as in the action of mental arithmetic. In such an example it is clear that concepts are in operation and therefore, by CPH this is evidence of self-consciousness. However, this is but one example of working memory in action. It is possible that working memory is available to organisms that might not be concept-bearing. For example, working memory might be employed for short-term spatial memory or for short-term retention of recent actions, neither of which necessitates the possession of concepts. Therefore, a demonstration of working memory alone is not sufficient evidence of self-consciousness.

### Long-Term Memory

Long-Term memory is divided into two major groups: *declarative* and *non-declarative* memory. To my knowledge, there is no suggestion of a link between any type of non-declarative memory and self-consciousness. Indeed, it is likely that these forms of memory do not even rely on any direct *conscious* involvement, let alone *self*-conscious. These forms of memory storage and retrieval occur at the subpersonal level. Also, since these forms of memory are in no way reliant on concepts, there is no connection to self-consciousness through CPH. In the case of procedural memory, it can be argued that concept possession is involved in the decision to acquire certain skills – e.g., a deliberate decision to practise music playing. But once having acquired those skills – that is, having suitably encoded them in memory storage – no mechanism of self-consciousness is required to express them. In light of

this, it would seem that non-declarative memory will not provide insight into self-consciousness.

Declarative memory is divided into two further subdivisions known as *semantic memory* and *episodic memory*, each of which are further examined below. Declarative memory is so-called because it refers to content that is truth-evaluable. In other words, the content involves a representation that can be evaluated as being true or false. According to Dokic [2001]: "...the information-link underlying factual memory is *doxastic*..." and "...since beliefs have conceptual contents, the information retained in factual memory is always conceptual..." (p218). Dokic is referring to humans (who we already know possess concepts), however if this also applies to organisms generally then this would seem to imply the possession of concepts in non-humans, too. If true, by CPH, I would accept a demonstration of declarative memory in animals as evidence of self-consciousness. In the case of episodic memory, a *direct* link with self-consciousness has been made [Suddendorf & Corballis 1997, 2008; McCormack & Hoerl 2001; Conway 2001a; Dokic 2001; Tulving 2005; Zentall 2005; Dere et al. 2008; Easton & Eacott 2008]. However, the picture is not so clear cut, as discussed next.

**Semantic Memory**

Semantic memory is memory for facts. The fact might be of the type: 'pressing this lever releases an item of food'. *Knowledge* of this fact implies possession of concepts – in this case the concepts PRESSING, LEVER, FOOD, etc. In the case of a rat, however, we should question whether the content described is *known* by the rat. In other words, does the rat actually *understand* that pressing the lever produces food, or is this information simply encoded as non-conceptual content? The rat, through operant conditioning, may have acquired the habit of pressing a lever to acquire food. This would then be a case of procedural memory rather than semantic memory. Informational content, we know, may be stored in memory via mechanisms such as associative learning in which concept possession need not be involved. For example, spatial memories, which allow animals to navigate through their environment, need not be conceptual in nature. Thus, we should always be careful when using cognitively loaded language, such as "the dog *knows* where it buried the bone." The same applies to similar terms like 'remembers' or 'believes', especially when applied to animals. Such expressions imply conceptual understanding on the part of the subject, but this

is not necessarily the case. Certainly, for normal human adults there exists conceptual semantic memory, which is presumably the specific meaning of Dokic's [2001] comment that "...factual memory is the retention of conceptual content" (p223). But just because the content of memory is declarative this does not entail that the possessor of that memory can do the declaring. Thus, I propose to subdivide semantic memory into two types: *non-conceptual semantic memory* and *conceptual semantic memory*. I introduce these terms to cater for the possibility of *non-conceptual mental states with motivational power*, to be contrasted with propositional attitudes such as 'belief', which imply concept possession.

By definition, conceptual semantic memory requires concept possession. This fact allows it to be used as a test for self-consciousness, based on the CPH as discussed in chapter 3. Non-conceptual memory (procedural memory and non-conceptual semantic memory) does not require concepts. If we wish to use conceptual semantic memory as a test for the existence of self-consciousness, we need to be able to distinguish behaviour driven by conceptual semantic memory from that driven by non-conceptual memory. This is no easy task, and is not helped by loose talk in descriptions of animal behaviours. For example, we may talk of a rat *remembering* the route through a maze that leads to a food reward but do we really mean that the rat has factual knowledge about the way through the maze? I would say not necessarily so. Training rats to run mazes for food rewards is a paradigm example of conditioning and may show nothing more than 'learned' procedural memory. Examples of this sort expose the difficulty of devising true tests for conceptual semantic memory: there is no easy way to distinguish between procedural or non-conceptual semantic memory and conceptual semantic memory in animals based on their behaviour. As discussed in chapter 1, animals can interact with the world in a great variety of ways through non-conceptual mechanisms such as associative learning or the detection of affordances. Any test for self-consciousness based on this paradigm will need to identify behaviour demonstrating conclusively that true conceptual semantic memory is involved, implying the presence of concepts in the subject. To do this, the test must show that the behaviour could not be explained based on non-conceptual mechanisms.

The difficulty boils down to being able to distinguish between a *remembered fact*, which implies an association with concept possession, and mere *representation*. It is possible to physically record representations of the world that do not require an organism to have concepts – indeed, there are many ways to physically record representations that do not

require an organism at all. For example, a piece of music can be represented in abstract form by sheet music or in analogue form by grooves in a vinyl record or in digital form on a compact disk or MP3 player. A piece of music encoded in an organism's memory can be physically represented in a brain or nervous system without necessarily implying concept possession by the organism. This content, when retrieved, can elicit specific behaviours in the organism: consider the ringing bell that caused Pavlov's dog to salivate as an example. In other words, a *stimulus* can interact with an organism's memory (physically stored representation) to elicit a *response*. This is a case of either learned procedural memory or non-conceptual semantic memory and we should resist the temptation to say "the dog knows for a fact that a ringing bell means food will soon appear."

I do not mean to dismiss claims of concept possession in dogs or rats out of hand. Given strong evidence I may indeed come to the conclusion that these animals are concept possessing. What I suggest is that any such evidence should be held against a high standard. For example, claims have been made implying inferential reasoning in rats (e.g., Bunsey & Eichenbaum [1996]), which if correct, on my view, implies concept possession. These and other similar claims are examined in a chapter 8. My point here in relation to memory is that semantic memory as a paradigm for research into self-consciousness is unsuitable given the difficulty in distinguishing behaviour driven by *conceptual* vs. *non-conceptual* semantic memory.

**Episodic Memory**

The other type of declarative memory, *episodic memory*, provides a more fruitful approach to the investigation of self-consciousness in animals, as I show below. As mentioned earlier, not only are concepts thought to be involved in episodic memory, but it has also been *directly* associated with self-consciousness (or rather, *autonoesis*, which, as I explain below, should be distinguished from self-consciousness). More importantly, given the current thinking about the nature of episodic memory, it seems eminently suitable as an experimental paradigm. If it is true, as widely thought, that episodic memory allows the subject to plan for its own future, then this allows the design of experiments that are less prone to misinterpretation of observed behaviour.

Given the superior potential of episodic memory as a tool for researching self-consciousness, the rest of this chapter is devoted to this topic. What needs to be established are the following. First, that episodic memory does in fact exist as a phenomenon in its own right distinct from other forms such as semantic memory. Second, the nature of episodic memory; how it is thought to operate and what are its distinct characteristics. Third, given the nature of episodic memory, can it be considered a valid indicator for self-consciousness. Finally, it what ways can it be utilised as a research paradigm. Each of these are investigated in the next three sections. Then, having established its credentials for use in self-consciousness research, the results of some episodic memory experiments are examined.

## 7.3 Characteristics of Episodic Memory

Endel Tulving first coined the term episodic memory in 1972. The following passage is an excerpt of a more recent definition of episodic memory from Tulving [2005]:

> Episodic memory is a recently evolved, late developing, and early deteriorating brain/mind (neurocognitive) memory system. It is oriented to the past, more vulnerable than other memory systems to neuronal dysfunction, and probably unique to humans. It makes possible mental time travel through subjective time – past, present and future. This mental time travel allows one, as 'owner' of episodic memory ('self'), through the medium of autonoetic awareness, to remember one's own previous 'thought-about' experiences, as well as to 'think about' one's own possible future experiences. The operations of episodic memory require, but go beyond, the semantic memory system ... The essence of episodic memory lies in the conjunction of three concepts – self, autonoetic awareness, and subjective time. (p9)

For my purposes in this chapter, the key claims made in this passage are: (i) a capacity for episodic memory implies the ability for mental time travel to past and future; (ii) episodic memory will likely not be found in non-human animals; and (iii) autonoetic awareness is necessary for episodic memory. If the first claim is true, that episodic memory implies the ability for mental time travel to the future, then this enables the design of experiments to show future planning has taken place by experimental subjects, which I believe can be made immune to misinterpretation. To me, this would be a significant improvement over other experimental paradigms in which the observed behaviours are often too easily explicable by non-cognitive accounts. Experiments based on future planning tasks could then be done to

test claim (ii) regarding the likelihood of finding episodic memory in animals. If (as I believe to be the case) episodic memory is found to exist in animals, then by claim (iii) this would establish the presence of autonoesis in those animals. Autonoesis, as I discuss below, is a highly developed form of self-consciousness in that it implies not only the concept of SELF but also the concepts PAST and FUTURE. I also show that the existence of episodic memory as valid evidence of self-consciousness is consistent with CPH. This argument is based on the assertion examined later that episodic memory requires concept possession, which by CPH implies the presence of self-consciousness.

**Is Episodic Memory Distinct from Semantic Memory?**

There has been an explosion of interest in the notion of episodic memory being distinct from semantic memory since Tulving [1972] coined the term *episodic memory*, although similar distinctions had been made earlier in the 20th century [Sutton 2010]. Whereas semantic memory refers to retained *factual* knowledge, what distinguishes episodic memory is the capacity to represent a specific *episode* from one's life. The knowledge of such an episode is itself factual knowledge and so in a sense is also semantic knowledge, but there is thought to be something special about episodic memory in that it is knowingly grounded in personal experience, while the origin of purely semantic knowledge is not necessarily retained[56]. There is a distinctiveness to episodic memory during retrieval – namely, that episodic memory recall is *experienced* as being episodic: when one remembers an experience from one's past, it feels different to when one is simply recalling a piece of factual knowledge.

There is not universal agreement that a true distinction between episodic memory and semantic memory exists. Martin [2001] is sceptical about a fundamental divide between memories of facts in general and of one's own past in particular. Although Martin acknowledges an experiential component to episodic memory, he believes the general condition on memory is that of preserving either past *knowledge* or past *apprehension*; these two being closely related in that the latter is the episodic counterpart to the former. Martin's main point is somewhat subtle and perhaps best expressed by the following quotation:

[56] The origin of a fact might be retained if it is associated with a particular learning episode, but this case should be considered as two separate (though linked) memories: the memory for the fact itself and the episodic memory of the learning episode. More often factual knowledge (such as that Paris is the capital of France) is retained without concomitant knowledge of how, when or where the memory was initially encoded.

"...episodic memory is not now apprehension of a past episode, but rather the retention of a past apprehension of that episode" (p267). In other words, he doubts that episodic memory recall amounts to a 'reliving' of the original experience. Cockburn [2001] is sympathetic to these ideas and further suggests that we can sometimes remember *what* happened without knowing *that* we remember. He also adds the threat of infinite regress: "If there *must* be a way in which he knows that he knows, then presumably there must also be a way in which he knows that he knows that he knows" (p397). The essence of these views is that only the *content* is different between semantic memory and episodic memory (one being *facts* and the other *personally apprehended events*); the mechanisms of encoding and retrieval are the same. If so, this weakening of the distinction between semantic memory and episodic memory threatens the proposal that episodic memory presents a potentially superior paradigm for research into self-consciousness. This is because if episodic memory cannot be distinguished from semantic memory then paradigms based on episodic memory will suffer the same vulnerabilities to misinterpretation. However there exists empirical evidence to suggest a marked difference between semantic and episodic memory.

Many authors agree that recollecting an episodic memory is phenomenologically different from recalling a factual memory. There is a sense in which recollecting an episodic memory is a reliving of a personally experienced event [Perner 2001]. Thus, there is 'something it is like' to remember a witnessed event [Hoerl 2001], which critically grounds the acquired knowledge in direct experience. The sense of *oneself existing in the past* signals to the subject that the mental representation is in fact a memory of an actually experienced event – not a fantasy, dream, plan or imagined image [Conway 2001b; Campbell 2001]. It appears that episodic memory inherently involves an element that signifies its status as an episodic memory. Some authors are convinced that during memory formation episodic memory signalling information is encoded along with the factual information[57]. By contrast, a piece of semantic knowledge may have *no link with the self in the past* whatsoever. Conceivably, an organism (a human infant, say) might be able to recall a piece of factual knowledge (from semantic memory) without knowing how or when it gained that knowledge and without involving a sense of self. Recollecting an episodic memory, however, involves knowledge of the *self* existing at the *time* of acquisition.

[57] The main contenders for the nature of this episodic memory specific coding are: *contextual information* [McCormack 2001]; *phenomenological records* [Conway 2001A, 2001B]; and *emotional content* [Dere, Zlomunica, Huston & De Souza Silva 2008].

A common conception of episodic memory is that it is an extension of semantic memory and metarepresentational in nature, in the sense that associated with the *facts* of an episode is a metarepresentational memory that the remembered episode comes directly from one's own past experience [Dokic 2001; Hoerl 2001; Zentall 2005][58]. Tulving suggests that episodic memory evolved 'out of' semantic memory and presents a table of empirically derived properties of semantic and episodic memory [Tulving 2005, table 1.1]. The table sets out a list of properties that are common to both semantic and episodic memory and a list of properties unique to episodic memory. This structure supports Tulving's assertion that episodic remembering implies semantic knowing but not vice versa, and that episodic memory has some special characteristics over and above semantic memory. Further support is provided by empirical evidence that semantic encoding processes are involved during episodic memory storage [Robertson and Köhler 2007].

The view that there is a fundamental difference between semantic memory and episodic memory is further supported by the fact that amnesics have a gross deficit in the ability to store new episodic memories in the absence of any semantic memory deficit. This is true not only of those who become amnesic later in life (i.e., after acquiring a store of factual knowledge) but also for those born amnesic, who are still able to acquire normal intelligence, language and semantic memory [Baddeley 2001].

Given the evidence, it seems likely that episodic memory is indeed distinct from and significantly different in nature to semantic memory. The defining characteristics of episodic memory over and above those of semantic memory can be summarised as follows. Firstly, there is a particular phenomenology associated with episodic memory when recalling an episode from one's past. This can be further broken down to the sense of both *self* involvement and the association with it being a *past* event. Secondly, episodic memory implies the existence of relatively sophisticated cognitive capacities such as metarepresentation and concept possession.

[58] This aspect of episodic memory, if true, is particularly relevant to my conception of self-consciousness as in chapter 1 I listed metacognition as a hallmark of self-consciousness.

**Summary**

The idea that episodic memory offers a more apt paradigm for research into self-consciousness relies on it being different to semantic memory in a significant way. Namely, that associated with episodic memory is a particular phenomenology that identifies it as such. The counterview is that there is no essential difference between semantic and episodic memory, at least in the way they are encoded and retrieved. However, there are good reasons to believe in a significant difference between semantic and episodic memory. Firstly, the evidence based on amnesics implies separate mechanisms are involved. Secondly it is undeniable that a phenomenological difference exists. Furthermore, the different nature of the content of semantic and episodic memory highlights an important consideration: the content of episodic memory always involves the *self* and the *past* whereas semantic memory content need not involve either. Thus, it is possible to conceive of semantic memory existing in the *absence* of concepts of self and past, but this is not the case for episodic memory. It is this fact that allows us to exploit episodic memory as a research tool for self-consciousness, as further discussed below.

## 7.4 Episodic Memory and Self-Consciousness

My aim in this section is to examine the connection between episodic memory and self-consciousness. Many authors already accept that episodic memory implies the existence of self-consciousness in a subject. In this section, not only do I validate that claim, but I also uncover a little-recognised differentiation between self-consciousness (at least as I use it) and the term usually applied in the context of episodic memory: *autonoesis*. I show that autonoesis is a more developed form of self-consciousness, in that autonoesis involves a sense of one's own existence through time while self-consciousness only requires a sense of one's own existence in the present.

### Autonoesis versus Self-Consciousness

According to Tulving [2005]: "The essence of episodic memory lies in the conjunction of three concepts – self, autonoetic awareness, and subjective time" (p9). The term 'autonoetic

awareness' is usually interpreted as 'self-knowing' and this could be interpreted as equivalent to the type of self-consciousness I described earlier as the target of my interest. However this is not quite the conception Tulving has in mind for autonoesis. Tulving [1985] gives the following description: "Autonoetic (self-knowing) consciousness is the name given to the kind of consciousness that mediates an individual's awareness of his or her existence and identity in subjective time extending from the personal past through the present to the personal future" (p1). Tulving later describes an autonoetic subject as a "projectable, or time-travelling or remembering" self [Tulving 2005]. Hence, Tulving describes autonoesis specifically in terms of an *awareness of subjective time*[59] and a *capacity to recall episodic memories*. I examine the claim that episodic memory implies a sense of subjective time (including past, present *and future*) in the following section; here I aim to establish not only that self-consciousness and autonoesis are conceptually distinct notions but also that they exist separately (in fact, that autonoesis is a more developed form of self-consciousness).

Tulving [2005] cites a case study in which he describes an amnesic subject ('KC') as 'conscious' and 'self-reflectively conscious', but yet *not* 'autonoetically conscious' and *not* having 'normal self-awareness' (where the latter is not defined). KC was 30 years old when he suffered brain damage as the result of an accident. It is clear from the descriptions of KC's memory deficiencies that he lacks certain cognitive capacities – specifically, he is unable to recall episodes of his own life from either before or after the accident although he has no apparent damage to his semantic recall or semantic memory encoding. KC's self-awareness may not be 'normal' as compared to other ordinary human adults, but it does not appear that he has any deficiency in his sense of himself as existing as a psychological subject, which is how I earlier described my conception of self-consciousness. Tulving & Kim [2009] have described beings such as amnesics and infants that do not have the capability of autonoetic memory as perfectly capable of learning from their past and 'knowing' the past in a fully conscious way, though 'noetically' and not 'autonoetically'. According to Tulving & Kim, what beings without autonoesis cannot do is 'remember', or consciously re-experience their past life events as previously experienced.

Since Tulving talks of amnesics like KC being *self-reflectively conscious* but not autonoetically conscious, Tulving explicitly allows for the type of fundamental self-

[59] Several authors now talk of 'subjective time' and I follow their usage here. It should be noted, though - as pointed out to me by Prof. John Sutton - that really what is meant by this is a *subjective understanding of objective time*.

consciousness I have previously described, in which an individual at least understands that he exists as a psychological subject (without *necessarily* possessing a sense of subjective time). Tulving himself [personal communication 2011] has confirmed this interpretation, remarking that his conception of autonoesis "...in no way implies that all forms of self-awareness must include awareness of personal times other than the present." Tulving's remark regarding KC's deficiency in self-awareness appears to be aimed more at describing a lack of *personal identity*, more about which I discuss below. It may be an unfortunate misusage, then, to describe autonoesis simply as 'self-knowing' as that would seem to preclude self-knowledge in the absence of a concept of subjective time. As the foregoing discussion on KC confirms, 'noesis' is also a kind of self-knowing, in the sense of self-reflection, even though a sense of subjective time is apparently absent.

Autonoesis, then, is tied to a sense of time and is best viewed as a particular type or level of self-consciousness that enables a concept of the self as existing at times other than the present. Some other authors have also interpreted autonoesis specifically in terms of its association with subjective time. For example, Conway [2001a] describes recollective experience as "...a form of autonoetic consciousness in which the current self becomes aware of itself in the past..." (p241); and Zentall [2005] describes autonoetic conscious awareness as simply 'remembering'. But if so, its usage within Tulving's [2005] definition of episodic memory might be somewhat redundant, as the concept of autonoesis is implied from the other two concepts (*self* and *subjective time*) that make up the tripartite constituency of episodic memory in the passage quoted at the start of section 7.3. Some other authors, though, do not explicitly link autonoesis with a sense of time. For example: "An animal, which is able to project itself in the position of another animal, i.e., knows what another animal knows in order to manipulate the other animal's knowledge, can be said to have a kind of self- or autonoetic awareness/consciousness" [Dere et al. 2008, p161]. It is quite possible that some authors do not see it as important whether the temporal aspect is explicitly involved in autonoesis or not and may simply use the terms 'autonoesis' and 'self-consciousness' interchangeably. However, to my mind, if autonoesis is to be seen as tied to subjective time, it is a different and probably more developed form of self-consciousness. In the more fundamental conception, I envisage self-consciousness as involving a self-concept in which the individual understands itself to exist as a psychological subject, with no temporal aspect implied. Self-consciousness does not necessarily imply understanding of one's existence at times other than the present. As such, I will continue to differentiate between the terms

'autonoesis' and 'self-consciousness', with the understanding that autonoesis means self-consciousness with the further capacity of a sense of subjective time.

**Does Episodic Memory Imply Self-Consciousness?**

By Tulving's very definition of episodic memory, autonoesis is implied and there is wide acceptance of this notion (e.g., Suddendorf & Corballis [1997]; Conway [2001a]; McCormack & Hoerl [2001]; Dokic [2001]; Gardiner [2001]; Zentall [2005]; Easton & Eacott [2008]; Dere et al [2008]). Indeed, it seems reasonable that if episodic memory involves a reaching back to a past personal experience then it requires a self-concept to perform this feat. In the first place there is the current knowledge of oneself as the experiencer of an event. Furthermore, Dokic [2001] argues that the subject must have (also) been self-conscious at the time of the remembered event, given the implausibility of 'seeing' through one's past life if it only consisted of "...first-order mental states and episodes, neither unified nor bound together by any reflection..." (p231). Then, both episodic memory encoding *and* retrieval requires the existence of self-consciousness in a subject. By all accounts, then, episodic memory does imply autonoesis. As earlier discussed, autonoesis is a more developed form of self-consciousness as I use the term. Therefore, a demonstration of episodic memory should be taken as evidence for the existence of autonoesis and hence also self-consciousness.

**Episodic Memory and the Concept Possession Hypothesis**

In the foregoing discussions I concluded that episodic memory is indeed a sufficient test for the existence of self-consciousness. This conclusion is consistent with CPH, which states that concept possession is sufficient for self-consciousness. It would be inconsistent if episodic memory were possible without concept possession, however, as we have already discussed, episodic memory requires not only a concept of the self but also of subjective time. Dokic [2001], for example, writes: "...factual memory [i.e. conceptual semantic memory] is the retention of conceptual content..." (p223). Dokic also agrees with Tulving [2005] that episodic memory can be seen as an *extension* of semantic memory: "...an episodic memory is just a factual memory associated with the further, metarepresentational memory that the

former memory comes directly from the subject's past experience" (p219). This implies that episodic memory, too, must involve conceptual content. Furthermore, as previously discussed, just in knowing that the episode comes from one's own past implies a concept of *self* in the sense of being the subject of the experience.

Despite the foregoing, there are instances where authors *appear* to write of episodic memory as if it was *non*-conceptual in nature. Dokic [2001] himself, for example, also writes: "...factual memory is the retention of conceptual content, whereas episodic memory carries non-conceptual information about the past" (p223). At first glance this might seem to contradict the earlier analysis that episodic memory does involve concepts. However, what Dokic means is that the *phenomenology* of episodic memory is non-conceptual; he is not here referring to the *content*, as the following passage makes clear: "Episodic memory is *non-inferential* in the intuitive sense that it immediately presents itself as episodic. I do not need to infer its episodic character from features of its content" (p223). In other words, it is the phenomenology (*episodic character*) of episodic memory that needs no inference. I believe this is the same notion that Clayton & Russell [2009] have in their proposal for a more minimalist view of animal episodic memory, despite initially describing it as focusing "... on the non-conceptual content of a re-experienced situation" (p2330). Later, they put it this way: "...whatever it is that gives episodic memory its phenomenology is non-conceptual and, given this...animals need not have limited episodic experience in virtue of their limited conceptual apparatus; though their re-experiencing will inherit the conceptual character of their experiencing" (p2336). So, while it makes sense to talk of the *phenomenology* of episodic memory as non-conceptual, knowledge that a memory is episodic *does* require concept possession – in particular the concepts of the *self* (in that the memory was personally experienced) and *past*. Accordingly, it is safe to say that the view that episodic memory indicates self-consciousness is consistent with the commitments of CPH.

**Levels of Consciousness: Tulving and Savanah**

Tulving [1985] describes a correspondence between types of memory systems and levels of consciousness. For him, procedural memory (*nondeclarative*) corresponds to *anoetic* consciousness, where the latter is described as "...temporally and spatially bound to the current situation. Organisms possessing only anoetic consciousness are conscious in the sense

that they are capable of perceptually registering, internally representing, and behaviourally responding to aspects of the present environment, both external and internal" (p3). Next comes semantic memory (*declarative*), which corresponds to *noetic* consciousness, where "Noetic consciousness allows an organism to be aware of, and to cognitively operate on, objects and events, and relations among objects and events, in the absence of these objects and events. The organism can flexibly act upon such symbolic knowledge of the world" (p3). Finally, there is episodic memory (*also declarative*), which corresponds to *autonoetic* consciousness.

The distinction between declarative and nondeclarative memory is reminiscent of the 'levels' analysis of phylogenetic development I presented in chapter 3. There, I described 'level 2' organisms as those that are conscious (that is, perceptually conscious) but not self-conscious. Level 3 organisms are self-conscious and are capable of conceptualisation. My 'level 2' corresponds to Tulving's organisms with *nondeclarative* memory while my 'level 3' corresponds to his organisms with *declarative* memory. Suddendorf & Corballis [2008] point out that a kind of 'prospection' is available via nondeclarative memory, but only in the sense that a current stimulus can trigger a response that is influenced by past experience. This maps onto my description of level 2 organisms as they are similarly bound by a (non-conceptual) stimulus-response paradigm. Organisms with declarative memory, like my 'level 3' organisms, are not stimulus-bound.

There are, of course, some pertinent differences in these models. While my model stops at level 3, in which I see organisms possessing self-consciousness, Tulving's splits this level into two. My level 3 organisms (those with declarative memory) for Tulving are divided into those with only (conceptual) semantic memory (noetic) and those with episodic memory (autonoetic), where it is assumed that the latter group also possess semantic memory, as episodic memory is considered by Tulving to represent an extension of semantic memory [Tulving 2005]. Organisms with episodic memory would for me constitute a 'level 4' category. However, as my concern is to investigate the existence of self-consciousness and not necessarily autonoesis (which as stated earlier is seen as self-consciousness with the addition of the more advanced cognitive capacity of a concept of subjective time), I have had no need for a level 4 in my schema. However, I introduce it here for the sake of comparison with Tulving's model: whereas my level 3 corresponds to Tulving's *noetic* organisms, my level 4 corresponds to Tulving's *autonoetic* organisms (see table 7.2). Note that in table 7.2

'semantic memory' refers to *conceptual* semantic memory. *Non-conceptual* semantic memory would be classed within the anoetic level. As Tulving has confirmed, although there is nothing in the concept of semantic memory that would exclude knowledge and awareness of oneself in the present, this "...does NOT imply that creatures that have semantic memory are aware of themselves... These are not all-or-none concepts" [Tulving, personal communication 2011].

Table 7.2: Correspondence between Tulving and Savanah taxonomies

| **Tulving** | | **Savanah** | |
|---|---|---|---|
| | | Level 1 | No consciousness present |
| Anoetic | Non-declarative memory | Level 2 | Consciousness present but no self-consciousness |
| Noetic | Semantic memory (declarative) | Level 3 | Self-consciousness present (*but no sense of persistence through time*) |
| Autonoetic | Episodic memory (declarative) | *'Level 4'* | *Sense of persistence through time* |

**Personal Identity**

It is worth comparing episodic memory with mirror self-recognition (MSR) as tests for self-consciousness. In human infants, MSR emerges at around 18 months [Amsterdam 1972], while episodic memory appears much later (as discussed in section 7.7 below). In chapter 5 I argued that MSR provides sufficient evidence for a self-concept. Episodic memory, however, does this and much more. Episodic memory not only provides evidence for the concept of self, but for even more sophisticated concepts such as 'persistence of self through time' (to be understood, at this point, as a concept of having existed in the *past*; in section 7.5 I examine the notion that this concept includes the self in the *future*). For some this latter concept might be considered necessary for a 'proper' sense of self (and so implicit in a concept of self), but I would rather make a distinction here between a fundamental self-consciousness as I have previously described, and the more sophisticated self-knowledge that comes with the concept of persistence of self through time. Episodic memory, in demonstrating a concept of persistence of self through time, might be indicative of a sense of *personal identity*. This is because personal identity is often associated with a subject's ability to construct a life history,

based on autobiographical memories [Neisser 1988; Bruner 1997; Bruner & Kalmar 1998; Nelson 2005]. But a concept of oneself in terms of personal identity is not necessary for a self-concept worthy of being called self-consciousness as I described it earlier. A similar position is taken on this issue by Suddendorf & Corballis [1997], who consider it necessary "...to dissociate a self-concept in the present from personal identity (or self-concept through time), the former being a prerequisite for mental time travel and the latter the consequence of mental time travel." Thus, MSR may be considered sufficient evidence for a fundamental self-consciousness in the sense I have been using that term (i.e., with no requirement for a sense of persistence through time), while episodic memory might be taken as evidence for a more sophisticated form of self-consciousness possibly approaching the level of personal identity. This sets the bar rather high and may be why there is resistance to the idea of episodic memory in animals (e.g., Tulving [2005]). I return to the issue of personal identity later.

**Summary**

To summarise, episodic memory appears to indicate an understanding of one's existence in the past and as this capacity implies (minimally) a self-concept, episodic memory can be taken as evidence of self-consciousness. Furthermore, episodic memory also indicates *autonoesis*, which can be seen as self-consciousness with the added cognitive capacity of a sense of subjective time (at least in regard to a sense of the *past*). Thus, if experiments can be devised to show episodic memory in animals then this can be taken as evidence of self-consciousness and autonoesis. In the next section I examine Tulving's claim that episodic memory implies a sense of subjective time (including past, present *and future*). This claim is critical to the later discussions because, as I show below, experiments based solely on *past* events cannot conclusively prove the existence of episodic memory, but those based on expectations of *future* events can.

## 7.5 Episodic Memory as 'Mental Time Travel'

According to Tulving [2005], episodic memory "Makes possible mental time travel in both temporal directions, past and future" (p11). *Prima facie*, it seems reasonable that a concept of

subjective time means an understanding of both past and future. If such an understanding means an ability to mentally project oneself into another time then it should not matter if that time is in the past or the future. However, it first needs to be established that episodic memory implies a sense of subjective time. An alternative conception would be that episodic memory requires only the ability to 'relive' a past experience, with no 'projection' as such taking place. A 'reaching back' into one's own past may be subserved by the existence of memory traces – neuro-cortical imprints of experienced events – that allow the rememberer to mentally replay the event. By contrast a projection into the future to 'pre-experience' [Atance & O'Neill 2001] a previously unexperienced event is an ability not necessarily connected with memory traces. Furthermore, the supposition that there is a unique phenomenology to episodic memory *recall* does not add weight to the speculation that future-oriented thinking works by the same mechanism.

Despite these objections, there is indeed evidence that episodic memory is simply one aspect of a general concept of subjective time. Tulving [1985] concluded from case studies of amnesic patients, which had a deficiency in episodic memory, that "...the lack of conscious awareness of personal time encompasses both the past and the future" (p5). A correlation between episodic memory and episodic future thinking is also revealed from studies comparing adults with autism spectrum disorder (ASD), in which the ASD subjects were deficient in their abilities for both [Lind & Bowler 2010]. The empirically established fact [e.g., Friedman 2001; Dere et al. 2008] that recalled episodic memories are not always strictly veridical but tend to show a significant element of construction (or reconstruction) lends further support. For example, Suddendorf & Corballis [2008] think that mentally constructing past episodes and mentally constructing future ones may be "...two sides of the same coin" (p31) and this flexibility in recollection may be a reflection of future function. Conway [2001b] considers the ability to manipulate memories in order to examine different possible outcomes as a basis of future thinking.

Event sequencing is an aspect of episodic memory that can also provide some support to the idea that episodic memory implies a sense of subjective time. Although it is accepted that time sequencing can often be done erroneously [Freidman 2001; Easton & Eacott 2008; McCormack & Hoerl 2001], it is enough that it can be done at all, as this indicates that "One grasps that there are systematic relationships between different events..." [McCormack &

Hoerl 2001]. This grasp of event sequencing is indicative of a grasp of subjective time, and may underlie an ability to sequence imagined events into the future [Dere et al. 2008].

There is evidence from brain imaging experiments of significant overlap in brain regions active during episodic memory recall and prospection [Buckner & Carroll 2006; Szpunar et al. 2007]. These regions are also active when a subject imagines the perspective of another (so-called *theory of mind* activity, as discussed in chapter 6). That the same brain regions are active during these activities may indicate that they are all part of a more general type of self-projection. Projection of oneself into the mind of another to gain their perspective may involve essentially the same process as projecting oneself into one's own mind at different times or situations. This could cover projection into one's own past to reconstruct experienced events as well as projection into an imagined future or hypothetical situation. After reviewing relevant studies, Buckner & Carroll concluded that "Thinking about the future, episodic remembering, conceiving the perspective of others (theory of mind) and navigation engage [a core brain network], which suggests that they share similar reliance on internal modes of cognition and on brain systems that enable perception of alternative vantage points" (p55). Brain imaging techniques are limited in both granularity of resolution as well as explanatory power, and Buckner & Carroll admit that "...we are far from understanding the specific relevant anatomy for prospection and related forms of self-projection..." (p53). Indeed, other experiments highlight the *differences* in neural signatures between episodic memory and future thinking [Weiler, Sucha & Daum 2010]. Thus, it would be unwise to rely on this evidence alone. Nevertheless, at the very least we can say that the evidence from brain imaging is not inconsistent with the idea of a link between episodic memory and future thinking.

The arguments and evidence above provide a good case that episodic memory implies the capacity for both past and future mental time travel. As such, we should take a demonstration of mental time travel to past *or* future as evidence for episodic memory and therefore for the existence of self-consciousness in a subject. In the next section, some experimental paradigms are considered.

## 7.6 Experimental Paradigms

I hope to have shown in the previous sections that episodic memory provides opportunities for definitive tests of self-consciousness. Although we still need to be cautious in our interpretation of animal behaviours, it should be possible to devise tests that show animals have a sense of subjective time. Specific tests can be engineered based on novel environments/conditions, the results of which might be explicable only by assuming a sense of subjective time in the subjects. As we saw earlier, a concept of subjective time applies not only to the capacity for 'reaching back in time' to relive past experiences, but also to *reaching forward to the future*. As I discuss below, the future-oriented test paradigms are the most promising.

There are a variety of experimental paradigms that have been applied to animals and/or human infants to test for the existence of episodic memory capability. In reviewing these, attention is paid to not only the theoretical validity of the test, but also to the possible interpretations of results. The main paradigms to be examined are *recollection of specific past events; the capacity to time-sequence events; 'What-When-Where' memory; the ability to keep track of time;* and *planning for the future.*

### Event Recollection and Time Sequencing of Events

Determining the ability of an individual to recollect a past experience is easy when the subject is a human adult; not so for animals. Determinations of this type generally rely on verbal reporting, a method unavailable for animal subjects. Even in human infants, a verbal report might not be definitive evidence of episodic memory. McCormack [2001] refers to reports in the literature of very young children verbally recalling specific past events and internally representing them as in some way 'past', but McCormack doubts their ability to quarantine off images as memories. An alternative explanation is that the event has been retained as a semantic memory rather than true episodic memory. Similarly, in animals, experiments may circumvent the requirement for verbal reports by observations of behaviour indicative of past event recollection, but may fall prey to the same type of objection, i.e. that episodic memory is not involved but rather some form of nondeclarative memory. For example, that an animal returns to the location of cached food need not necessarily indicate a

memory of the caching event itself, particularly if such caching behaviour is innate to the species.

Dere et al. [2008], citing evidence of chimpanzees recalling events after 16 hours of others hiding objects, assert that this "...indicates that apes can give an unprompted report of a personal experience, suggesting that they have the ability of conscious recollection..." (p162). Since chimpanzees are not naturally food-caching animals the observed behaviour seems impressive, but I am sceptical of interpreting this as evidence of episodic memory as the information regarding the location of the object could have be encoded as semantic information.

Zentall [2005] cautiously proposes that animals may not be 'stuck in time' and presents evidence that animals (e.g., pigeons) can make choices based on the relative recency of two events. The behaviour to be reported was whether the pigeon had recently pecked or had refrained from pecking a response key. The pigeons were trained to choose a red comparison stimulus if they had recently pecked an initial stimulus and to choose a green comparison stimulus if they had recently refrained from pecking. This was analogous to training the pigeons to answer the question, "What did you just do?" where the appropriate answer would be, "I just pecked" if they chose red or "I just refrained from pecking" if they chose green. Following this training a task was introduced using different coloured keys, with the subsequent 'unexpected' inclusion of the red/green keys to elicit an answer to the question "what did you just do?" The pigeons successfully answered correctly, indicating memory for the past sequence of their own recent actions. However, Zentall remarks that the repeated training trials required suggest that a simpler, rule-learning account may be responsible for the accurate performance of this task. Furthermore, since the retention period was quite short, Dere et al. [2008] suggest that the correct performance might be mediated by working memory rather than episodic memory.

The problems with experimental paradigms that test for recollected or time-sequenced events remain as significant barriers. Without verbal testimony, it will always be difficult to establish episodic memory over (non-conceptual) semantic memory in these cases. This issue is compounded by rule-learning as a possible obfuscating factor in some cases and, if the periods in question are short, by the possibility of working memory being the main operative medium.

**What-Where-When (WWW)**

'What-Where-When' (WWW) memory as an indication for the existence of episodic memory was first suggested by Tulving [1972] and was picked up as a challenge by Clayton & Dickinson [1998] in a series of experiments on scrub jays. The suggestion was that episodic memory is the integration of the WWW of an event. Clayton et al. showed that scrub jays are able to keep track of the length of time since food items were stored by allowing them to recover either perishable worms or non-perishable peanuts that they had previously cached at specific locations. The jays had first been trained to recognise the 'shelf-life' of the worms prior to the experiment. Jays searched for their preferred food (worms) when allowed to recover them shortly after caching. However, they avoided searching for worms after a longer interval during which the worms had decayed, and instead searched for the peanuts. This experiment demonstrated that jays are able to keep track of time ('When'), as well as to integrate the 'What' (type of food) and the 'Where' (cache location) of the memory. Clayton & Dickinson gave this as evidence that scrub jays have 'episodic-like' memory. They and their colleagues furthermore suggest that this capacity is unlikely to be unique to humans and food-caching birds and is probably important to survival in a number of species [Clayton, Griffiths, Emery & Dickinson 2001].

Testing for WWW memory using similar techniques as that used for scrub jays have been applied to other animals with varying levels of perceived success. Dere et al. [2008] report that an integrated WWW memory for unique experiences has not yet been shown in non-human primates and in rats WWW is suggestive only in experiments following extensive training. Bird et al. [2003] report radial maze experiments on food-hoarding rats (Rattus norvegicus) based on a similar technique to the scrub jays, where cheese was used as the preferred yet faster-degrading food. They report memory for what and where but not for when. The same paradigm was also used by Hampton et al. [2005] to test for WWW memory in rhesus monkeys and again the subjects demonstrated memory for what and where but not when.

Dere et al. [2008] report on an experiment purported to demonstrate WWW memory in mice (and repeated later in rats) that did not rely on food caching and did not rely on training or food rewards. The experiment instead relied on a 'one-trial' object-recognition paradigm. In essence, 'recognition' of familiar and novel objects was assessed based on the time spent exploring them, including after spatial and temporal displacements of the test objects: "...the

mice spent more time exploring two old familiar objects relative to two recent familiar objects, reflecting memory for what and when, and concomitantly directed more exploration at a spatially displaced old familiar object, reflecting memory for what and where...” (p170). It is doubtful, however, that such marginal behavioural differences can relate in a meaningful way to cognitive capacities of the supposed sophistication required for episodic memory. In any case, and more importantly, it is highly questionable as to whether WWW is a suitable indication of episodic memory, as discussed next.

Several authors have objected to WWW as valid criteria for episodic memory on various grounds. In the first place, there are counter-examples, such as knowing the WWW of one’s own birth or a historical event without the possibility of episodic memory being involved [Hampton, Hampstead & Murray 2005: Zentall 2005, 2008; Suddendorf & Corballis 2008]. Even more damaging is the objection that the ‘When’ knowledge is not necessary for episodic memory [Easton & Eacott 2008] and, for that matter, neither is the ‘Where’ [Campbell 2001]. Subjects can recall episodes from their past, accompanied by the phenomenology of a recollection, without being able to confidently state either the temporal or spatial location of the event. It these cases it seems that only the ‘What’ criterion is necessary for episodic memory – for example, I can recall distinctly having watched a particular movie without consciously remembering exactly when or where it happened. Of course, the ‘What’ criterion alone cannot depict an instance of episodic memory, as this is effectively reducing episodic memory to the content of semantic memory. What is still required is the associated phenomenology of episodic memory to definitively characterise it as such. Moreover, it is arguably only the phenomenology that is required. For example, I am surely not alone in experiencing a situation in which I think to myself something like “I can distinctly remember doing something really important yesterday afternoon at 1:00pm just after finishing lunch, but I can’t remember what it was or where I did it.” The foregoing example includes only the ‘When’ component of WWW along with the phenomenology, yet certainly counts as an instance of episodic memory. Thus, I very much doubt that WWW can be trusted as criteria for establishing the existence of episodic memory.

### Keeping Track of Time

Despite the foregoing objections, it could be argued that even if the scrub jay experiments do not directly demonstrate episodic memory based on WWW, they at least do demonstrate a concept of subjective time in that the jays are able to keep track of the passage of time since an event. However, we need to consider alternative explanations that do not rely on a concept of subjective time. It is possible for instance, that animals have a variety of 'hard-wired' mechanisms that allow them to keep track of time, which, due to an anthropomorphic bias, we might erroneously attribute to human-like episodic memory [Vonk & Povinelli 2006]. This objection is made more poignant by the fact that the scrub jay task tapped into innate food caching and recovery behaviour rather than a novel task [Easton & Eacott 2008]. An example of such a hard-wired mechanism might be a kind of internal 'stop-watch' [McCormack 2001] akin to circadian rhythms. Although this type of explanation seems implausible given the long period tracked by the jays (several days), Hampton, Hampstead & Murray [2005] point out that what constitutes a 'short' or 'long' delay interval differs widely among species and the scrub jay behaviour may be explicable simply by working memory. Roberts [2002, 2006] suggested an even simpler mechanism: that a bird might associate a weak memory of cached worms with worm decay.

Given the alternative explanations for animals' ability to keep track of time, this does not seem suitable as a test for episodic memory. As such, the scrub jay experiments described (so far) are inadequate as evidence for episodic memory (but see the segment on scrub jays in section 7.7).

### Planning

As discussed earlier, there is good reason to believe that episodic memory is simply one aspect of a general *concept of subjective time*, interpreted as a capacity for mental time travel to the future as well as the past. Mental time travel to the future means, in this context, being able to project oneself into the future and take the perspective of one's future self. This possibility allows for ways in which mental time travel can be detected. We can discount behaviours that can be explained by species-specific genetic pre-dispositions (such as food gathering for long term storage) or cued by environmental triggers (such as day/night cycles). However, future-oriented behaviours of the right kind (i.e. planning in novel situations)

should be taken as sufficient evidence of a sense of subjective time. Based on this idea, Tulving [2005] devised a protocol to test for the existence of episodic memory in animals. The test, which he dubbed the *spoon test*, requires an organism to plan for its own future circumstances. Doing so shows it is capable of mental time travel, thereby showing a sense of subjective time and hence the capacity for episodic memory. This in turn, as previously discussed, provides sufficient evidence for not just self-consciousness but for the more cognitively sophisticated *autonoesis*.

The spoon test is so-called because it alludes to an old Estonian folk tale in which a child must plan ahead for attendance at a party by bringing her own spoon for pudding. Put simply, the intention is to demonstrate mental time travel abilities by showing that the subject is able to plan for its own future. However, Tulving imposed several constraints in order to eliminate the possibility of misinterpreting animal behaviour as evidence of planning. The first of these is that the behaviour must not be instigated by a present need or be governed by current physiological states, but rather should satisfy a need that will be realised at a future time. This constraint addresses what Suddendorf & Corballis [1997] call the Bischof-Köhler hypothesis: that animals can only act toward the satisfaction of a future need if cued by their present motivational state. Tulving's second constraint is that the behaviour not be triggered by specific environmental stimuli present in the original learning situation, since this makes it difficult to rule out the possibility of the behaviour being governed by associative learning. Lastly, Tulving's spoon test requires that the future intention be directed at something that happens in a different place to where the preparatory action happens, once again in order to minimise the influence of present situational cues.

### 7.7 Experiments on Future-Orientation

In this section I examine the results of research into planning in animals and human infants. Despite the fact that planning represents a superior paradigm for investigation into self-consciousness, the majority of experiments are still inconclusive. Nevertheless, I argue that at least one experiment, involving scrub jays, shows good evidence of mental time travel.

**Arthropod Navigation**

Cheng [2012] describes arthropod navigation in terms of place-finding servomechanisms, control systems that have ongoing control by appropriate stimuli. Here is a description of the servomechanism by Cheng:

> The system can be characterized as having a standard, a specification of a target place, to which it "aims." The current state of affairs (input data) is compared with the standard, and the difference is an error. The system is designed to move so as to reduce the error.

Nothing in this description indicates conscious or self-conscious behaviour. It appears to describe programmatic behaviour, presumably underpinned by genetic 'hard-wiring' and overlaid with the capacity to take in and process current information from the environment. However, it presents a good opportunity to discuss examples of behaviour that *seem* like planning but most likely are not, such as can be seen in the complex behaviour of ants and foraging bees.

A particularly interesting case described by Cheng [2012] is on 'route planning' by jumping spiders of the genus *Portia*. In novel (laboratory) settings these spiders can navigate a route to a perceived prey (a dead spider of a different species or a fake lure) even when the route includes a section forcing the spider to lose visual contact with the prey. In experiments conducted by Tarsitano & Jackson [1997] the spider had to 'remember' which of two simulated tree trunks led to the prey and it did so at significantly above chance levels. In subsequent experiments, Tarsitano & Andrew [1999] were able to examine the scanning movements of the spider's eyes during the route 'planning' phase in which the spider had to choose between two apparent routes, one of which was broken (i.e. non-viable). Initially the spider scanned the gap in the broken route a lot but then spent more time scanning the unbroken route before embarking on the journey.

It is tempting to interpret the *Portia* behaviour as a case of planning. It seems as though the spider is consciously checking out its options and working out a path to the prey: it notices the gap in one of the potential paths, realises that path is not viable and so discounts it. It then concentrates on scanning the viable path to memorise the route it needs to take. Irrespective of the actual goings-on in the spider's control system, I must admit to being impressed with this behaviour. Nevertheless, explanations are available that do not rely on the conclusion that planning is taking place, in the sense being discussed here (i.e. in which a sense of

subjective time is required). In the course of evolution this genus has become genetically pre-disposed to 'recognise' viable routes to prey based on the topology of the landscape – even if the landscape is artificial. Topological features that prevent successful navigation to prey (such as gaps or barriers) would soon be deselected by natural selection. Another way to express it is to say that the spider is able to detect the affordances in its environment that represent a successful pathway to its prey. The pathway is then encoded within the spider's nervous system, perhaps as a set of 'instructions' that the spider's control system later executes. The details here are not known, although Cheng [2012] suggests that it might involve a set of 'beacons' (effectively, landmarks) that allow the full route to be broken down into sub-routes. An appropriate set of beacons along the route would act as 'secondary objectives' and allow a line of sight to the next beacon in each section of the route, explaining how the spiders are able to navigate all the way to the prey even when losing direct sight of it. Furthermore, Tarsitano & Andrew [1999] note that the spiders do not solve the route 'all at once' but change secondary objectives when the immediate secondary objective cannot not be reached, thereby giving the spider a way to complete routes in a complex environment without having to pay a high cost in cognitive processing in the central nervous system.

The foregoing discussion shows how behaviour that appears to involve complex planning may in fact be nothing of the sort. Thus I maintain that we should remain sceptical whether the subjects are relatively simple organisms such as arthropods or phylogenetically highly developed such as primates.

**Rats**

Cook, Brown & Riley [1985] describe an experiment with rats where 'prospection' (i.e. forward planning) seems to be taking place. A 12-arm radial maze was used in this experiment in which the rats were required to visit each arm to retrieve a food item. At staggered intervals of visitations (i.e. after having visited 2, 4, 6, 8 or 10 arms) the rats would be removed for a 15 minute interval and replaced. The question being asked was, do the rats visit the unvisited arms based on which arms they had already visited (retrospective memory), or based on which arms they were yet to visit (prospective cognition). Results showed that the rats performed the worst at this task when they were removed at the halfway point (after 6 arms had been visited). The conclusion consistent with these results is that the

rats used whichever method engendered the least effort at the time. Thus, in the first half it is easier to remember which arms had been visited and then avoid them for subsequent searches, but in the second half it is easier to remember which arms had not yet been visited and then search in those ones. The latter technique is thought to represent a case of prospection in that the retained information is 'which arms I *will* visit next' rather than 'which arms I have visited'.

This behaviour probably should not count as an instance of planning. In the first place it does not conform to Tulving's criterion that the behaviour should not be instigated by a present need (the rats were hungry). But even if we discount all of Tulving's spoon test criteria there are other reasons to reject this behaviour as an example of planning. If the delay interval was increased to 60 minutes no evidence of prospection was obtained. Whatever was encoded in the rats' brains in regard to unvisited arms may have utilised working memory. It has been claimed that animals are 'stuck in time' [Roberts 2002] – meaning that they remain 'in the present'. However, clearly, there must be some leeway in this: there must be a short span of time both forward and backward that constitutes 'being in the present' and this might vary amongst different animals. An immediate motivational state may be cause for an action seconds or minutes later, but we should not consider this a case of planning; the action caused is still a present action and not really a 'future' one, even if it is executed some time after the initiating stimulus.

**Primates**

Evidence of future planning has been reported in bonobos, orangutans (Mulcahy & Call, 2006) and chimpanzees (Dufour & Sterck, 2008). Mulcahy and Call used a tool transportation paradigm to determine if bonobos and orangutans can plan for a future need. The animals were trained to use a tool to retrieve a food reward in the 'test room'. They were then ushered into a waiting room from where they observed experimenters removing all tools from the test room before being allowed back in after an hour. To solve the task, the animals had to choose the right tool (from a variety) to bring with them when leaving the test room and then bring those tools back with them when re-entering an hour later. Bonobos and orangutans (five of each) performed successfully at well above chance levels. In a separate experiment one bonobo and one orangutan were tested on the same procedure with the delay

period increased to 14 hours (overnight). Both failed on the first trial, but then succeeded at well above chance on the following trials.

There are doubts as to whether the Mulcahy & Call [2006] results conform to Tulving's spoon test criteria. Suddendorf & Corballis [2007] question whether the tool transportation action was instigated by a present need. As the present need was not controlled for, the Bischof-Köhler hypothesis (that non-human animals are unable to differentiate future states from present ones) cannot definitively be discounted.

Dufour & Sterck [2008] performed future-orientation experiments on chimpanzees. Their first set of experiments were designed to discover if planning took place in a social setting, given that chimpanzees are highly social animals. The experiments employed an exchange paradigm, in which chimpanzees learned that certain objects (straws, branches, etc.) could be exchanged with a human partner for a treat, and were required to plan ahead by selecting and bringing appropriate objects to a test room. Subjects failed to show future-oriented behaviour using this paradigm. Dufour & Sterck speculated that the reason for failure may be due to a progressive loss of motivation for the task or because it was complicated and required more complex cognitive capacities. The task involved planning based on a calculation of which tool to use, where it should be used and when. The social aspect possibly also contributed to failure: primate social structures and interactions are not straightforward and in this experiment further complications may have been introduced as non-chimpanzees were involved in the social interaction.

Following the failure of the exchange paradigm experiments, Dufour & Sterck [2008] attempted to replicate the Mulcahy & Call [2006] experiments with chimpanzee subjects. In these experiments the subjects had to select an appropriate a tool (hook) to be transported one hour later into a test room to retrieve a bottle of juice. Three out of seven subjects succeeded in the task and Dufour & Sterck reported their results as successful. Water was freely available during the tool selection period to ensure that thirst was not a factor when planning was taking place. However, as Dufour & Sterck readily admit, the possibility cannot be ruled out that the sight of the hook at the tool collection period may by association have driven the chimpanzees into collecting them out of current desire for juice, so the Bischof-Köhler objection remains unresolved.

Overall, the experiments involving great apes is suggestive of some future-oriented abilities, but given the uncertainties around the Bischof-Köhler hypothesis it may be too early to declare a true understanding of subjective time by these animals.

**Scrub Jays**

Arguments have been raised based on observations of scrub jays that they have an ability to plan for the future (e.g., de Kort, Dickinson & Clayton [2005]; Emery & Clayton [2001]; Raby, Alexis, Dickinson & Clayton [2007]; Correia, Dickinson & Clayton [2007]). Although in most cases, as usual, plausible alternative explanations exist I argue that in at least one experiment [Correia, Dickinson & Clayton 2007] the scrub jays were shown to dissociate their future need from their current motivational state and thus essentially passed the spoon test. The birds were fed on two occasions with an opportunity to cache food during the intervening gap. Usually, scrub jays pre-fed with one type of food (pine nuts or kibbles) will preferentially cache the *other* type, due to an effect known as specific satiety. This conforms to the aforementioned Bischof-Köhler hypothesis, in that the future need is cued by the current motivational state (i.e., being sated with one type of food gives the other food type a higher incentive value causing the birds to cache that other type). However, in the experiment, some jays (the 'Same' group) were trained to expect the same type of food in a secondary feeding session while others (the 'Different' group) were trained to expect the other food type (see figure 7.2). The Same group are fed with pine nuts on both day 1 and day 2 while the Different group are fed pine nuts on day 1 but kibbles on day 2. According to the Bischof-Köhler hypothesis, both groups should cache the other food type from what they were fed in the first feeding session. That is, since both groups were sated with pine nuts on day 1, they should both cache kibbles on day 2 due to the specific satiety effect. However, the Different group preferentially cached the *same* type of food (i.e., pine nuts) as they were fed in the first feeding session, overriding the specific satiety effect in favour of providing for greater food variety in the second feeding session. In other words, since the Different group knew they would get kibbles on day 2 they cached pine nuts (even though they had just been sated on pine nuts). Thus, this group of scrub jays planned for a future need by overriding a current need and thus disproved the Bischof-Köhler hypothesis.

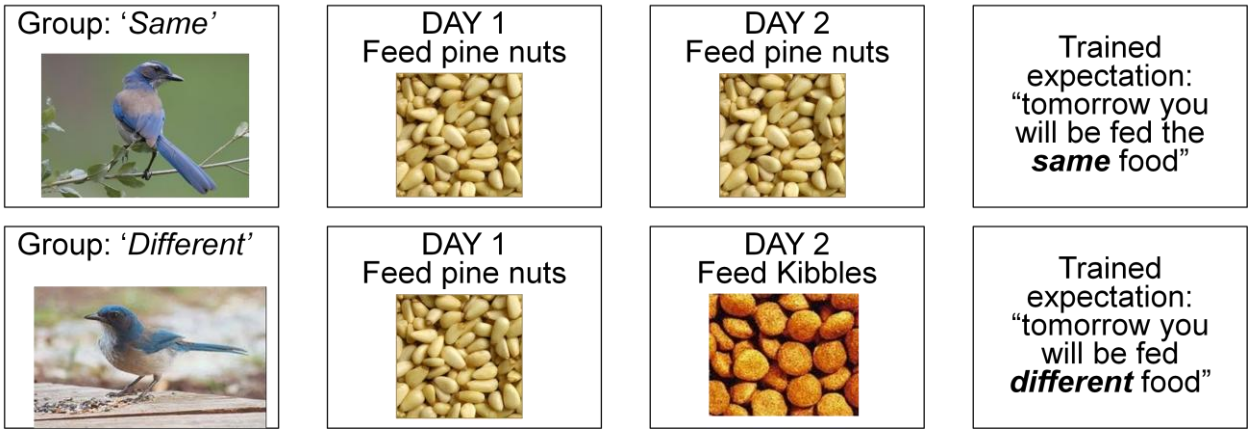



Figure 7.2: Schematic depiction of the training phase of the episodic memory experiment on scrub jays by Correia, Dickinson & Clayton [2007]. The Bischof-Köhler hypothesis predicts that *both* groups of subjects will cache kibbles due to the specific-satiety effect, since both are sated with pine nuts on day 1. However, the Different group preferentially cached pine nuts because they knew they would get kibbles on day 2.

The scrub jay experimenters remain, to my mind, overly cautious about these results, as exemplified by the following passage:

> In the absence of language, there is no knowing whether this reflects episodic future thinking, in which the bird is projecting itself into tomorrow morning's situation, or semantic future thinking, in which the jay takes prospective action, but without personal mental time travel into the future. However, in either case it shows that these birds must have the capability to plan for a future motivational state over a timescale stretching at least into tomorrow [Raby et al. 2007, p920]

To me and some other commentators [e.g., Feenders & Smulders 2008], the scrub jay results provide strong evidence for mental time travel capacity in the test subjects. The experiment satisfies all of Tulving's spoon test criteria except the third requirement: a relocation of the future activity from the location of the preparatory activity. Of course, this criterion could not possibly be met within the confines of this experimental paradigm, in which caching food at a specific location is central. Whether failure to meet this constraint is sufficient grounds to reject the results as successful is open for debate. If an experiment relies upon a recollection during the future-occurring activity then the constraint makes sense, as one would want to preclude the possibility that the recollection was cued by situational cues. In the scrub jay case, this would mean that it is significant to the experimental result that the food retrieval

activity is cued by associations made with the cache location. In other words, this would be important if it was the *recollection* that was the key behaviour in the experiment. But in this experiment it is not the recollection activity that is most important; rather, it is the decision made at the food caching event. The decision by the birds to cache a particular type of food is what indicates the ability to plan for the future, irrespective of the situational parameters. It might very well be that, by setting the location requirements, Tulving has set the bar unnecessarily high for the spoon test. In this experiment the present and future motivational states of birds in the test and control group were successfully controlled to demonstrate that it was primarily the *future motivational states* that drove the planning activity of the subjects. This should be taken as strong evidence of planning by scrub jays. The (perhaps extraordinary) conclusion we are forced to draw from these results is that there is a strong case for ascribing self-consciousness and autonoesis to scrub jays.

**Human Infants**

In humans it is taken for granted that self-consciousness and autonoesis will emerge during the course of normal child development. The open questions here are, perhaps, which of these cognitive developments occurs earlier and at what age they occur. In chapter 5 I argued that mirror self-recognition provides sufficient evidence of self-consciousness, and this capability develops around 18 months. Autonoesis, which adds the further cognitive capacity of a sense of subjective time, appears to develop sometime between 3-5 years [Atance & O'Neill 2005]. Suddendorf & Busby [2005] conducted an experiment to more narrowly determine the age group in which planning occurs in children. The age groups they used were 3 years (29-45 months), 4 years (51-57 months) and 5 years (63-69 months), with a sample size of 16 in each group. The task was to plan which toys to bring to an otherwise empty room that only had a puzzle board with no puzzle pieces. The 3 year olds in both experiment and control groups performed at chance levels: 50% choosing the puzzle pieces. Both older groups performed above chance for the experimental groups, indicating that planning capability emerges around 4 years of age. A similar experiment by Russell, Alexis & Clayton [2010] yielded much the same results. This is consistent with the claim made earlier that autonoesis is a more developed form of self-consciousness.

## 7.8 Conclusion

I argued, based on the Concept Possession Hypothesis, that declarative memory (conceptual semantic memory and episodic memory) is linked to self-consciousness, but that conceptual semantic memory is unsuitable as a research paradigm due to the apparently impossible task of discriminating between conceptual semantic memory and non-conceptual semantic memory based solely on observations of animal behaviour. Episodic memory is distinct from semantic memory and implies a sense of subjective time including the ability to project oneself into the future. Episodic memory therefore implies the possession of not only the self-concept (and hence self-consciousness) but also of the concept of the self existing in the past and future (and hence autonoesis). Autonoesis is commonly defined simply as 'self-knowing' but it importantly comprises a sense of subjective time and is therefore quite distinct from self-consciousness. Autonoesis can be seen as a more developed form of self-consciousness.

Demonstrating episodic memory in a subject by showing the subject has a sense of subjective time should be taken as evidence of autonoesis (and therefore also self-consciousness). Experiments intended to do this using past events (such as event recollection, time sequencing, What-Where-When memory, and time tracking) all suffer the same pitfall as for conceptual semantic memory: observed behaviour can be accounted for with alternative explanations. Tulving [2005], however, devised a paradigm (the 'spoon test') based on future planning that is immune to this problem. Apart from failing on one of Tulving's criteria (which I consider non-critical) an experiment on scrub jays conformed to the spoon test and showed that the subjects were able to plan for a future (not current) need. Thus, the scrub jays demonstrated a sense of subjective time and hence are autonoetic and self-conscious.

To claim that scrub jays are self-conscious may seem extraordinary. However, other members of the corvid family (magpies) have passed the mirror self-recognition test [Prior, Schwarz & Güntürkün 2008]. Although scrub jays have not yet been tested for MSR, I predict that they too will pass; there is already tantalising indications of MSR in scrub jays [Dally et al. 2010]. Furthermore, there is other evidence of corvid intelligence (in which concept possession is evident); for example, the behaviour of the crow in Aesop's famous fable of the *Crow and the Pitcher* [60] has been confirmed in rooks [Bird & Emery 2009]. It is even more extraordinary to claim that scrub jays are autonoetic, as this would place them at a similar

[60] In this fable, a thirsty crow chances upon a pitcher partially filled with water but out of reach of its beak. To get to the water the crow drops in pebbles until the water level rises enough for it to take a drink.

cognitive level to at least 3-5 year old infants. Nevertheless, primate and corvid self-consciousness might represent a case of convergent evolution and although much further research is needed to confirm this conclusion, there is no reason to discount it on present evidence.

Even so, we should be cautious about the consequences of this result. Earlier, I mentioned that episodic memory, in showing a sense of subjective time, might be indicative of a sense of personal identity. However, there may be something of a gap between a basic capacity for episodic memory, such as that demonstrated by scrub jays, and a more sophisticated capacity for autobiographical memory. Although some authors use 'episodic memory' and 'autobiographical memory' more-or-less synonymously [e.g., Dere et al. 2008], others [e.g., Conway 2001a, 2001a, 2008], make a clear distinction between them and assume episodic memory feeds into autobiographical memory. Simply having a concept of one's own persistence through time does not guarantee that one has the ability to weave a personal history from which one derives a sense of personal identity. Despite their extraordinary capacity for episodic memory, we are not quite ready yet, I submit, to confer personhood upon scrub jays.

# Chapter 8: Rats and Rationality

## 8.1 Introduction

In the earlier chapters of part 2 I examined various research paradigms to address the question: "is this research paradigm a valid way to determine the existence of self-consciousness in animals?" The focus on those chapters was on specific research paradigms (mirror self-recognition, imitation and episodic memory), and applied across several animal species. In this chapter I take the reverse approach and concentrate on a particular species rather than a particular research paradigm. Here I address the question: "are rats, as a species, self-conscious organisms?" Rats are extensively tested against a wide variety of paradigms, so they make a good choice for this analysis. To my knowledge no researchers are directly making claims of *self-consciousness* in rats, but many are making claims of *rationality.* The question of whether rats are rational is a significant one and worthy of close scrutiny independent of its connection with self-consciousness. However, if it is true then I would be further committed to the view that rats as a species are self-conscious. This is a consequence of the Concept Possession Hypothesis (CPH) presented in chapter 3, which claims that concept possession is sufficient evidence of self-consciousness. In chapter 4 I discussed various ways in which concept possession might be determined and one of these was rationality, since inferential reasoning requires concept possession. I show, however, that the evidence for rationality in rats is far from conclusive and research results can be explained using non-conceptual accounts.

Many researchers are claiming cognitive capacities in rats that can be interpreted (explicitly or implicitly) as rationality. Of course, the literature on rat experimentation is vast, but I have chosen examples of five key research paradigms to examine in this regard: spatial navigation; metacognition, transitive inference; causal reasoning; and goal orientation. Each of these are analysed in turn. Note that the issue here is not simply a case of researchers using 'loose talk' or using 'rationality' in one of the weaker senses described in chapter 4. Nor are they using 'rationality' metaphorically: they are ascribing *actual rationality* – in the sense of PP-rationality, implying the ability for conceptual, inferential thinking – to their rat subjects. For example, Eichenbaum [2000] claims that rats display *problem-solving* skills in certain spatial navigation tasks. According to Eichenbaum, the processing of declarative memory in rats

"...includes complex cognitive rules and concepts..." (p47). Foote & Crystal [2007] assert that rats make *adaptive decisions* about future behaviour. That is, rats make a *reasoned* choice based on supposed knowledge of their own ability to pass a test: "Presumably, an animal that knows that it does not know the answer to a test question will decline to take the test" (p551). Bunsey & Eichenbaum [1996] suspect that rats are capable of representations that can be expressed indirectly and *inferentially* due to the 'transitive inference' effect. In transitive inference, the type of reasoning thought to be involved can be expressed as 'A is better than B; B is better than C; *therefore* A is better than C' (although, again, there is no suggestion that the rats are capable of thinking those thoughts in propositional terms). Blaisdell, Sawa, Leising & Waldman [2006] claim the rats in their experiments are capable of *inferential*, causal reasoning: "Rats made causal inferences in a basic task" (p1020). Finally, in discussing the apparent goal orientation of rats, Dickinson [1985] claims that rats have knowledge in *propositional-like* form. However, in each case, after in-depth analysis, I conclude that the observed results can be adequately explained by associative theories or other non-conceptual accounts and therefore an ascription of rationality to the rats is unwarranted. That does not mean I *completely* rule out the possibility of self-awareness in rats. It is possible that some further evidence from future experiments will be sufficiently convincing. Indeed, in earlier chapters I have already countenanced the likelihood of self-awareness in other non-primates (such as elephants, dolphins and certain corvid species), so I remain open to this possibility in other species. Nevertheless, we must keep the bar high and currently the results of rat experiments can be explained without invoking rationality.

## 8.2 Spatial Navigation

I begin with spatial navigation because although at first glance this would not seem to be a facility that need rely on any conceptualisation or reasoning power, nevertheless some authors still use language implying rationality with respect to it. Virtually all animals navigate through their environment and this would seem to be inbuilt; probably species-specific, but nevertheless selected for by evolution. The hippocampus in rats has long been known to play a significant role in spatial navigation [Moser, Kropff & Moser 2008] and Eichenbaum [2000] has suggested that "...the hippocampus may be required for new *problem solving* in familiar environments" (p45, emphasis added). Eichenbaum describes experiments using the 'Morris water maze' in which a tank of water has a single column standing just

below the water surface as a refuge for rats to escape from the water: "...when rats with hippocampal damage that have successfully learned to locate the escape platform from a single start position are tested from new start positions, they fail to readily locate the platform. In contrast, normal animals swim directly to the escape locus on each new probe trial" [Eichenbaum 2000, p45]. I do not see the need to presume that this shows problem solving – at least, not in the strong sense required for a case of inferential reasoning. At most this shows that hippocampus must be involved in memory formation and/or retrieval in certain spatial tasks. Cheng [1986] has shown that the rat's metric frame (or 'cognitive map' – see the discussion in section 3.2) specifies locations primarily by their geometric relations to environmental shape. Presumably the hippocampus is involved in the formation and maintenance of the metric frame, but it is by no means obvious that in using the metric frame for navigation, as in the water maze experiment, that inferential reasoning is needed to 'solve' the problem of locating the escape platform. The metric frame maps the localised world geometry in the rat's hippocampus but this is information encoding and does not necessarily represent *knowledge* (i.e. conceptual understanding).

As described in chapter 7 on Episodic Memory, Cook, Brown & Riley [1985] conducted an experiment into the rats' spatial navigation of a 12-arm radial maze. The researchers were investigating prospection (future thinking) in the rats, but the results could be construed as indicating reasoning by the rats. The reasoning seemed to be along the lines of 'at the start of my explorations I will keep track of where I have already been but toward the end I will keep track of what is left to explore, as that will minimise my cognitive load'. Of course I do not expect the rat to have thought in those propositional terms, but the question is whether the content of the rat's mental state could be of that form. Once again, my answer is that there is no reason to suspect the rat's thoughts to be so sophisticated. Foraging is normal species-specific behaviour for rats and even though radial mazes are not natural environments for rats, it is quite probable that their foraging practices have been optimised for efficiency by natural selection. Thus, this might be considered a case of E-rationality but not PP-rationality.

### 8.3 Metacognition

Tests for metacognition in animals have relied on a 'bail-out' paradigm (e.g., Smith [2005]; Hampton [2005]; Foote & Crystal [2007]). For example, Smith [2005] investigated the ability

for rhesus macaques to judge their own confidence at visual density discrimination tasks. After a suitable period of training, the monkeys are 'asked' in a series of trials to judge whether a box of pixels on a computer screen is dense or sparse according to a set threshold, and are rewarded for a correct answer. For a lesser but guaranteed reward the subject also has the option to decline the test, in effect to answer with a response of 'uncertain'. Judgements of this type are deemed by the experimenters to be acts of metacognition because the monkey is making a decision not on the density of the box but (apparently) on knowledge of its own ability to succeed at the task. In Hampton's [2005] version monkeys were required to remember an image that had been presented to them on a previous computer screen by matching to sample from an array of possibilities on a subsequent screen. The delay in presentation of the choice array was variable, and the monkeys were allowed to opt out of the test after the delay but prior to the display of the choice array. Opting out (for a guaranteed but lesser reward) was deemed to imply a metacognitive thought in the monkey equivalent to 'I cannot remember the image'. The monkeys tended to opt out more often as the length of the delay increased (so that remembering became more difficult).

It can be argued that metacognition construed in this way (the ability to monitor one's own ability to discriminate between stimuli) represents a case of PP-rationality. The subject may be having thoughts expressible (by us) as "I cannot tell whether the pattern is dense or sparse so I might choose the wrong option. Therefore I will bail-out and get the lesser reward instead." (Again, there is no suggestion or requirement that the subject itself can express the thought linguistically.) To argue against PP-rationality it is necessary to provide possible alternative explanations for the subjects' behaviour that do not assume the need for such reasoning to have occurred. I show below that the results of experiments using the bail-out paradigm can be explained by (first-order) associative learning without assuming any metacognitive abilities. Smith [2005] dismisses associative learning as an explanation, but his arguments rely on assuming that the subjects' associations to reward are dipolar – e.g., based on options such as 'Dense/Sparse' or 'Same/Different', etc. I argue below that multiple associations are available to the subjects. Further, Smith notes that his earlier attempts to show metacognition in rats failed, and suggests that the reason for this failure is that these tasks "...seem to be psychologically structured in some way that leaves rats out...but leaves humans, monkeys and dolphins in" [Smith 2005, p261]. However, Foote & Crystal [2007] subsequently conducted bail-out experiments on rats and replicated the monkey results, invalidating Smith's argument. Of course, it could be just that rats need to be added to the list

of animals capable of metacognition. However, I use Foote & Crystal's results on rats as an example for my argument that no assumption of metacognition is necessary. The essence of my argument is that the bail-out option itself represents a third stimulus that becomes associated with the lesser reward. This argument is then compared to the behavioural economic model (BEM) presented by Jozefowiez, Staddon & Cerutti [2009], in which the results are explained in terms of a mathematical pay-off maximisation model.

**Metacognition in Rats**

Foote & Crystal's [2007] rats were trained using tones of eight different durations ranging from 2 to 8 seconds. The four shorter tones were associated with a 'Left' lever for a reward while the four longer ones were associated with the 'Right' lever for the reward[61]. The two tones in the middle of the range were the most difficult to classify as long or short. In the test phase, after presentation of a tone the rat had the option of entering one of two apertures. In aperture 1 ('take-the-test') were the two levers, Left and Right, in which pressing the correct lever produced a reward of six food pellets. In aperture 2 ('decline-the-test') the rat obtained a guaranteed but lesser reward of three food pellets (see figure 8.1). For tones near each extreme (long or short) rats were more likely to enter aperture 1 (presumably indicating confidence of their own ability to make a correct lever selection). For the harder to discriminate middle tones the rats were more likely to enter aperture 2 (presumably indicating lower confidence and so taking the option for the lower guaranteed reward).

[61] My descriptions of experiments in this chapter are simplifications. The experimenters implemented all the necessary controls to avoid unwanted biases, such as reversing the levers for some rats, etc.

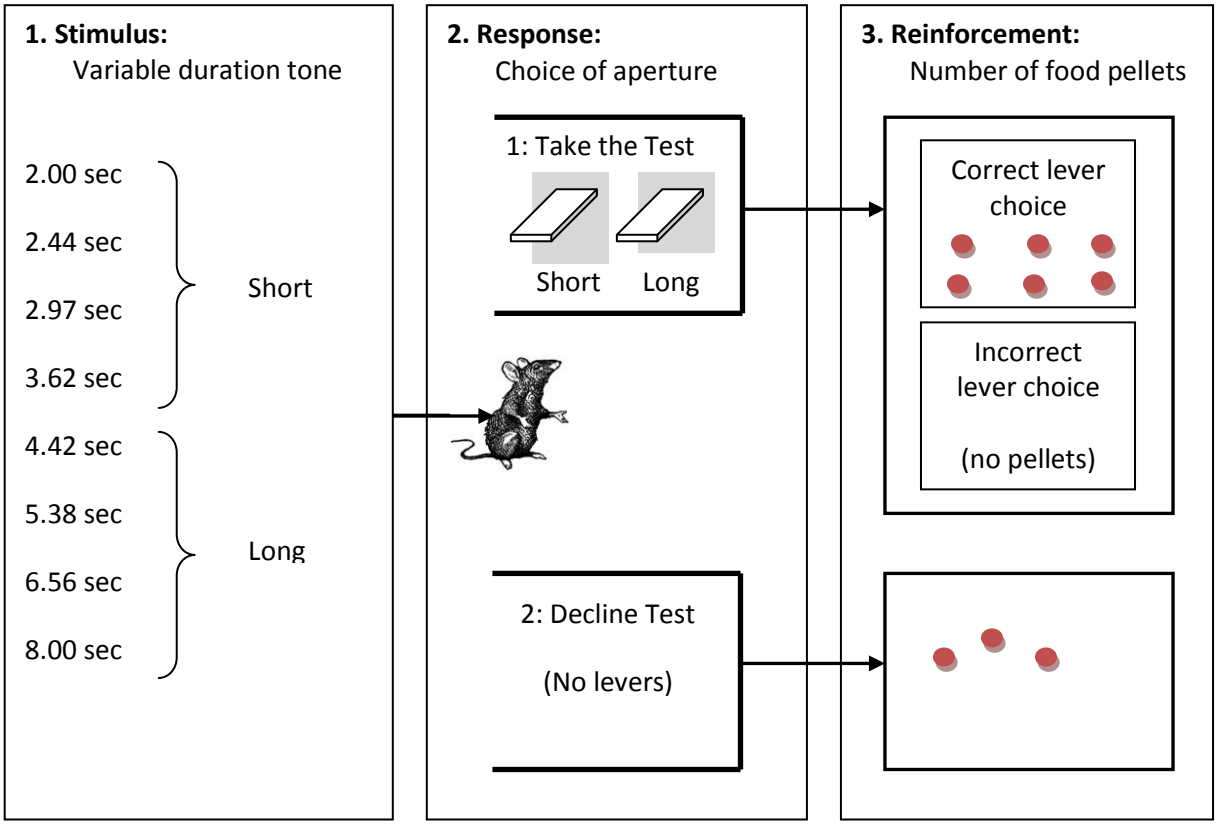



Figure 8.1: Schematic depiction of rat metacognition experiment

Strength of association is a function of reward value and reward reliability. A stimulus-response that is rewarded every time will result in a stronger association than for one that is rewarded only half the time. Thus, a tone duration at either extreme (long or short) which the subject can easily discriminate reliably results in a high value reward, resulting in a strong association with the required response (correct lever selection). However, the middle duration tones are harder to discriminate between long or short and result in more errors being made. In those cases they are just as likely to get a reward as to *not* get a reward no matter which lever they press, therefore it is unlikely those durations will form associations with *either* lever. So, during the training phase, while strong associations will be formed for both the very long and very short tones, there is a third possibility that emerges in which *no* association is formed for the middle tones.

Although the intention was to train the rats to associate rewards with either long or short tones, there is no reason why they might not actually form *multiple* associations, including an (unintended) association for the middle tones. Training does not end with the initial training phase, but continues into the test phases when the reward structure is different. During the test phase, the rat is able to form new associations involving aperture 2, which *always* yields

a (lesser) reward. These latter associations, being of lesser reward, are not likely to override the previously established stronger associations for long/short (provided, as is the case, those associations remain well rewarded). However, these new associations can be established for the intermediate duration tones because those ones did not form associations previously. So, along with the strong associations of the long and short tones with their paired levers and food rewards, a new association is free to be established between the middle tones and aperture 2. Instead of a choice of either 'long' or 'short' plus a metacognitive understanding of 'I can't tell', the choices trained may have actually been 'long', 'short' or 'intermediate', with no metacognition involved. Including 'intermediate' as a possible discrimination, table 8.1 lists out all the associations possible according to the available behaviour options.

Table 8.1: The full range of possible associations in the Foote & Crystal [2007] experiment

| **Tone duration** | **Behaviour** | | | **Reward level** | **Trained effect** |
|---|---|---|---|---|---|
| | **ID** | **Aperture** | **Lever** | | |
| Long | L1 | 1 | Right | High | Strong association |
| Long | L2 | 1 | Left | Zero | Negative association |
| Long | L3 | 2 | N/A | Small | Weak association |
| Intermediate | I1 | 1 | Right | (Indeterminate) | No effect |
| Intermediate | I2 | 1 | Left | (Indeterminate) | No effect |
| Intermediate | I3 | 2 | N/A | Small | Weak association |
| Short | S1 | 1 | Right | Zero | Negative association |
| Short | S2 | 1 | Left | High | Strong association |
| Short | S3 | 2 | N/A | Small | Weak association |

Table 8.1 shows that there are three behaviour options for each of the three available discriminations: *long, short* or *intermediate.* The final column in the table indicates the strength of incentive for each of the available actions following discrimination. For a discrimination of 'long', the strongest incentive is for behaviour L1 (enter aperture 1 and select the *right* lever). Similarly, for a discrimination of 'short' behaviour S2 is preferred (enter aperture 1 and select the *left* lever). For a discrimination of 'intermediate', I3, though weakly incentivised compared to L1 and S2, is nevertheless more favourable than the available alternatives. This associative model correctly predicts the observed behaviour of the rats. As such, there is no need to assume metacognitive abilities in rats to account for these experimental results.

**Carruthers on Metacognition**

Carruthers [2008] has also argued against a metacognitive explanation of the bail-out experiments, raising three main points with which the first two I concur and the last I challenge. Firstly, Carruthers presents an argument not too dissimilar to the one presented above, though his example was Smith's [2005] dense/sparse discrimination experiment. In regard to the subject's reaction to presentation of the intermediate stimulus (i.e. one that is difficult to distinguish between dense/sparse), Carruthers remarks: "...the competition between pressing or not pressing D and pressing or not pressing S is in a four-way tie, whereas there exists an unopposed motive to press 'don't know', so as to avoid a time out. Hence that, accordingly, is what the animal does" (p65). This analysis is not dissimilar to my notion of an association made with the bail-out option. Secondly, Carruthers notes that "...undergoing a feeling of uncertainty needn't mean being aware that one is feeling uncertain, as such (which would be a metacognitive state)" (p68). Thus, the feeling of uncertainty can be associated with the bail-out option without the need to assume metacognition.

Lastly, Carruthers explains apparent metacognitive behaviour in bail-out experiments in terms of *first-order* beliefs and desires that have varying strengths. The crux of the argument is that first-order beliefs/desires can elicit behaviours matching the observed experimental results but without any intervention of self-reflective thought. The fact that the beliefs/desires are first-order allows Carruthers to dismiss metacognition as an explanation. But this line of reasoning will not do for me; although metacognition can be discounted on this argument the ascription of beliefs to the subjects in my view implies that they are capable of possessing concepts, which, by CPH implies self-awareness. It is not necessary, however, to describe the motivational forces at work as beliefs. For example, Carruthers ascribes the following strong belief to the subject: "...if the pattern is dense and 'D' is pressed, food will result." Rather than construing this mental content as a belief, it should be seen simply as an association between the elements, viz. <dense pattern>; <D button>; <food delivery>. E-rationality (i.e. natural efficiency-maximising mechanisms) can fully explain why this association is preferentially selected as a motivational factor. There is no need to assume that the subject can conceptualise any of the elements in the association in order for it to motivate behaviour.

**Behavioural Economic Model (BEM)**

Jozefowiez, Staddon & Cerutti [2009] claim that the Foote & Crystal results can be explained in terms of pay-off maximisation (the Behavioural Economic Model, or BEM). This mathematical model is based on only two assumptions: (i) when confronted with a stimulus a subject emits the behaviour associated with the higher pay-off; and (ii) the perception of the stimulus is noisy. The 'noisiness' is effected with a Gaussian distribution function to simulate the spread of possible stimuli in the variable tone duration experiment. On adjusting certain parameters, such as the level of 'risk aversion' assumed for the subject and the width of the Gaussian probability curve, the model does a fair job of predicting the behaviour observed in the rat experiments. Although the quantitative fit is not exact, Jozefowiez et al. argue that metacognition is not needed to explain the observed results in rats:

> ...BEM, which lacks any metacognitive ability — only basic discrimination processes — satisfies the two generally accepted criteria for metacognition: that the probability of picking the uncertain response increases with the difficulty of the task...and that the subject is more accurate on free-choice trials then on forced-choice trials (p33)

In chapter 4 I also argued that mathematical models implemented on computers were examples of (non-cognitive) programmatic behaviour. This example provides further support for the view that the bail-out experiments are insufficient evidence of metacognition.

## 8.4 Transitive Inference

The phenomenon of 'second-order' Pavlovian conditioning has been known for several decades [Savastano & Miller 1998] and can account for what we might call 'transitive associations'. That is, in which two different pairings containing one common element elicits an association between the two unpaired elements (e.g., (i) A paired with B; and (ii) B paired with C; causes (iii) A associated with C). That such cross-associations can be trained in animals is no cause (on their own) to suggest the presence of inferential reasoning (e.g., "...there seems to be no sense in which implicit grasp of any *logical* principle is involved" [Bermũdez 2003, p113]). However, it has been suggested that rats are capable of transitive *inference*, in which the specific relationships between the unpaired elements are being inferred. A familiar example of transitive inference in mathematics is in the usage of the

'greater than' symbol (>). If it is known that A > B and that B > C then we can infer that A > C. Analogues of this transitive inference paradigm have been designed with the aim of demonstrating transitive inference by animals. Elements A and B are paired such that A is rewarded but not B, thereby training the subject to prefer A over B. Then B is paired with C such that B is rewarded but not C, so that B is preferred to C. Then the subjects are presented with A and C to determine which is preferred. In this paradigm it could be argued that subjects will prefer A over C since C has *never* been rewarded, and this explanation does not rely on transitive inference [Allen 2006]. For this reason it is common practice in more recent transitive inference experiments to use five elements (A, B, C, D and E) and to compare B and D so as to ensure that the compared elements have had equal measures of reinforcement and to avoid any other first/last in series effects. One such experiment is examined in detail in the next section.

**The Five-Element Transitive Inference Paradigm**

Eichenbaum [2000] describes an experiment in which rats were trained with four odour pairs to prefer one odour over the other in each pair (see figure 8.2). Labelling the odours as A, B, C, D and E, and using the '>' symbol to represent preference, the trained preferences can be expressed as A > B; B > C; C > D; D > E. In probe trials following the training, rats showed a preference of B over D (B > D) despite the fact that both were equally rewarded in their pairings. The strength of Eichenbaum's claim that this amounts to a case of transitive inference (as opposed to mere transitive association) rests on two key factors. Firstly, B and D are not merely associated; rather, the inherent *relationship* between B and D appears to have been inferred based on the hierarchy established during training. Secondly, while normal rats were able to perform this inference, a second group of rats with induced hippocampal damage were not (even though this latter group were able to acquire the individual pairings at normal rate).

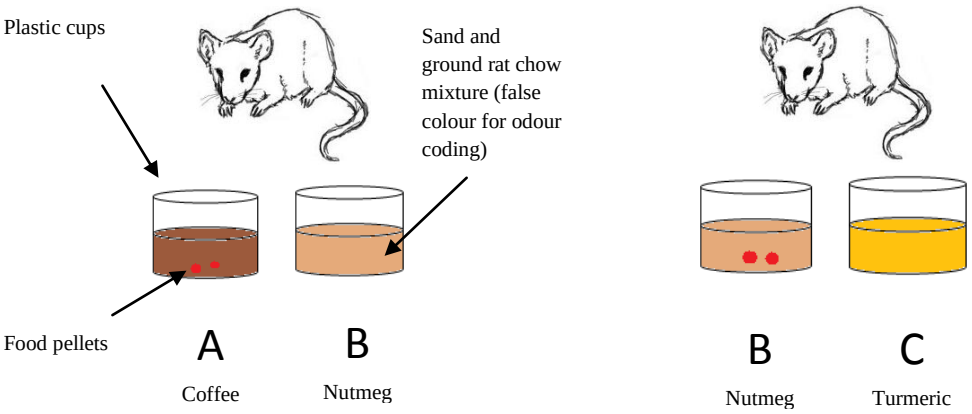



Figure 8.2: Schematic depiction of the 'five-element' transitive inference experiment (for simplicity only three elements are shown). Rats uncover a reward in cup A but not in cup B and therefore come to prefer odour A (coffee) over odour B (nutmeg). Later they are similarly trained to prefer odour B (nutmeg) over odour C (turmeric).

Given the importance of the hippocampus in humans for higher cognitive capacities such as declarative memory, it is tempting to ascribe the same or at least similar functionality for the hippocampus in animals (at least in those animals not too far removed from humans such as other mammals). Bunsey & Eichenbaum [1996], who conducted the transitivity experiments on rats and compared their results to analogous studies on humans, write: "...in both humans and animals, stimuli can be associated independently of hippocampal function but the establishment of representations that can be expressed indirectly and inferentially is critically supported by the hippocampus" (p257). However, Eichenbaum [2003] himself concedes that the brain is plastic and brain areas putatively specialised for one function may be commandeered for use of other functions (such as when visual processing areas are used for non-visual processing). Even though there are similarities between mammalian brain structures, the overall plasticity of brains means we cannot draw too strong a parallel between human experience and animal experience. At best we can assume (based on supporting evidence from MRI studies) that similar brain areas can perform similar mechanistic functions, such as encoding information transmitted to the brain from external signals via the sense organs. Therefore, we should be cautious about ascribing higher cognitive capacities such as inferential reasoning to animals based on the similarity of human/animal brain components. I cover this point in more detail later.

**The Value Transfer Theory**

Despite Eichenbaum's claim that the transitivity experiment indicates inference, alternative associative theories are still plausible. Allen [2006] describes one such theory, known as the 'value transfer theory': "...even though B and D are individually rewarded at the same rate, B is seen in association with A, which is always a winner. This is hypothesized to give B a positive boost in comparison to D" (p177). Zentall [2001] describes experiments on pigeons that positively demonstrated this effect. In this experiment two pairs of stimuli were differentially rewarded: stimulus A was rewarded 100% of the time over B that was never rewarded; and C was rewarded 50% of the time over D that was never rewarded (denoted $A_{100}B_0$; $C_{50}D_0$). When B was tested against D, B was preferentially selected, as predicted by the value transfer theory. Zentall concluded that "These results support value transfer theory and suggest that it may not be necessary to posit an ordered representation of the stimuli experienced during transitive-inference training" (p74).

Further pigeon experiments by Zentall [2001] deserve much greater attention than they have so far garnered. These disclose the fact that the value transfer theory alone is inadequate and that there are other quite subtle associative effects that need to be accounted for in the context of the transitive inference paradigm. The value transfer theory predicts that just as A transfers positive value to B, B should also transfer *negative* value to A. To test for *negative* value transfer Zentall tested pigeons on ($A_{100}B_0$; $C_{100}D_{50}$; test A vs. C). Negative value transfer predicts that C would be preferentially selected over A since the low-value B should 'drag down' the value of A by association, but Zentall found no such effect. Zentall then increased the pigeons' experience with the negative stimuli B and D by reinforcing them in unpaired conditions: ($A_{100}B_0$; $C_{100}D_{50}$; $B_0$; $D_{50}$; test A vs. C). However, instead of a preference of C over A as predicted by negative value transfer, the result was a preference for A over C! Zentall offers as a possible explanation that rather than a negative value transfer from B to A, the effect is the result of 'positive contrast'. That is, the value of A is enhanced by *contrast* with B (A=100 vs. B=0) as compared with the differential between C and D (C=100 vs. D=50). Testing ($A_{100}B_0$; $C_{50}D_0$; $B_0$; $D_0$; test B vs. D) yielded preference for D over B indicating a *negative* contrast effect (i.e., B is much worse compared to A than D is compared to C).

The upshot of these results is that the five-element test paradigm as it stands is inadequate to yield conclusive evidence of inferential reasoning in rats. Zentall's pigeon tests need to be

performed in rats and should inspire further variations. The results so far are still far too open to plausible interpretation by alternative theories until many more rigorous tests are performed. What (as Allen [2006] remarked) would a transitive inference theory predict about a circular rather than a hierarchical series – that is, in which the trained pairs are as follows: A>B; B>C; C>A? Given the uncertainties surrounding this experimental paradigm we should discount the results obtained as sufficient evidence of rationality in rats.

### 8.5 Causal Reasoning

Blaisdell, Sawa, Leising & Waldman [2006] claim that rats are capable of causal reasoning. Causal reasoning could be interpreted as a primitive ability to 'grasp' the causal power of objects in a non-conceptual sense [Hoerl 2011], such as that an event B is always followed by event A. This is a case of associative learning despite the temporal separation between the associated events. However, if 'reasoning' is here meant to involve inferential thinking then this would be a case of rationality. According to Blaisdell et al., in their experiments rats made causal *inferences* that cannot be explained by associative theories but are consistent with Bayes net theories. The references to *reasoning* and *inferences* imply that a case for rationality in rats can be made on this basis. As such I examine the experiment to determine if a conclusion of rationality is warranted. My conclusion is that while a reasonable case can be made for causal 'reasoning' in Hoerl's non-conceptual sense, this does not entail a case of rationality. *Pace* Blaisdell et al., the observations can indeed be explained using associative theories.

Blaisdell et al. [2006] trained rats such that subsequent to a ten-second flickering light (L) they were presented with *either* a ten-second tone (T) *or* ten seconds of sucrose delivery (F). The rats were also trained to associate a ten-second noise (N) with simultaneous delivery of ten seconds of sucrose (F) (see figure 8.3). In the test phase rats were allocated to one of four conditions in which a lever was available (the lever was not available during the training). These were: intervene-T; observe-T; intervene-N; and observe-N. In the intervene-T condition, rats were presented with a ten-second T on pressing the lever, while in observe-T

pressing the lever had no effect but T was presented 'randomly'[62]. The same rules held for intervene-N and observe-N (see left side of figure 8.4). The number of nose pokes into the food aperture was recorded as a measure of the rats' expectation of F

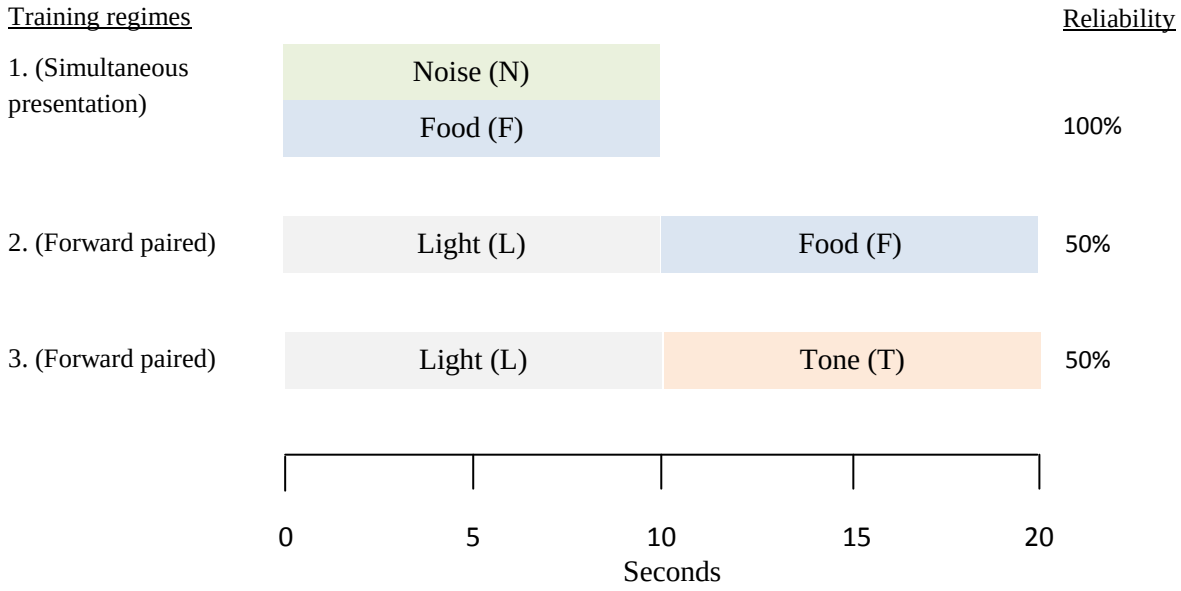



Figure 8.3: Training phase of the causal reasoning experiment

According to Blaisdell et al. [2006] rats in the observe-T condition should reason that T was caused by L (but L was 'missed') so F should also be present as F is also caused by L. Rats in the intervene-T condition should reason that since the cause of T was a lever press rather than L, then F would not be present. If so, this predicts a lower rate of nose pokes for intervene-T than for observe-T. By contrast, there should be no difference in nose pokes between intervene-N and observe-N, since the rats would reason that N is a direct cause of F irrespective of lever presses. The average number of nose pokes per ten-second presentation of T or N were as follows: intervene-T 10; observe-T 15; intervene-N 19; observe-N 20 (see right side of figure 8.4). The difference in nose pokes was marginal between the intervene-N and observe-N conditions, but significant between the intervene-T and observe-T conditions, consistent with predictions.

[62] The presentations of T in the observe-T condition occurred at exactly the same instances as for the intervene-T: the compartments were yoked such that the intervene-T rat's lever presses produced tones in both compartments simultaneously [Blaisdell, personal communication].

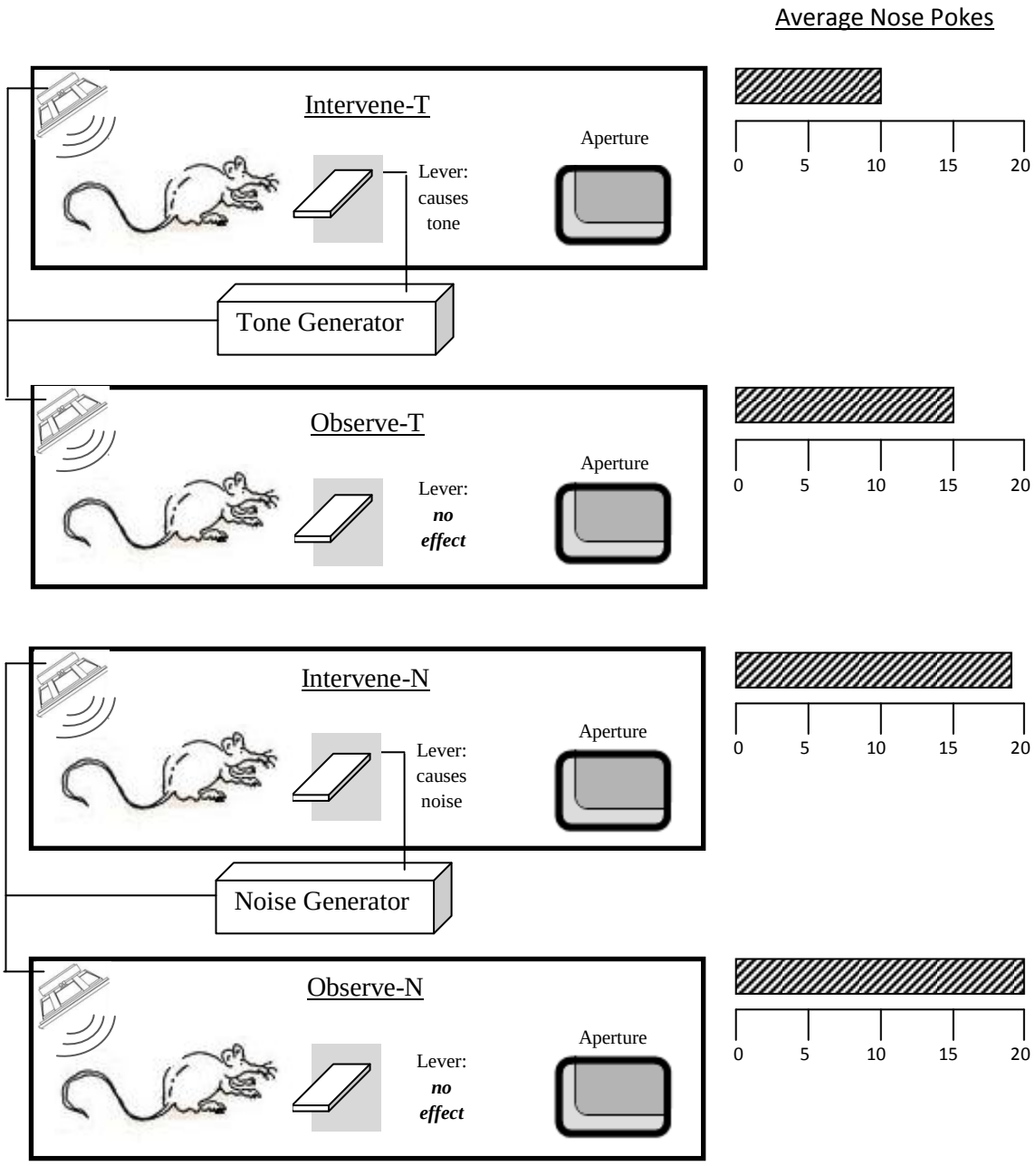



Figure 8.4: Test (extinction) phase of the causal reasoning experiment

There are at least two areas in which an explanation of inferential reasoning can be applied. The first relates to the fact that in the test phase the rats searched for a food presentation when T occurred (even though L was never presented) and despite the fact that during training either T or F was presented following L but not both. Transitive association can account for the fact that T caused the rats to search for F, but in regard to the absence of L, Blaisdell et al. comment that "Apparently, in the initial phases of learning, rats tend to conservatively treat

the absent but expected events as possibly present but missed" (p1021). Thus, the reasoning during the observe-T condition can be construed along the following lines:

**Observe-T**

(1) L causes T
(2) L causes F
(3) T occurred, therefore L must have occurred (and I must have missed it)
(4) Therefore F will occur

The second case of possible inferential reasoning is in regard to the intervene-T condition in which the rat attributes the occurrence of T to its own lever-pressing action rather than to a 'missed' L. This can be construed along the following lines:

**Intervene-T**

(5) T was caused by my lever press so there was no L
(6) Therefore F will not occur

The results of the experiment, however, can be accounted for using associative theories as I explain next. Furthermore, the associative explanation also accounts for effects not adequately accounted for by the inferential reasoning explanation: (i) why the observe-T nose pokes were significantly less than the observe-N nose pokes; and (ii) why the intervene-T nose pokes were significantly above zero.

The effect of the training was to form the following variable-strength associations in the rats:

(A) N => F. Strong association: F was presented at 100% probability following N
(B) L => T. Weak association: T was presented at 50% probability following L
(C) L => F. Weak association: F was presented at 50% probability following L
(D) (L=>) T<=>F. Weak association, transitively implemented via (B) and (C). (The notation is intended to show a weak association between T and F irrespective of the presence or absence of L)

In addition, during the test phases the following new associations are introduced:

(E) Lever => T (in the intervene-T condition)
(F) Lever => N (in the intervene-N condition)

As there has been no trained association of (G) 'lever => T<=>F', any expected association by the rat must be transitively derived based on (D), but (G) must be weaker than (D) since it is derivative of (D) and incorporates the previously unencountered element of the lever. Thus, this is the weakest association and results in the least number of nose-pokes (intervene-T). An

association of (H) ‘lever => N => F’ is also transitively derived, this time based on (A) and will be weaker than (A) as observed (fewer nose pokes for intervene-N than for observe-N). The fact that the difference in average nose pokes between intervene-N (19) and observe-N (20) is small while that between intervene-T (10) and observe-T (15) is relatively large can be accounted for by the relative strengths of the original associations. (A) is a strong association, while (D) is weak, so the weakness of the derived association is magnified in the case of (G), but in (H) is virtually swamped by the overall strength of (A).

If this explanation is correct it can be easily demonstrated with a slight modification of the experiment, in which during training F is presented at (say) 50% probability of N rather than 100%. Additionally, N should be forward-paired with F (rather than simultaneously presented) to more closely resemble the other trained associations. Then the association strength of (A) would more closely match that for (D). The explanation provided above predicts that in this set-up the profile of nose pokes for intervene-N and observe-N would closely resemble that for intervene-T and observe-T respectively.

Now, to account for the two observed effects left unaccounted for as mentioned earlier: (i) why the observe-T nose pokes (15) were significantly less than the observe-N nose pokes (20); and (ii) why the intervene-T nose pokes (10) were significantly above zero. These can be explained by the relative strengths of associations much better than an explanation based on inferential reasoning. Effect (i) is not predicted by the inferential reasoning explanation but is predicted on the associative account based on the fact that association (A) is stronger than (D). For effect (ii), an inferential reasoning explanation predicts zero nose pokes, since if the rat reasoned that L had not occurred during intervene-T there would be *no* reason to check for F rather than a ‘partial reason’ to check. However, the actual non-zero nose pokes is predicted by the association strength analysis: (G) has a weak but *non-zero* association strength.

Several other factors give reason to doubt that the rats are PP-rational. For instance, consider that an alternative explanation for the non-zero nose pokes in the intervene-T condition (i.e. effect (ii) mentioned above) compatible with the inferential reasoning model is that “...even if the rats thought it unlikely that food would be there they would have a look-see just to be sure” [Blaisdell, personal communication]. However, one must question how many such nose pokes would be conducted before a rational rat came to the conclusion that there would *never* be any food there. The *more* nose pokes the *less* likely is an inferential explanation, and the

average number of nose pokes for intervene-T (10) was only a third fewer than that for observe-T (15). Similarly, note that the rats never grasp the fact that when L has caused T then F does not occur and vice versa, instead, as mentioned earlier, there appears to be a transitively derived association between T and F. Blaisdell et al.'s explanation is that "With few learning trials, rats tend to integrate individual learning relations into a coherent integrated model. Only after many trials do rats encode the explicit absence of the nonpresented cues" (p1021). In other words, the association between T and *lack* of F (and vice versa) requires more extensive training. I have no reason to doubt that this is correct but this is an *associative* type explanation, not an inferential one. As a final example, in the observe-T condition why should we accept the supposition that the rats assume that L was present but simply missed? Are PP-rational creatures likely to continue to make this error time and again?

Blaisdell et al. suggest an explanation should be sought compatible with Bayes net theories in lieu of associative theories but the two should not be considered mutually exclusive – quite the reverse. Bayes Theorem provides a mathematical procedure for calculating conditional probabilities. The process begins with an assumed 'prior' probability of an event (say, the existence of food at a particular time and place) that is modified by subsequent incoming data to produce an updated 'posterior' probability. Naturally, whether applied to human behaviour or to animal behaviour, there is no suggestion that the subjects in question are consciously performing mathematical calculations; the brain does that at the subpersonal level. It is not unreasonable to assume that priors are determined in animals via their evolutionary history and individual previous experiences [McNamara, Green & Ollson 2006]. Current experiences then provide the incoming data that can modify the prior to produce a posterior. In the context of rat experiments, the reinforcement schedule provides the incoming data. The fact that the rat's behaviour modifies as a result of learned associations is evidence of a posterior having been derived. Thus, associative and Bayesian explanations are not incompatible. Bayesian approaches can be considered statistical inference rather than inferential reasoning. Statistical inference can be modelled on algorithmic computers (e.g., see Pearl [1988]), and as such can be instantiated without genuine concept possession (and hence not involving PP-rationality).

Given the doubts raised above, I conclude that these experiments are insufficient evidence of PP-rationality in rats. Nevertheless, rats do appear to perceive some causal relations such as

that *L is a cause of T* and *L is a cause of F* and probably also that lever-pressing is a cause of T (in the intervene-T condition). To the extent that this can be called causal reasoning it is possibly a species of E-rationality. According to Bermũdez [2003] there is no reason to believe that non-linguistic (and hence – to Bermũdez – non-conceptual) creatures understand causation: "The proposal is simply that (at least some) nonlinguistic creatures have a basic capacity to track causal relationships holding between events or facts and that this basic capacity allows them to engage in a primitive form of conditional reasoning" (p146). Bermũdez [2003] emphasises that this level of causal reasoning is widespread in the animal kingdom and is just what we would predict on evolutionary grounds. Thus, there is no reason to assume concept possession for the type of 'causal reasoning' displayed by the rats in these experiments.

Hoerl [2011] maintains that there are primitive abilities to 'grasp' the causal power of objects that do *not* imply concept possession, and this applies to the rats in these experiments. However, Hoerl suggests that there is also a type of causal reasoning that *does* rely on concept possession. According to Hoerl, "In this type of reasoning, causal relations are themselves conceived of as being subject to certain conditions, such that it makes sense to ask how A might be prevented from causing B, how one might enable A to cause B, or how else the relationship between A and B might be interfered with." Certainly the rats in this experiment have not met Hoerl's criterion for concept-based causal reasoning.

## 8.6 Goal-Orientation

In the early 1980s experiments on instrumental learning in rats overturned the prevailing dogma that lever-press acquisition was controlled solely by sensorimotor learning involving a process of stimulus-response (S-R) association and instead suggested that animals are capable of a "more elaborate form of encoding based on the response–outcome (R–O) association" [Balleine & O'Doherty 2010]. Anthony Dickinson, who was involved in many of these experiments, dislikes the overuse of the term 'response' preferring instead 'action' to distinguish between habitual behaviour and "truly purposeful and goal-directed behaviour," and similarly prefers to view 'outcomes' as 'goals' [Dickinson 1985]. At least some behaviour in rats, claims Dickinson, is under teleological control and cannot be explained at the psychological level in terms of internal associations:

> Rather, we argue that the knowledge about the action-goal relation must be encoded in a propositional-like form so that it can be operated on by a practical inference process to generate the instrumental performance. In this sense actions are inherently rational in a way that responses can never be. [Dickinson 1985, p78]

It seems by this that Dickinson not only thinks rats are rational, but that they perhaps entertain thoughts in propositional form. In the quotation above, Dickinson uses the term 'encoded', which might indicate that he means to imply only that *information* is encoded in the lever-pressing rat's brain. However, Dickinson [1985] continually refers to the rat having *knowledge* (by which we can assume he rather means 'belief') about the instrumental relations between its actions and their consequences. If rats are capable of entertaining thoughts in propositional form, as Dickinson apparently believes, then they must possess concepts. If they possess concepts they are capable of inferential reasoning and may be deemed PP-rational organisms. Furthermore, by CPH, as concept possessing organisms they must be considered self-conscious. Here, as earlier, I argue that *information* relating actions to consequences can be encoded in a rat's brain and can motivate behaviours, but that there is no reason to ascribe *knowledge* to the rat. This information can be expressed in propositional terms by we humans, but there is no need to assume that the rat can entertain thoughts in propositional form. There is no proposition that the rat believes. Instrumental learning can be explained in terms of association theories and does not require the assumption of PP-rationality in rats.

When rats are given limited (not extended) training using a simple food reward paradigm, subsequent devaluation of the reward causes the rat to decrease its response (or, *action*). For example, Dickinson trained rats to press a lever for one type of food while they received another type non-contingently. Then only the contingent food was devalued by giving them access to it in the absence of the lever on alternate days after training was complete, and making them mildly ill after food consumption with injections of lithium chloride. Subsequently, during the test (extinction) phase, lever pressing was much reduced compared to a control group. According to Dickinson [1985] this shows that lever pressing is an action (i.e. under the teleological control of the rat) rather than a response, since it demonstrates that the rat had knowledge about the relation between the action and the goal (i.e. that it is the action of pressing the lever that causes the presentation of food). An account based on S-R paradigms, according to Dickinson, is inadequate to explain this behaviour since it should predict no difference in lever pressing frequency as the strength of association between lever-

pressing and food presentation was trained to be the same for both the subject rats and the control group.

It is not in dispute that these results show that the rats have made an association between the pressing of the lever and the presentation of a reward. What is disputed is the interpretation of the nature of that association. By application of Morgan's canon, an interpretation that assumes purposive actions by the rat is unwarranted since a non-cognitive explanation is available. Information that food presentations follow lever pressing is encoded in the rat's brain and acts as a motivator. This information was acquired through training the rat to associate a lever press with a subsequent reward. It was not acquired by the rat making inferences; if it was acquired that way it would not have required several days and dozens or hundreds of repetitions to instil this information.

The fact that the subject rats pressed the lever less frequently than the controls during extinction seems to indicate that they underwent a process of reasoning along the following lines: 'this food makes me sick; I don't want to be sick; pressing the lever produces this food; therefore I will not press the lever'. But again, a line of reasoning like this should lead to the rats not pressing the lever at all rather than just comparatively less than the control group. Alternatively, the results can be explained adequately by an associative account relating to the strength of reinforcement. Although the devaluation processes were conducted separately from the original training, the food stimulus was a common element linking back to the previous association. If we use '=>' to represent 'causes', then we have the following trained associations:

(i) lever-press => food pellet;
(ii) food pellet => sickness;

which together forms:

(iii) lever-press => food pellet => sickness.

In other words, the sickness is transitively associated with the lever-press (in similar fashion to that previously discussed in the earlier section of transitive inference). Obviously, since the sickness is a *negative* reinforcer, this would tend to reduce the reinforcement strength of the lever press. Accordingly, the frequency of lever presses would be reduced compared to the control group but would still be non-zero, as observed.

Further support for the idea that conditioning is sufficient explanation for instrumental learning is provided by Robert Rescorla. In trying to understand the nature of instrumental behaviour, Rescorla [1990] compared it to standard Pavlovian (S-R) conditioning. Rescorla surmised that instrumental training may result in more elaborate hierarchical structures involving not just the stimulus (S) and the response (R) but also the outcome (O) in an analogous way to the standard S-R paradigm. For example, "...one might consider a structure of the form S-(R-O) in which the R-O association itself becomes associated with the stimulus. Then, that stimulus might functionally activate not an element but rather an association between elements" [Rescorla 1990, p262]. To test this idea, Rescorla noted that in Pavlovian conditioning a conditioned stimulus (CS) that gives information about the unconditioned stimulus (US) develops a stronger association with the US. Rescorla considered whether an instrumental discriminative stimulus that provides information about the R-O relation also develops better control over responding: if so then this suggests that the S-(R-O) model is a valid Pavlovian paradigm. Rescorla successfully designed and tested S-(R-O) analogues to three standard Pavlovian phenomena in this respect[63] and found that:

> ...three of the major phenomena central to modern conceptions of Pavlovian conditioning have analogies in instrumental learning in a fashion consistent with the view that learning entails the development of an association between the stimulus and the R-O relation. [Rescorla 1990, pp268-269]

Given this evidence and the availability of alternative associative-based explanations, it is not necessary to ascribe conceptual abilities to rats to account for instrumental learning and apparent goal-orientation.

As part of his argument for goal-orientation in rats, Dickinson [1985] points to the fact that overtraining can override the devaluation effect and cause a habit to form. In this case, lever-pressing is not decreased during the extinction phase. According to Dickinson [1985]: "...when the animals are over-trained, they no longer experience the behaviour-reward correlation with the result that their performance is no longer controlled by knowledge about this relation. In the absence of such knowledge, reward devaluation can have no effect and a habit has been established" (p75). No justification is given, however, for the claim regarding what the rats are *experiencing* (which, based on the paper overall, I take as meant literally rather than in the sense of simply registering the correlation). After all, there is no empirical

[63] These are the Kamin blocking effect; the importance of CS-US contingency; and the role of relative stimulus validity.

evidence we can get that will enable us to know 'what it's like to be a rat' (or a bat [Nagel 1974]). It is thus pure conjecture that the effect of overtraining is to decrease the salience of the behaviour-reward correlation, and indeed it is counterintuitive to surmise that this 'knowledge' should decrease rather than increase with continued exposure.

Daw, Niv & Dayan [2005] have provided an explanation of the overtraining effect based on rat brain studies. They show that there is competition between two computational systems, controlled by two different areas of the brain, and that the final arbitration is achieved by a Bayesian process. The dorsolateral striatum supports habitual control (in which the action has become insensitive to the value of the outcome), while prefrontal cortex supports 'goal-directed' behaviour (in which the action is sensitive to the value of the outcome). Thus, according to Daw et al. [2005], in effect overtraining is a process of transferring control from the prefrontal cortex to the dorsolateral striatum. The basic idea is that during initial training it is more efficient to use the prefrontal system despite the extra cost in processing because of the relative uncertainty of the outcome values. Once the uncertainty is reduced by continued training, it becomes more efficient to transfer control to the dorsolateral striatum, which has a lower cost of processing. Again, this transfer is not a conscious or intentional process. Rather, it appears to be a case of E-rationality in that the control system processing is simply being optimised via a Bayesian process based on continued feedback. On the other hand, in humans the pre-frontal cortex is associated with 'executive control' (e.g., Koechlin & Summerfield [2007] and so the possibility exists that prefrontal areas of a rat's brain has similar functionality, as discussed next.

Performance of decision-making tasks in humans has also shown evidence of competition between brain areas homologous to that found in rats [Balleine & O'Doherty 2010]. There is a temptation to anthropomorphise the experience had by rats based on this brain homology and behavioural correspondence: since homologous prefrontal brain systems are responsible for goal-directed behaviour in both rats and humans, it seems reasonable to assume that they might undergo similar experiences. The presumed phylogenetic continuity of brain development adds to this temptation: similar brain structures are likely to continue to perform the same functions as more developed organisms evolve. While this is true, we still have to figure out where to draw the line. For example, it is currently thought that outcome values are set by association with emotional feedback and hence that the amygdala plays a critical role in goal-directed action [Balleine & O'Doherty 2010]. Now, it is probably not unreasonable to

assume that a rat's emotional responses during conditioning is not too different to a human's given corresponding brain structures (i.e. the rat and human amygdala). But while we can assume a rat experiences (say) fear, should we assume any further cognitive correspondence, such as that the rat also is capable of the metacognitive capacity to *understand* (has a concept of) fear? Although there are *similarities* between human/rat brains *and* behaviours there is still an enormous gap. Humans solve mathematical equations and send rockets to other planets; rats do not. The evidence of homologous brain areas and comparable conditioning behaviour between rats and humans ought to lead us to the conclusion that humans are like rats rather than rats are like humans.

## 8.7 Conclusion

In defining what should be taken as rationality in animals as discussed in chapter 4 I have deliberately set the bar high. This was not done to imply that there is no 'middle ground' between non-rational and rational; undoubtedly this is not a quantum jump and there most likely is a gradation of increasing rational capacity. However, when the question is whether an animal has self-consciousness in much the same way as experienced by human adults, a determination of 'sort of' will not do. If CPH is correct and concept possession implies self-consciousness then we need a very convincing case of concept possession, such as true inferential reasoning, or what Kacelnik [2006] has labelled PP-rationality.

To determine whether rats are PP-rational I have examined five paradigmatic examples of research. The first to be considered, spatial navigation (such as used by rats to locate the escape column in a Morris water maze or to explore arms in a radial maze) is part of the species' repertoire and does not appear to rely on inferential reasoning. In regard to the other four paradigms, the main competition is between associative (non-inferential) theories and inferential reasoning theories. Regarding the bail-out tests for metacognition, a plausible explanation of the observed behaviour is the association of the bail-out option itself with a particular reward. Although that reward is of lesser value than for 'correct' discriminations, natural (non-conscious) mechanisms can explain the overall maximisation of food delivery efficiency. At best, this might meet Kacelnik's criterion for E-rationality but not PP-rationality.

The experiments on cross-associations that Zentall [2001] conducted on pigeons indicate the deep subtleties involved in transitive associations. Relative reward values for different stimuli can be carried across to untrained stimulus pairs in surprising ways. Zentall demonstrated not only the value transfer effect but also positive and negative contrast, and suspects further yet-to-be discovered similar phenomena. These experiments should also be conducted in rats: the effects can account for the apparent transitive inference demonstrated in the five-element test. Rather than 'inference' these experiments show transitive association.

Transitive association can also account for so-called causal reasoning and goal-orientation. The results of the causal reasoning experiments can be accounted for based on the relative association strengths transferred to newly introduced associations (i.e. the lever=>tone/noise in the intervene-T/N tests). That relative strengths of association have a significant impact on behaviour was amply demonstrated by the aforementioned Zentall [2001] experiments. Furthermore, the association strength analysis also accounts for other effects in Blaisdell et al.'s experiments that go unexplained by inferential theories.

In the goal-orientation experiments described by Dickinson [1985] an inferential explanation is given for why the rats that were averted from the contingent food reduce their lever presses. But this behaviour can adequately be explained by the transitive association formed between lever, food and sickness and the fact that the reinforcement strength is reduced by the aversion. There is no need to assume propositional-like thoughts in rats as Dickinson does to explain the experimental results. Furthermore, very little can be concluded by the existence of homologies between rat and human brains with respect to the phenomenal experiences rats undergo. Rather, these homologies can explain why humans sometimes behave non-rationally, like rats, since humans must have inherited and retained some brain functions from phylogenetically common ancestors; it would be imprudent, I suggest, to assume the reverse. It would be well for researchers to keep this in mind when, for example, using rats as models to explore human psychology (e.g see Quirk & Beer [2006]).

Common to many of these experiments is the fact that differential effects emerge after averaging out results over many trials. Thus, for example, lever presses may be reduced compared to a control group but (though statistically significant) usually not by enormous degrees. This fact is not adequately explained by inferential accounts (in which a reduction to

near *zero* would be more convincing)[64]. However, these results are to be expected if the information encoded and associations formed by rats are embedded as Bayesian probabilities. In this respect I largely concur with Kim Sterelny's assessment:

> But for better or worse, while rats are good probabilistic reasoners, it is my hunch that they cannot adequately change their reasoning dispositions. To learn about reasoning, you need to be able to represent and evaluate your own reasoning capacities. [Sterelny 2006, p310]

The point to emphasise here is the need for metacognitive abilities; when one is reasoning, one *knows* that that is what one is doing. Humans can and sometimes do perform actions without having undergone a reasoning process, and sometimes these actions may turn out to be the best option available. When then asked why the action was taken the response might be "I don't know, it was just instinct." But when a human makes a decision through reasoning she is (self-)conscious of the fact that she is reasoning, even if the process is flawed. Your reasoning may lead to incorrect conclusions, and you might not be able to spot the flaw, but as long as you undergo a process of inference then it counts as PP-rational behaviour. There is insufficient evidence to conclude that rats are undergoing this process.

[64] This objection might be levelled at the scrub jay experiments described in chapter 7 (in which I argued that inferential reasoning did take place) where scrub jays cached food according to what they expected to be fed the next day. However, in that experiment the objection does not hold. The scrub jays preferentially cached the same food they had been sated with rather than exclusively caching the same food. A scrub jay that had reasoned it would be fed exclusively on one type of food would not necessarily decide to only cache the other type of food. It might still wish to increase the availability of the expected food and so cache more of the unexpected food and still a little of the expected food. This action is not inconsistent with the scrub jay having performed explicit inferential reasoning. The situation is different with the rat experiment as in that case the rat has no reason to nose poke if it has come to a reasoned conclusion that no food would be there.

# Chapter 9: Closing Comments

The topic of this thesis is self-consciousness and how we can determine its existence in animals and human infants. The thesis is divided into two parts. The first part is theoretical and explores the nature of self-consciousness while the second analyses experimental studies of self-consciousness in animals and human infants. In chapter 1 I explained what I mean by 'self-consciousness'. I emphasised that I apply the notion to organisms that comprehend their own existence in the *psychological* sense, not just as physical objects (bodies). I argued that this entails an understanding of one's intentional agency and involves a capacity for metacognition. This discounts what might be called 'primitive' self-consciousness. For example, an organism might be in a particular non-conceptual mental state that is in some way self-specifying (such as pain), but this does not count as self-consciousness as I use the term. The standard here is rather high; indeed I characterised my notion of self-consciousness as 'something like that possessed by human adults'. Nevertheless, this notion of self-consciousness does not extend to a sense of personal identity or self-worth.

In chapter 2 I made a distinction between two views on the nature of self-consciousness (the 'Fundamental Dichotomy') and argued in favour of the view that self-consciousness is relational rather than intrinsic. The intrinsicist view has it that a subject can be inherently self-conscious without the mediation of an inference or self-observation. The relationalist view, by contrast, sees self-consciousness as always involving a relation of some sort – such as that between a self-conscious thought and the thinker of the thought. I argued in favour of the relationalist side, in part by suggesting that introspection is an action requiring a subject to perform the action and an object on which the action is performed (even though the subject and object are both one and the same self). To validate this position in the wider context of the philosophy of self-consciousness, I measured it up against the known properties of self-consciousness as well as the major 'classic' theories such as *immunity to error* and the *essential indexical.* I found that relationalism adequately meets the challenge issued by Kriegel [2007]: that the 'peculiarities of self-consciousness' must be accounted for in the context of a general theory of self-consciousness.

The relationalist view of self-consciousness underpins the central claim made in this thesis, the Concept Possession Hypothesis of Self-Consciousness (CPH), presented in chapter 3. The relationalist view acknowledges that there are *some* forms of self-access that are indeed

intrinsic – that is, direct and unmediated – but claims that these primitive, non-conceptual forms of self-access do not constitute self-consciousness proper. According to CPH, what separates organisms that are self-conscious ('level 3' organisms) from organisms that are *conscious* but not *self*-conscious ('level 2' organisms) is concept possession. Level 2 organisms may have access to certain mental states (such as fear or pain) but are incapable of conceptual mental states and therefore unable to possess a *self*-concept. Not only are level 3 organisms concept possessing, but just by being concept possessing they must possess a *self*-concept. I presented three considerations in support of this thesis. First, I suggested an analogy between conception and perception to illustrate that, just as perception always involves the self-percept, conception always involves the self-concept. Second, I argued that every concept could be linked to the self-concept via a web of necessarily associated concepts. Finally, I used the theory of non-conceptual content, as championed by Bermũdez [1998], to argue that the only factor separating organisms that are conscious but not self-conscious from organisms that are both conscious and self-conscious is the fact that the latter are capable of mental states with conceptual content.

If CPH is correct then it provides a way to determine the existence of self-consciousness in animals and human infants: if a convincing argument can be made that an animal possesses concepts, then this is good evidence for self-consciousness. Thus CPH provides another weapon in researchers' armoury for investigating self-consciousness and claiming its existence in their experimental subjects. Of course, proving concept possession is no easy task itself. Nevertheless, even when keeping the standard high, there should be some observable behaviours in animals that are explicable only in terms of concept possession. In chapter 4 I discussed which behaviours might qualify. These are any behaviours that demonstrate rationality, symbol-mindedness or propositional thinking (as well, of course, as any behaviour that directly indicates a particular concept possessed by the experimental subject). I argued that it is extremely difficult to make a convincing claim of propositional thinking in non-linguistic organisms, but that both rationality (inferential reasoning) and symbol-mindedness provide opportunities for research paradigms. With these theoretical underpinnings in place, I was able in part 2 to evaluate a variety of research paradigms investigating self-consciousness in animals and human infants.

I devoted the first three chapters of part 2 to the evaluation of three research paradigms to determine whether any of them qualified as valid demonstrations of concept possession and

hence self-consciousness. In chapter 5 I examined mirror self-recognition (MSR) and concluded that it does demonstrate self-consciousness on the grounds that it entails the subject is 'symbol-minded'. Subjects that recognise themselves in a mirror cannot be taking the mirror image as a physical entity located behind the mirror. This is because the image is distal and unconnected to their actual body. Instead they must understand that what they see is a *representation* of their bodies. In effect, the subject demonstrates an ability to ascribe a meaning (the self) to the representation (the mirror image) thereby showing they are symbol-minded. The protocol developed by Gallup [1970], known as the *mark test*, was found suitable for experiments on animals and has been used to convincingly demonstrate MSR in chimpanzees (with strong indications of MSR in dolphins, elephants and magpies).

Imitation was examined in chapter 6 as one way of determining Theory of Mind (ToM). I argued that ToM indicates self-consciousness as it implies the ability to infer the mental state of others. All theories of how ToM works rely, to varying degrees and in different ways, on using the *self* as a model for understanding the minds of others. Thus, other-awareness (in the psychological sense rather than the physical sense) depends on self-awareness. Furthermore, since the capacity for ToM indicates that the subject is *inferring* the mental state of the other, this implies concept possession and hence self-awareness by CPH. Imitation comes in a variety of different forms, many of which are reflex-like and – at least in primates – most likely underpinned by hard-wired circuits (i.e. mirror neurons). Consequently I reconfigured the question to 'what *type* of imitation can be taken as evidence of self-consciousness?' Based on some ingenious experiments by Harris & Want [2005], I concluded that Selective Imitation should be considered good evidence of self-consciousness. In this paradigm the imitators, by selecting only a certain subset of the demonstrator's actions, display an understanding of the demonstrator's intentions, therefore showing ToM.

The topic explored in chapter 7, episodic memory, has long been associated with self-consciousness, though usually the term used is 'autonoesis', as coined by Endel Tulving [1972]. Autonoesis and self-consciousness are used mostly interchangeably in the literature, but I argued that there is a significant difference not fully appreciated by most authors on the subject. Implicit in autonoesis is a conceptual understanding of the self existing in past, present and future. But in my conception of self-consciousness it is only necessary to understand the self as existing in the present. Thus the concept of autonoesis is distinct from self-consciousness. Nevertheless, a demonstration of autonoesis obviously encompasses self-

consciousness. Correia et al. [2007] observed episodic memory in scrub jays that could overcome a natural tendency to act according to current motivations by planning for a future need. Instead of caching food of a *different* sort from that on which they had been sated, they preferentially cached the *same* type of food on the expectation of receiving the other type the next day. I took this experiment as good evidence that scrub jays are autonoetic – they have an understanding of themselves persisting into the future. Thus scrub jays are not only self-conscious, they are capable of episodic memory, which in infants is usually not expected to develop until about age 3. This apparently extraordinary conclusion is supported by independent data, such as the fact that other species (magpies) of the same family (corvids) have passed the test for mirror self-recognition.

In chapter 8 I took a different approach to the question of animal self-consciousness. Instead of examining one particular research paradigm, I chose to examine one animal *species* on which many different types of experiments are done. The animal I chose to examine is the lab rat, which many researchers are claiming (explicitly or implicitly) are rational. No researchers (to the best of my knowledge) are claiming *self-consciousness* in rats. However, any truly rational organism (i.e. any organism capable of inferential reasoning) must possess concepts and hence according to CPH must be self-conscious. Therefore, if any of the claims for rationality in rats are correct I would be committed to ascribing self-consciousness to rats. I selected five research paradigms in which claims of rationality (or claims that indirectly imply rationality) in rats have been made: spatial navigation; metacognition; transitive inference; causal reasoning; and goal-orientation. In each of these, following in-depth analysis, I concluded that explanations based on associative theories could not be ruled out. As such, there is no reason based on the available evidence to ascribe rationality to rats. Consequently there is as yet no evidence of self-consciousness in rats.

## 9.1 Future Research

There are of course many avenues for further research on both theoretical and empirical topics. On the empirical side I have already mentioned some of these. In chapter 5 I concluded that if CPH is correct then symbol-mindedness must arise in human infants prior to or concurrently with MSR and I suggested longitudinal studies to validate this claim. In chapter 8 regarding the Blaisdell et al. [2006] experiments I claimed that the differences in

nose-pokes between the ‘intervene/observe-N’ and ‘intervene/observe-T’ conditions could be explained by the varying strengths of associations formed during the training phase. I then suggested suitable modifications to the training phase to validate or invalidate this claim.

The Harris & Want [2005] experiment on Selective Imitation was conducted on human infants and is yet to be performed on animals. The key factor was an exclamation of ‘oops!’ by the demonstrator to signal that an incorrect action had been taken. It might be possible to adapt the method for use with animals, by pre-training the subjects to associate a particular signal (such as an exclamation of ‘oops!’) with a (generic) mistake. This might be achieved if a sufficient number of incorrect actions recognisable to animal subjects can be determined. For example, primates might recognise actions such as dropping food or falling from branches (etc.) as ‘incorrect’ actions and by training come to associate a signal (such as an accompanying ‘oops!’) with *any* incorrect action. Then the Harris & Want experiment may be applied to those subjects and a control group. If the trained subjects (but not the controls) showed Selective Imitation based on an exclamation of ‘oops!’ by the demonstrator, this would then be good evidence that the training had been successful. Not only would this indicate that the subjects could infer the demonstrator’s intentions and/or state of mind, but it would be evidence that the subjects had learned that ‘oops!’ indicates *any* incorrect action. That would indicate that the subjects had acquired a concept of ‘incorrect action’, rather than only associating each particular trained incorrect action with the exclamation. As evidence of concept possession this would, by CPH, be additional evidence of self-consciousness.

As mirror self-recognition (MSR) is the gold standard for self-consciousness, I would be keen to see the mark test applied to many more animals in which the protocol could be effectively applied. For example, many rodents have sufficiently flexible forelimbs to be able to touch unseen marks on their foreheads. Therefore, I suggest lab rats should be tested for MSR, which would be a direct determination of self-consciousness compared to seeking evidence of rationality in rats as I did in chapter 8. Octopi are suspected of high intelligence and might also be morphologically suitable subjects for the mark test. The mark test has been successfully applied to magpies and should be applied to other birds, especially other corvid species (such as scrub jays, crows, rooks and ravens). Corvids have relatively large brains compared to body weight and are already known to display intelligent behaviour, so they are primary contenders for self-awareness according to CPH.

On the theoretical side there are many related issues that are too far outside the scope of this thesis to have been addressed in detail but may be worthy of further research. How does the brain generate consciousness and self-consciousness (often referred to as the 'Hard Problem' of consciousness)? How did consciousness and then self-consciousness evolve? Though very difficult questions, there are some promising lines of inquiry. The advent of consciousness could be explained by *emergence* theories, in which novel phenomena arise as a consequence of increased system complexity, where 'system' could refer to configurations of matter such as the brain (e.g., Braddon-Mitchell [2007]; Hempel & Oppenheim [2008]). As a simple example of emergence, the phenomenon of organic chemistry arose as a consequence of inorganic molecules self-organising into the more complex organic molecules. Conceivably, consciousness emerged as a result of the increasing complexity of brains (or nervous systems) reaching a critical threshold.

Self-consciousness, being a (rather special) extension of consciousness, may have been an adaptation driven by natural selection. Many authors speculate that higher cognitive capacities evolved in primates generally and in hominins in particular due to their social complexity (for a summary, see Byrne [2000]; for a more recent formulation, see Sterelny [2012]). In these theories the problems posed by the need for complex interactions with social companions was the principal selective pressure for development of higher cognitive capacities. I suggest an alternative point of view, that self-consciousness may have arisen as a result of self-oriented problems; problems faced by individuals where actions derived from rational thought (that is, inferential reasoning) would prove advantageous over mechanistic (i.e. stimulus-response type) behaviour. So, it may have been the need to solve what might be called 'I-problems' – environmental challenges impacting individual survival as opposed to challenges impacting *species* survival – that selected for self-awareness. This view is not incompatible with the social complexity theory: the I-problems presumably included those that arose due to the proliferation of social companions.

Of course the preceding speculation still leaves much to be explained. However an intriguing proposal has been made by Povinelli & Cant [1995] that is illustrative of the idea I am suggesting. Povinelli & Cant suggest that some form of self-consciousness developed as a result of 'arboreal clambering' where individual apes needed a sense of the self in order to safely navigate through the fragile canopy. In this view the origin of self-conception occurred as the result of our ancestors' remaining in the trees long after their increased body size

favoured descent. Povinelli & Cant suggest that a sense of personal agency, and hence self-conception, emerged in the context of locomotion. Clambering induced by body weight drove the evolution of this sense of personal agency, and this psychological capacity supports the behaviour of self-recognition. This idea is a reasonable example of what I mean by the selection pressure of the need to solve 'I-problems'.

Another area of research is the possible creation of *artificial* self-consciousness. Concept possession is the central issue in CPH, but there is a question as to how concepts are instantiated at the neurophysiological level. This is critical if we are to create true artificial intelligence. 'Artificial intelligence' is a misnomer as currently applied to computers since even the most sophisticated computers (such as IBM's Deep Thought and Watson) are not truly intelligent. If they were truly intelligent – that is, if they possessed concepts – then by CPH they would also be self-conscious entities. But clearly, as consciousness and self-consciousness are supervenient on the physical brain, there must be a way in principle to create artificial self-consciousness by simulating brain processes. Current research in this area focuses on replicating neural circuitry, which is an obvious and sensible place to begin. However, is not clear that this aspect of the brain will suffice to do the job. What happens in the brain involves a lot more than just the neural network architecture. For example, nearly half of the human brain is white matter, the 'cabling' (axons) that interconnects the different neuronal regions of grey matter, and this is no longer thought to be passive tissue [Fields 2008]. It is now suspected that myelination (the build up of fatty material called myelin that sheaths axons), which continues well into adulthood, is an important factor in the plasticity of the brain [Fields 2005]. Although much research is engaged in searching for the *neural* correlates of consciousness, very little takes into account the contributions made by other factors such as myelin and other components in the brain. Thus, even if (or when) we are able to replicate the neural architecture of the brain in all its complexity, we may still not have managed to create artificial self-consciousness. We may need to replicate all non-neural aspects of the brain as well. As enormous a challenge as this is, I fully condone putting research effort into this grand endeavour.

# References


Allen, Colin (2006). Transitive Inference in Animals: Reasoning or Conditioned Associations? In Hurley & Nudds, (eds.) *Rational Animals?* Oxford University Press, London, England, pp175-185

Allport, Gordon W (1943). The Ego in Contemporary Psychology. *Psychological Review*, Vol. 50, pp451-478

Amsterdam, Beulah. (1972). Mirror Self-Image Reactions Before Age Two. *Developmental Psychobiology* Vol. 5, pp297-305

Anderson, James (1984). The Development of Self-Recognition: A Review. *Developmental Psychology*, Vol. 17, no. 1, pp35-49

Anderson, J. R. & Gallup, G. G. (2011). Do rhesus monkeys recognize themselves in mirrors? *American Journal of Primatology*, Vol. 73, pp1-4

Andrews, Kristin, (2011). Animal Cognition. *The Stanford Encyclopedia of Philosophy* (Summer 2011 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2011/entries/cognition-animal/>

Anisfeld, Moshe (2005). No Compelling Evidence to Dispute Piaget's Timetable of the Development of Representational Imitation in Infancy. In Hurley & Chater, *Perspectives on Imitation, Vol 2: Imitation, Human Development, and Culture*. A Bradford Book, MIT Press, Cambridge Mass. & London, England, pp 107-132

Armstrong, D M (1968/1994). Introspection. In Cassam (ed.), *Self-Knowledge*. Oxford University Press, Oxford, UK., pp109-117

Asendorpf, Jens (2002). Self-Awareness, Other-Awareness, and Secondary Representation. In Meltzoff & Prinz, *The Imitative Mind: Development, Evolution, and Brain Bases.* Cambridge University Press, Cambridge, UK , pp 63-73

Atance, C & O'Neill, D (2001). Episodic future thinking. *Trends in Cognitive Sciences*, Vol.5, no. 12

Atance, C & O'Neill, D (2005). The emergence of episodic future thinking in humans. *Learning and Motivation,* Vol. 36, pp126–144

Aydede, Murat, (2010). The Language of Thought Hypothesis, *The Stanford Encyclopedia of Philosophy (Fall 2010 Edition)*, Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/fall2010/entries/language-thought/

Baddeley, Alan (2001). The Concept of Episodic Memory. In Baddeley, Aggleton & Conway (eds.), *Episodic Memory: New Directions in Research*, Oxford University Press, pp1-10

Balleine, B & O'Doherty, J (2010). Human and Rodent Homologies in Action Control: Corticostriatal Determinants of Goal-Directed and Habitual Action.

*Neuropsychopharmacology Reviews* Vol. 35, pp48–69

Baron-Cohen, S; Leslie, A & Frith, U (1985). Does the autistic child have a "theory of mind"? *Cognition*, Vol.21, pp37-46

Bayne, Tim (2004). Self-Consciousness and the Unity of Consciousness. *The Monist*, Vol. 87, no.2, pp219-236

Berg, K; Delgado, S; Cortopassi, K; Beissinger, S & Bradbury, J (2011). Vertical transmission of learned signatures in a wild parrot. *Proceedings of the Royal Society, B: Biological Sciences*. doi: 10.1098/rspb.2011.0932

Bermũdez, José Luis (1998). *The Paradox of Self-Consciousness*, MIT Press, Cambridge, Massachusetts and London, England

Bermũdez, José Luis (2003). *Thinking Without Words*, Oxford University Press, Oxford, UK

Bermũdez, José Luis (2006). Animal Reasoning and Proto-Logic. In Hurley & Nudds, (eds.) *Rational Animals?* Oxford University Press, London, England, pp128-137

Bermũdez, José & Cahen, Arnon (2012). Non-conceptual Mental Content. *The Stanford Encyclopedia of Philosophy (Spring 2012 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2012/entries/content-non-conceptual/>

Bertenthal, B & Fischer, K (1978). Development of Self-Recognition in the Infant. *Developmental Psychology*, Vol. 14, no. 1, pp44-50

Bigelow, Ann (1981). The Correspondence Between Self- and Image-Movement as a Cue to Self-Recognition for Young Children. *Journal of Genetic Psychology*, Vol. 139, pp11-26

Bird, C & Emery, N (2009). Rooks Use Stones to Raise the Water Level to Reach a Floating Worm. *Current Biology*, Vol. 19, no. 16, pp1410-1414

Bird, L; Roberts, W; Abroms, B; Kit, K & Crupi, C (2003). Spatial Memory for Food Hidden by Rats (*Rattus norvegicus*) on the Radial Maze: Studies of Memory for Where, What, and When. *Journal of Comparative Psychology*, Vol. 117, no. 2, pp176–187

Blaisdell, A; Sawa, K; Leising, K & Waldmann, M (2006). Causal Reasoning in Rats. *Science*, Vol. 311, pp1020-1022

Block, Ned (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*. Vol. 18, pp227-287

Bloom, P & German, T (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, Vol. 77, ppB25-B31

Bolton, Neil (1977) *Concept Formation*, Pergamon Press, Oxford, England

Braddon-Mitchell, David (2007). Against Ontologically Emergent Consciousness. In

McLaughlin, B & Cohen, J (Eds.) *Contemporary Debates in Philosophy of Mind*, Blackwell Publishing Ltd., USA, UK & Australia, pp287-299

Brewer, Bill (2005). Do Sense Experiential States have Conceptual Content? In E. Sosa and M. Steup (Eds.), *Contemporary Debates in Epistemology*, Blackwell, Oxford, UK

Brook, Andrew (2001). Kant, Self-Awareness and Self-Reference. In Brook & DeVidi, (Eds.) *Self-Reference and Self-Awareness*, John Benjamins Publishing Company, Amsterdam/Philadelphia, pp 9-30

Brown, Harold (1986). Sellars, Concepts and Conceptual Change. *Synthese*, Vol. 68, no.2 p.275

Bruner, Jerome (1997). A Narrative Model of Self-Construction. In Snodgrass & Thompson, *The Self Across Psychology*, Annals of the New York Academy of Sciences, Volume 818, New York, pp145-161

Bruner, Jerome & Kalmar, David A (1998). Narrative and Metanarrative in the Construction of Self. In Ferrari & Sternberg, (eds.) *Self-Awareness*, The Guildford Press, New York, pp308-331

Buckner, R & Carroll, D (2006). Self-projection and the brain. *Trends in Cognitive Sciences*, Vol.11, no.2

Bunsey, M & Eichenbaum, H (1996). Conservation of hippocampal memory function in rats and humans. *Nature*, Vol. 379, 18 Jan.

Burge, Tyler (1988/1994). Individualism and Self-Knowledge. In Cassam (ed.), *Self-Knowledge*. Oxford University Press, Oxford, UK., pp65-79

Byrne, Alex (2004). Perception and Conceptual Content. In Steup, M & Sosa, E (Eds.) *Contemporary Debates in Epistemology*, Blackwell Publishing Ltd.

Byrne, Richard (2000). Evolution of Primate Cognition. *Cognitive Science: A Multidisciplinary Journal*, Vol. 24, no. 3, pp543-570

Byrne, Richard (2002). Seeing Actions as Hierarchically Organized Structures: Great Ape Manual Skills. In Meltzoff & Prinz, *The Imitative Mind: Development, Evolution, and Brain Bases.* Cambridge University Press, Cambridge, UK, pp122-140

Byrne, Richard (2005). Detecting, Understanding, and Explaining Imitation by Animals. In Hurley & Chater, *Perspectives on Imitation, Vol 1: Mechanisms of Imitation and Imitation in Animals*. A Bradford Book, MIT Press, Cambridge Mass. & London, England, pp 225-242

Camp, Elisabeth (2007). Thinking With Maps. *Philosophical Perspectives*, Vol 21, *Philosophy of Mind*

Campbell, John (2001). Memory Demonstratives. In Hoerl & McCormack, *Time and Memory*, Clarendon Press, Oxford, UK, pp169-186

Carey, Susan (2009). *The Origin of Concepts.* Oxford University Press, Oxford, UK

Carruthers, Peter (2008). Meta-cognition in Animals: A Skeptical Look. *Mind & Language*, Vol. 23, no. 1, pp58–89

Carruthers, Peter (2009a). Higher-Order Theories of Consciousness. *The Stanford Encyclopedia of Philosophy* (Fall 2009 Edition), Edward N. Zalta (ed.), forthcoming URL = http://plato.stanford.edu/archives/fall2009/entries/consciousness-higher/

Carruthers, Peter (2009b). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences*, Vol. 32, pp121–182

Carruthers, Peter (2009c). An Architecture for Dual Reasoning. In J.Evans and K.Frankish (eds.), *In Two Minds: dual processes and beyond*. Oxford University Press, Oxford, UK

Carruthers, Peter (2010). Introspection: Divided and Partly Eliminated. *Philosophy and Phenomenological Research.* Vol. 80, no. 1, pp76-111

Carruthers, Peter (2011). Language in Cognition. In E.Margolis, R.Samuels, and S.Stich (eds.), *The Oxford Handbook of Philosophy of Cognitive Science*. Oxford University Press, Oxford, UK

Carruthers, Peter (2012). The Fragmentation of reasoning. In P. Quintanilla (ed.), *La coevolución de mente y lenguaje: Ontogénesis y filogénesis*, Lima: Fondo Editorial de la Pontificia Universidad Católica del Perú

Carruthers, P; Fletcher, L & Ritchie, B (2012). Evolving Self-Consciousness. *4th Consciousness Online Conference*, <http://consciousnessonline.com/2012/01/17/co-4-papers/>

Carruthers, P & Smith, P (Eds.) (1996). *Theories of Theories of Mind*. Cambridge University Press, Cambridge UK

Castañeda, Hector-Neri (1966/2001). 'He': A Study in the Logic of Self-Consciousness. In Brook & DeVidi (Eds.) *Self-Reference and Self-Awareness*, John Benjamins Publishing Company, Amsterdam/Philadelphia, pp 51-79

Castañeda, Hector-Neri (1969/1994). On the Phenomeno-Logic of the I. In Cassam (ed.), *Self-Knowledge*. Oxford University Press, Oxford, UK., pp160-166

Chalmers, David J (1995). Facing Up to the Problems of Consciousness. *Journal of Consciousness Studies*, 2 (3), pp 200-219

Chalmers, David J (1996). *The Conscious Mind.* Oxford University Press, Oxford, UK

Chandler, Michael J & Carpendale, Jeremy (1998). Inching Toward a Mature Theory of Mind. In Ferrari & Sternberg, (eds.) *Self-Awareness*, The Guildford Press, New York, pp148-190

Cheng, Ken (1986). A purely geometric module in the rat's spatial representation. *Cognition*, Vol. 23, pp149-178

Cheng, Ken. (2012). Arthropod navigation: Ants, bees, crabs, spiders finding their way. In E. A. Wasserman & T. R. Zentall (Eds.) *Handbook of Comparative Cognition,* Oxford University Press, Oxford, UK, pp347-365

Chisholm, Roderick M (1969/1994). On the Observability of the Self. In Cassam (ed.), *Self-Knowledge*. Oxford University Press, Oxford, UK., pp94-108

Christie, R. M. (2001). *Colour Chemistry*. The Royal Society of Chemistry, Cambridge, UK

Claxton, Guy (2005). Against Copying: Learning When (and Whom) Not to Ape. In Hurley & Chater, *Perspectives on Imitation, Vol 2: Imitation, Human Development, and Culture*. A Bradford Book, MIT Press, Cambridge Mass. & London, England, pp 199-202

Clayton, N & Dickinson, A (1998). Episodic-like memory during cache recovery by scrub jays. *Nature letters*, Vol. 395, no.17

Clayton, N; Griffiths, D; Emery, N & Dickinson, A (2001). Elements of Episodic-like Memory in Animals. In Baddeley, Aggleton & Conway, *Episodic Memory: New Directions in Research,* Oxford University Press, pp232-248

Clayton, N & Russell, J (2009). Looking for episodic memory in animals and young children: Prospects for a new minimalism. *Neuropsychologia,* Vol. 47, pp2330–2340

Cockburn, David (2001). Memories, Traces and the Significance of the Past. In Hoerl & McCormack (eds.), *Time and Memory*, Clarendon Press, Oxford, UK, pp393-410

Conway, Martin (2001a). Phenomenological Records and the Self-Memory System. In Hoerl & McCormack (eds.), *Time and Memory*, Clarendon Press, Oxford, UK, pp235-256

Conway, Martin (2001b). Sensory-perceptual episodic memory and its context: autobiographical memory. In Baddeley, Aggleton & Conway (eds.), *Episodic Memory: New Directions in Research,* Oxford University Press, pp53-70

Conway, Martin (2008). Exploring Episodic Memory. In Dere, Easton,Nadel, & Huston (eds.), *Handbook of Episodic Memory*, Elsevier, Amsterdam, The Netherlands, pp19-30

Cook, R; Brown, M & Riley, D (1985). Flexible Memory Processing by Rats: Use of Prospective and Retrospective Information in the Radial Maze. *Journal of Experimental Psychology: Animal Behavior Processes*, Vol. 2, no. 3, pp453-469

Correia, S; Dickinson, A & Clayton, N (2007). Western Scrub-Jays Anticipate Future Needs Independently of Their Current Motivational State. *Current Biology* Vol. 17, pp856–861

Crane, Tim (1988). The waterfall illusion. Analysis, Vol. 48, pp142–147

Custance, Deborah & Bard, Kim A (1994). The Comparative and Developmental Study of Self-Recognition and Imitation: The Importance of Social Factors. In Parker, Mitchell & Boccia, (Eds.) *Self-Awareness in Animals and Humans: Developmental Perspectives*. New York: Cambridge University Press, pp207-226

Dally, Joanna; Emery, Nathan & Clayton, Nicola (2010). Avian Theory of Mind and counter espionage by food-caching western scrub-jays (*Aphelocoma californica*). *European Journal of Developmental Psychology*, Vol. 7, no. 1, pp17-37

Davidson, Donald (1987/1994). Knowing One's Own Mind. In Cassam (ed.), *Self-Knowledge*. Oxford University Press, Oxford, UK., pp43-64

Davidson, Donald (1999). The Emergence of Thought. *Erkenntnis*, Vol. 51, pp7-17

Daw, N; Niv, Y & Dayan, P (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, Vol. 8, No. 12, pp1704-1711

Deacon, Terrence (1997). *The Symbolic Species.* Allen Lane, The Penguin Press, London, UK.

Decety, J & Chaminade, T (2005). The Neurophysiology of Imitation and Intersubjectivity. In Hurley & Chater, *Perspectives on Imitation, Vol 1: Mechanisms of Imitation and Imitation in Animals*. A Bradford Book, MIT Press, Cambridge Mass. & London, England, pp119-140

DeLoache, Judy S (2004). Becoming Symbol-Minded. *Trends in Cognitive Science*, Vol 8, no.2, pp66-70

Dennett, Daniel, (1991). *Consciousness Explained*, Little, Brown and Company, Toronto and Boston

Dere, E; Zlomuzica, A; Huston, J & De Souza Silva, M (2008). Animal Episodic Memory. In Dere, Easton, Nadel, & Huston, *Handbook of Episodic Memory*, Elsevier, Amsterdam, The Netherlands, pp 155-184

Dickinson, Anthony (1985). Actions and habits: the development of behavioural autonomy. *PhilosophicalTransactions of the Royal Society of London. Series B, Biological Sciences*. Vol. 308, pp67-78

Dickinson, A & Balleine, B W (2000). Causal Cognition and Goal-Directed Action. In Heyes & Huber, (eds.) *The Evolution of Cognition*. A Bradford Book, MIT Press, Cambridge, Mass. & London, England, pp185-204

Dokic, Jerome (2001). Is Memory Purely Preservative? In Hoerl & McCormack (eds.), *Time and Memory*, Clarendon Press, Oxford, UK, pp213-234

Donald, Merlin (2005). Imitation and Mimesis. In Hurley & Chater (eds.), *Perspectives on Imitation, Vol 2: Imitation, Human Development, and Culture*. A Bradford Book, MIT Press, Cambridge Mass. & London, England, pp283-300

Downing, H & Jeanne, R (1988). Nest construction by the paper wasp, Polistes: a test of stigmergy theory. *Animal Behaviour*, Vol. 36, no. 6, pp1729-1739

Dretske, Fred (1981). *Knowledge & the Flow of Information*. A Bradford Book, MIT Press, Cambridge, Massachusetts

Dretske, Fred (2006a). Perception without Awareness. In Gendler & Hawthorne, (eds.) *Perceptual Experience.* Clarendon Press, Oxford, UK, pp147-180

Dretske, Fred (2006b). Minimal Rationality. In Hurley & Nudds, (eds.) *Rational Animals?* Oxford University Press, London, England, pp107-115

Droege, Paula (2005). Higher-Order Theories of Consciousness. *Internet Encyclopedia of Philosophy*, August 13, 2005, URL=http://www.iep.utm.edu/consc-hi/

Dufour, V & Sterck, EHM (2008). Chimpanzees fail to plan in an exchange task but succeed in a tool-using procedure. *Behavioural Processes* Vol. 79, pp19–27

Easton, Alexander & Eacott, Madeline (2008). A new working definition of episodic memory: replacing 'when' with 'which'. In Dere, Easton, Nadel, & Huston, *Handbook of Episodic Memory*, Elsevier, Amsterdam, The Netherlands, pp185-196

Eichenbaum, Howard (2000). A Cortical-Hippocampal System for Declarative Memory. *Nature Reviews, Neuroscience*, Vol. 1, pp41-50

Eichenbaum, Howard (2003). Learning and Memory: Brain Systems. In Squire, L; Bloom, F; McConnell, S; Roberts, J; Spitzer, N & Zigmond, M. *Fundamental Neuroscience.* Amsterdam/Boston. Academic Press, pp1299-1327

Elsner, Birgit (2005). What Does Infant Imitation Tell Us about the Underlying Representations? In Hurley & Chater, *Perspectives on Imitation, Vol 2: Imitation, Human Development, and Culture*. A Bradford Book, MIT Press, Cambridge Mass. & London, England, pp 191-194

Emery, N. J., & Clayton, N. S. (2001). Effects of experience and social context on prospective caching strategies by scrub jays. *Nature*, Vol. 414, pp443–446

Engelbert, M. & Carruthers, P. (2009) Introspection. In Nadel, L. (ed.) *Wiley Interdisciplinary Reviews: Cognitive Science*, Hoboken, NJ: John Wiley & Sons

Epstein, R; Lanza, R.P & Skinner, B.F (1981). 'Self-Awareness' in the Pigeon. *Science*, Vol. 212, pp695-696

Evans, Gareth (1982). *The Varieties of Reference*. Oxford University Press, Oxford, UK

Feenders G, Smulders TV (2008) Episodic-like memory in foodhoarding birds. In: Dere E, Easton A, Nadel L, Huston JP (eds) *Handbook of behavioral neuroscience,* vol 18. The Netherlands, Elsevier, pp 197–216

Fernandez-Duque, D., J. A. Baird and M. I. Posner (2000). Executive Attention and

Metacognitive Regulation. *Consciousness and Cognition* Vol. 9, no.2, pp288-307

Fields, Douglas (2005). Myelination: An Overlooked Mechanism of Synaptic Plasticity? *The Neuroscientist.* Vol 11, no. 6, pp528-531

Fields, Douglas (2008). White Matter Matters. *Scientific American,* March 2008

Fodor, Jerry (1975). *The Language of Thought*, Crowell, New York

Fodor, Jerry (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. A Bradford book, MIT Press.

Foote, Allison & Crystal, Jonathon (2007). Metacognition in the Rat. *Current Biology*, Vol. 17, no. 6, pp551-555

Frankish, Keith (2009). Systems and levels: Dual-system theories and the personal–subpersonal distinction. In Evans & Frankish (Eds.) *In Two Minds: Dual Processes and Beyond*, Oxford University Press, UK, pp.89-107

Friedman, W (2001). Memory Processes Underlying Humans' Chronological Sense of the Past. In Hoerl & McCormack, *Time and Memory*, Clarendon Press, Oxford, UK, (pp139-168)

Frith, Christopher (1992) *The Cognitive Neuropsychology of Schizophrenia*, Lawrence Erlbaum Associates, Hove (UK) & Hillsdale (USA)

Gallagher, Shaun (1995). Body Schema and Intentionality. In Bermũdez, Marcel & Eilan, (Eds.) *The Body and the Self*. A Bradfird Book, MIT Press, Cambridge, Massachusetts and London UK, pp225-244

Gallagher, Shaun & Zahavi, Dan (2006). Phenomenological Approaches to Self-Consciousness. *The Stanford Encyclopedia of Philosophy (Fall 2006 Edition), Edward N Zalta (ed.),URL*=http://plato.stanford.edu/archives/fall2006/entries/self-consciousness-phenomenological/

Gallese, Vittorio (2005). 'Being Like Me': Self-Other Identity, Mirror Neurons, and Empathy. In Hurley & Chater (eds.), *Perspectives on Imitation, Vol 1: Mechanisms of Imitation and Imitation in Animals*. A Bradford Book, MIT Press, Cambridge Mass. & London, England, pp101-118

Gallup, Gordon. (1970). Chimpanzees: Self-Recognition. *Science* Vol. 167, pp86-87

Gallup, Gordon. (1975). Towards an Operational Definition of Self-Awareness. In Tuttle, (Ed.) *Socioecology and Psychology of Primates*. The Hague, Netherlands: Mouton

Gallup, Gordon. (1994). Research Strategies and Experimental Design. In Parker, Mitchell & Boccia, (Eds.) *Self-Awareness in Animals and Humans: Developmental Perspectives*. New York: Cambridge University Press, pp35-50

Gallup, Gordon. (1997). On the Rise and Fall of Self-Conception in Primates. In Snodgrass &

Thompson, *The Self Across Psychology*, Annals of the New York Academy of Sciences, Volume 818, New York, pp73-83

Gallup, Gordon. (1998a). Can Animals Empathize? Yes. *Scientific American Presents, Exploring Intelligence*, Jan 1998

Gallup, Gordon (1998b). Mirrors and Radical Behaviourism: Reflections on C. M. Heyes. Peer commentary on Heyes (1998), *Behavioral and Brain Sciences*, Vol. 21, no. 1, p119

Gallup, G G, Anderson, J R & Platek, S M. (2011). Self-recognition. In S. Gallagher (Ed.), *The Oxford Handbook of the Self*. New York: Oxford University Press, pp80-110

Gallup, G G, & Povinelli, D J (1993). Mirror, Mirror on the Wall, Which is the Most Heuristic Theory of them All? A Response to Mitchell. *New Ideas in Psychology*, Vol. 11, pp327-335

Gallup, G G & Suarez, S D (1986). Self-awareness and the Emergence of Mind in Humans and Other Primates. In Suls & Greenwald (eds.) *Psychological Perspectives on the Self Vol.3*, Hillsdale, NJ: Erlbaum, pp3-26

Gardiner, J, M (2001). Episodic memory and Autonoetic Consciousness. In Baddeley, Aggleton & Conway (eds.) *Episodic Memory: New Directions in Research*, Oxford University Press, pp11-30

Gattis, M; Bekkering, H & Wohlschläger, W (2002). Goal-Directed Imitation. In Meltzoff & Prinz (eds.), *The Imitative Mind: Development, Evolution, and Brain Bases.* Cambridge University Press, Cambridge, UK, pp183-205

Gennaro, Rocco (1996). *Consciousness and Self-consciousness: A Defense of the Higher-Order Thought Theory of Consciousness.* John Benjamins, Amsterdam Netherlands and Philadelphia USA.

Gennaro, Rocco (2005). Consciousness. *The Internet Encyclopedia of Philosophy*. http://www.iep.utm.edu/consciou/#H6, Oct 2005.

Gennaro, Rocco (2009). Animals, consciousness, and I-thoughts. in *Philosophy of Animal Minds*, Robert Lurz (ed.), Cambridge University Press, pp184-200

Gibson, J, J (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

Gluck, M; Mercado, E & Myers, C (2007). *Learning and Memory: From Brain to Behavior*. Worth Publishers, New York

Goldman, Alvin (2005). Imitation, Mind Reading, and Simulation. In Hurley & Chater (eds.), *Perspectives on Imitation, Vol 2: Imitation, Human Development, and Culture*. A Bradford Book, MIT Press, Cambridge Mass. & London, England, pp79-94

Gopnik, Alison & Meltzoff, Andrew (1994). Minds, Bodies and Persons. In Parker, Mitchell & Boccia, (Eds.) *Self-Awareness in Animals and Humans: Developmental Perspectives*.

New York: Cambridge University Press, pp166-186

Gordon, Robert (2005). Intentional Agents Like Myself. In Hurley & Chater (eds.), *Perspectives on Imitation, Vol 2: Imitation, Human Development, and Culture*. A Bradford Book, MIT Press, Cambridge Mass. & London, England, pp 95-106

Grice, HP (1957). Meaning. *The Philosophical Review*, vol. 66, no. 3, pp. 377-388

Hampton, Robert R (2005). Can Rhesus Monkeys Discriminate Between Remembering and Forgetting? In Terrace & Metcalfe, (eds.) *The Missing Link in Cognition; origins of Self-Reflective Consciousness*. Oxford University Press, Oxford, UK, pp272-295

Hampton, R; Hampstead, B & Murray, E (2005). Rhesus monkeys (*Macaca mulatta*) demonstrate robust memory for what and where, but not when, in an open-field test of memory. *Learning and Motivation,* Vol. 36, pp245–259

Harris, Paul & Want, Stephen (2005). On Learning What Not To Do: The Emergence of Selective Imitation in Tool Use by Young Children. In Hurley & Chater (eds.), *Perspectives on Imitation, Vol 2: Imitation, Human Development, and Culture*. A Bradford Book, MIT Press, Cambridge Mass. & London, England, pp 149-162

Hart, Daniel & Fegley, Suzanne (1994). Social Imitation and the Emergence of a Mental Model of Self. In Parker, Mitchell & Boccia, (eds.) *Self-Awareness in Animals and Humans: Developmental Perspectives*. New York: Cambridge University Press, pp149-165

Hauser, M D; Kralik, J; Botto-Mahan, C; Garrett, M & Oser, J (1995). Self-Recognition in Primates: Phylogeny and the Salience of Species-Typical Features. *Proceedings of the National Academy of Sciences, USA*, Vol. 92, pp10811-10814

Hauser, M; Miller, C; Liu, K & Gupta, R (2001). Cotton-Top Tamarins (Saguinus oedipus) Fail to Show Mirror-Guided Self-Exploration. *American Journal of Primatology,* Vol 53, no. 3, pp131–137

Heck, Richard (2000). Non-conceptual Content and the "Space of Reasons". *The Philosophical Review,* Vol. 109, no. 4, pp483-523

Hempel, C & Oppenheim, P (2008/1965). On the Idea of Emergence. In Bedau & Humphreys (Eds.) *Emergence: Contemporary Readings in Philosophy and Science,* A Bradford Book, MIT Press, Cambridge MA, USA and London UK, pp61-68

Heyes, Cecelia (1994). Reflections on self-recognition in primates. *Animal Behavior.* Vol 47, pp909-919.

Heyes, Cecelia (1998). Theory of Mind in Nonhuman Primates. *Behavioral and Brain Sciences*, Vol. 21, no. 1, pp101-14

Heyes, Cecelia (2005). Imitation by Association. In Hurley & Chater (eds.), *Perspectives on Imitation, Vol 1: Mechanisms of Imitation and Imitation in Animals*. A Bradford Book, MIT Press, Cambridge Mass. & London, England, pp157-176

Hoerl, Christoph (2001). The Phenomenology of Episodic Recall. In Hoerl & McCormack (eds.), *Time and Memory*, Clarendon Press, Oxford, UK, (pp315-337)

Hoerl, Christoph (2011). Causal Reasoning. *Philosophical Studies.* Vol.152, no. 2, pp167-179

Hurley, Susan (2005) The Shared Circuits Hypothesis: A Unified Functional Architecture for Control, Imitation, and Simulation. In Hurley & Chater, *Perspectives on Imitation, Vol 1: Mechanisms of Imitation and Imitation in Animals*. A Bradford Book, MIT Press, Cambridge Mass. & London, England, pp177-194

Hurley, Susan (2006). Making Sense of Animals. In Hurley & Nudds, (eds.) *Rational Animals?* Oxford University Press, London, England, pp139-171

Hurley, Susan & Chater, Nick (eds.) (2005). *Perspectives on Imitation, Vol 1: Mechanisms of Imitation and Imitation in Animals*. A Bradford Book, MIT Press, Cambridge Mass. & London, England

Iacoboni, Marco (2005). Understanding Others: Imitation, Language, and Empathy. In Hurley & Chater (eds.), *Perspectives on Imitation, Vol 1: Mechanisms of Imitation and Imitation in Animals*. A Bradford Book, MIT Press, Cambridge Mass. & London, England, pp 77-100

Jackendoff, Ray (1992). *Languages of the Mind.* A Bradford Book, MIT Press, Cambridge MA, USA & London, UK

Jones, Susan (2005). The Role of Mirror Neurons in Imitation. In Hurley & Chater (eds.), *Perspectives on Imitation, Vol 1: Mechanisms of Imitation and Imitation in Animals*. A Bradford Book, MIT Press, Cambridge Mass. & London, England, pp205-209

Jozefowiez, J; Staddon, JER & Cerutti, DT (2009). Metacognition in animals: how do we know that they know? *Comparative Cognition and Behavior Reviews*, Vol. 4, pp29-39

Kacelnik, Alex, (2006). Meanings of Rationality. In Hurley & Nudds (eds.) *Rational Animals?* Oxford University Press, London, England, pp87-106

Kagan, Jerome (1998). Is There A Self in Infancy? In Ferrari & Sternberg, (eds.) *Self-Awareness*, The Guildford Press, New York, pp137-147

Kim, Jaegwon (2005). *Physicalism, or Something Near Enough*, Princeton University Press, New Jersey, USA and Oxfordshire, UK.

Kinsbourne, Marcel (2005a). Overlapping Brain States While Viewing and Doing. In Hurley & Chater (eds.), *Perspectives on Imitation, Vol 1: Mechanisms of Imitation and Imitation in Animals*. A Bradford Book, MIT Press, Cambridge Mass. & London, England, pp 210-214

Kinsbourne, Marcel (2005b). Imitation as Entrainment: Brain Mechanisms and Social Consequences. In Hurley & Chater (eds.), *Perspectives on Imitation, Vol 2: Imitation, Human Development, and Culture*. A Bradford Book, MIT Press, Cambridge Mass. & London, England, pp 163-172

Kirsch, Irving & Lynn, Steven Jay (1998). Dissociation Theories of Hypnosis. *Psychological Bulletin,* Vol. 123, no. 1, pp100-115

Koechlin, E & Summerfield, C (2007). An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences,* Vol 11, no. 6, pp229-235

de Kort, S; Dickinson, A & Clayton, N (2005). Retrospective cognition by food-caching western scrub-jays. *Learning and Motivation* Vol. 36, pp159–176

Kriegel, Uriah (2004). Consciousness and Self-Consciousness. *The Monist* Vol.87 no.2

Kriegel, Uriah (2007). Self-Consciousness. *The Internet Encyclopedia of Philosophy*, http://www.iep.utm.edu/, Dec 2007

Levine, Jospeh (1983). Materialism and Qualia: The Explanatory Gap. *Pacific Philosophical Quarterly,* Vol. 64, pp354-361

Lewis, Michael (1994). Myself and Me. In Parker, Mitchell & Boccia, (eds.) *Self-Awareness in Animals and Humans: Developmental Perspectives*. New York: Cambridge University Press, pp20-34

Lewis, Michael (1997). The Self in Self-Conscious Emotions. In Snodgrass & Thompson, *The Self Across Psychology*, Annals of the New York Academy of Sciences, Vol. 818, New York, pp119-141

Lewis, M & Brooks-Gunn, J (1979). *Social Cognition and the Acquisition of Self*. Plenum Press, New York and London.

Lind, S & Bowler, D (2010). Episodic Memory and Episodic Future Thinking in Adults With Autism. *Journal of Abnormal Psychology,* Vol. 119, no. 4, pp896-905

MacPhail, Euan (2000). The Search for a Mental Rubicon. In Heyes & Huber, (Eds.) *The Evolution of Cognition*. A Bradford Book, MIT Press, Cambridge, Mass. & London, England, pp253-271

McCormack, Teresa (2001). Attributing Episodic Memory to Animals and Children. In Hoerl & McCormack, *Time and Memory*, Clarendon Press, Oxford, UK, pp285-314

McCormack, T & Hoerl, C (2001). The Child in Time: Temporal Concepts and Self-Consciousness in the Development of Episodic Memory. In Moore & Lemmon, *The Self in Time: Development Perspectives*, Lawrence Erlbaum Associates, New Jersey, USA, pp203-228

McDowell, John (1994). *Mind and World*, Harvard University Press, Cambridge MA, USA

McDowell, John (2009). *Having the World in View: Essays on Kant, Hegel, and Sellars.* Harvard University Press, Cambridge MA, USA and London UK

McNamara, J; Green, F & Olsson, O (2006). Bayes' theorem and its applications in animal behaviour. *OIKOS,* Vol. 112, pp243-251

Marino, L, Reiss, D & Gallup, G G (1994). Mirror Self-Recognition in Bottlenose Dolphins: Implications for Comparative Investigations of Highly Dissimilar Species. In Parker, Mitchell & Boccia, (Eds.) *Self-Awareness in Animals and Humans: Developmental Perspectives*. New York: Cambridge University Press, Ch 25.

Marler, P; Evans, C & Hauser, M (1992). Animal Signals: Motivational, Referential or Both? In Papoušek, H; Jürgens, U & Papoušek, M (eds.) *Nonverbal Communication: Comparative and developmental approaches.* Cambridge University Press, Cambridge, UK.

Marten, K & Psarakos, S (1994). Evidence of Self-Awareness in the Bottlenose Dolphin (Tursiops truncates). In Parker, Mitchell & Boccia, (Eds.) *Self-Awareness in Animals and Humans: Developmental Perspectives*. New York: Cambridge University Press, pp361-379.

Martin, M (2001). Out of the past: Episodic Recall as Retained Acquaintance. In Hoerl & McCormack (eds.), *Time and Memory,* Clarendon Press, Oxford, UK, pp257-284

Mascolo, Michael F & Fischer, Kurt W (1998). The Development of Self Through the Coordination of Component Systems. In Ferrari & Sternberg, (eds.) *Self-Awareness*, The Guildford Press, New York, pp332-385

Mayberry, Rachel (2002). Cognitive development in deaf children: the interface of language and perception in neuropsychology. In Segalowitz and Rapin (eds.) *Handbook of Neuropsychology, 2nd Edition, Vol. 8, Part I,* Ch.4, Elsevier Science BV.

Mellor, D. H. (1988). Crane's Waterfall Illusion. *Analysis,* Vol. 48, pp147-150

Meltzoff, Andrew (2002). Elements of a Developmental Theory of Imitation. In Meltzoff & Prinz (eds.), *The Imitative Mind: Development, Evolution, and Brain Bases.* Cambridge University Press, Cambridge, UK, pp19-41

Meltzoff, Andrew (2005). Imitation and Other Minds: The "Like Me" Hypothesis. In Hurley & Chater (eds.), *Perspectives on Imitation, Vol 2: Imitation, Human Development, and Culture*. A Bradford Book, MIT Press, Cambridge Mass. & London, England, pp 55-78

Millikan, Ruth (2001). The Myth of Mental Indexicals. In Brook and DeVidi, (eds.) *Self-Reference and Self-Awareness*, John Benjamins Publishing Company, Amsterdam/Philadelphia, pp163-177

Millikan, Ruth (2005). Some Reflections on the Theory Theory-Simulation Theory Debate. In Hurley & Chater (eds.), *Perspectives on Imitation, Vol 2: Imitation, Human Development, and Culture*. A Bradford Book, MIT Press, Cambridge Mass. & London, England, pp182-189

Millikan, Ruth (2006). Styles of Rationality. In Hurley & Nudds, (eds.) *Rational Animals?* Oxford University Press, London, England, pp117-126

Mitchell, R. W (1993). Mental Models of Mirror Self-recognition: Two Theories. *New Ideas in Psychology* Vol. 11, pp295-325

Mitchell, R. W (1994). Multiplicities of Self. In Parker, Mitchell & Boccia, (eds.) *Self-Awareness in Animals and Humans: Developmental Perspectives*. New York: Cambridge University Press, pp166-186

Mitchell, R. W (1997a). Kinesthetic-Visual Matching and the Self-Concept as Explanations of Mirror-Self-Recognition. *Journal for the Theory of Social Behaviour*, Vol. 27, no. 1, pp17-39

Mitchell, R. W (1997b). A Comparison of the Self-Awareness and Kinesthetic-Visual Matching Theories of Self-Recognition: Autistic Children and Others. In Snodgrass & Thompson (eds.), *The Self Across Psychology*, Annals of the New York Academy of Sciences, Volume 818, New York, pp39-63

Morin, Alain & DeBlois, Sandra (1989). Gallup's Mirrors: More Than an Operationalization of Self-Awareness in Primates? *Psychological Reports*, Vol.65, pp287-291

Moser, E; Kropff, E & Moser, M-B (2008). Place Cells, Grid Cells, and the Brain's Spatial Representation System. *Annual Review of Neuroscience*, Vol. 31, pp69-89

Mossman, Kaspar (2007). I, Elephant. *Scientific American Mind*, Headlines, Feb/Mar 2007

Mulcahy, N.J. & Call, J. (2006). Apes save tools for future use. *Science* Vol. 312, pp1038–1040.

Murphy, Gregory (2002). *The Big Book of Concepts.* A Bradford Book, The MIT Press, Cambridge MA, USA and London, UK.

Nagel, Thomas (1971). Brain Bisection & the Unity of Consciousness. *Synthese* Vol. 22, pp396-413

Nagel, Thomas (1974). What Is It Like To Be a Bat? *The Philosophical Review*, Vol.83, pp435-450

Neisser, Ulric (1988). Five Kinds of Self-Knowledge. *Philosophical Psychology*, Vol. 1, no. 1, pp35-59

Neisser, Ulric (1997). The Roots of Self-Knowledge: Perceiving Self, It and Thou. In Snodgrass & Thompson (eds.), *The Self Across Psychology*, Annals of the New York Academy of Sciences, Volume 818, New York, pp19-33

Nelkin, Norton (1996). *Consciousness and the origins of thought*. Cambridge University Press, Cambridge UK.

Nelson, Katherine (1997). Finding Oneself in Time. In Snodgrass & Thompson (eds.), *The Self Across Psychology*, Annals of the New York Academy of Sciences, Volume 818, New York, pp103-117

Nelson, Katherine (2005). Emerging Levels of Consciousness in Early Human Development. In Terrace & Metcalfe, (Eds.) *The Missing Link in Cognition; origins of Self-Reflective Consciousness*. Oxford University Press, Oxford, UK, pp116-141

Nichols, Shaun (2001). The Mind's 'I' and the *Theory of Mind's* 'I': Introspection and Two Concepts of Self. *Philosophical Topics,* Vol. 28, pp171-199

Noë, Alva (1999). Thoughts and Experience. *American Philosophical Quarterly*, Vol 36, no. 3, pp 257-265

Nuttley, W; Atkinson-Leadbeater, K & van der Kooy, D (2002). Serotonin mediates food-odor associative learning in the nematode *Caenorhabditis elegans. Proceedings of the National Academy of Sciences,* Vol 99, no. 19, pp12449-12454

O'Brien, Lucy (2007). *Self-Knowing Agents*. Oxford University Press, Oxford, UK

Onishi, Kristine & Baillargeon, Renée (2005). Do 15-Month-Old Infants Understand False Beliefs? *Science*, Vol. 308, no. 5719, pp. 255-258

Parker, Sue Taylor & Milbraith, Constance (1994). Contributions of Imitation and Role-Playing Games to the Construction of Self in Primates. In Parker, Mitchell & Boccia, (eds.) *Self-Awareness in Animals and Humans: Developmental Perspectives*. New York: Cambridge University Press, pp108-128

Parsell, Mitch (2009). Quinean social skills: empirical evidence from eye-gaze against information encapsulation. *Biology and Philosophy*, Vol. 24, no.1, pp1-19

Parsell, Mitch (2011). Sellars on thoughts and beliefs. *Phenomenology and the Cognitive Sciences*, Vol. 10, pp261-275

Patterson, F G & Cohn, R H (1994). Self-Recognition and Self-Awareness in Lowland Gorillas. In Parker, Mitchell & Boccia, (eds.) *Self-Awareness in Animals and Humans: Developmental Perspectives*. New York: Cambridge University Press, pp273-290

Peacocke, Christopher (1992). *A Study of Concepts*. A Bradford Book, MIT Press, Cambridge, Massachusetts and London, England

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Mateo, Calif.:Morgan Kaufmann

Pepperberg, Irene; Garcia, Sean E.; Jackson, Eric C. & Marconi, Sharon. (1995). Mirror Use by African Grey Parrots (Psittacus erithacus). *Journal of Comparative Psychology* Vol. 109, pp182-195

Perner, Josef (1991). *Understanding the Representational Mind*. A Bradford Book, MIT Press, Cambridge, Massachusetts and London, England

Perner, Josef (1996). Simulation as explicationof predication-implicit knowledge about the mind: arguments for a simulation-theory mix. In Carruthers & Smith, (eds.) *Theories of Theories of Mind*. Cambridge University Press, Cambridge UK, pp90-104

Perner, Josef (2001). Episodic Memory: Essential Distinctions and Developmental Implications. In Moore & Lemmon, *The Self in Time: Development Perspectives*, Lawrence Erlbaum Associates, New Jersey, USA, pp181-202

Perry, John (1979/1994). The Problem of the Essential Indexical. In Cassam (ed.), *Self-Knowledge*. Oxford University Press, Oxford, UK., pp167-183

Plotnik, Joshua M, de Waal, Frans B M & Reiss, Diana (2006). Self-Recognition in an Asian Elephant. *Proceedings of the National Academy of Sciences*, Vol. 103, no. 45, pp17053-17057

Povinelli, Daniel J. (1989). Failure to Find Self-Recognition in Asian Elephants (Elephas maximus) in Contrast to Their Use of Mirror Cues to Discover Hidden Food. *Journal of Comparative Psychology* Vol. 103, pp122-131

Povinelli, Daniel J. (1994). How to Create Self-Recognising Gorillas (But Don't Try It On Macaques). In Parker, Mitchell & Boccia, (Eds.) *Self-Awareness in Animals and Humans: Developmental Perspectives*. New York: Cambridge University Press, pp291-300

Povinelli, D J & Cant, J G, (1995). Arboreal Clambering and the Evolution of Self-Conception. *The Quarterly Review of Biology*, Vol. 70, no. 4, pp393-421

Povinelli, D J; Gallup, GG; Eddy, T J; Bierschwale, D T; Engstrom, M C; Perriloux, H K & Toxopeus, I B (1997). Chimpanzees recognize themselves in mirrors. *Animal Behavior*, Vol. 53, pp1083–1088

Povinelli, D J & Prince, Christopher G (1998). When Self Met Other. In Ferrari & Sternberg, (eds.) *Self-Awareness*, The Guildford Press, New York, pp37-107

Premack, D & Woodruff, G (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, Vol 4, pp515-526

Pribram, Karl, H & Bradley, Raymond (1998). The Brain, the Me and the I. In Ferrari & Sternberg, (eds.) *Self-Awareness*, The Guildford Press, New York, pp273-307

Prinz, Jesse J (2005). Imitation and Moral Development. In Hurley & Chater (eds.), *Perspectives on Imitation, Vol 2: Imitation, Human Development, and Culture*. A Bradford Book, MIT Press, Cambridge Mass. & London, England, pp 267-282

Prinz, Wolfgang (2005a). An Ideomotor Approach to Imitation. In Hurley & Chater (eds.), *Perspectives on Imitation, Vol 1: Mechanisms of Imitation and Imitation in Animals*. A Bradford Book, MIT Press, Cambridge Mass. & London, England, pp141-156

Prinz, Wolfgang (2005b). Construing Selves from Others. In Hurley & Chater (eds.), *Perspectives on Imitation, Vol 2: Imitation, Human Development, and Culture*. A Bradford Book, MIT Press, Cambridge Mass. & London, England, pp 180-182

Prior H, Schwarz A & Güntürkün O (2008). Mirror-induced behavior in the magpie (*Pica pica):* Evidence of self-recognition. *PLoS Biol* Vol. 6, no. 8, e202, doi:10.1371 / journal.pbio.0060202

Quirk, G & Beer, J (2006). Prefrontal involvement in the regulation of emotion: convergence of rat and human studies. *Current Opinion in Neurobiology*

Vol. 16, no. 6, pp723-727

Raby, C; Alexis, D; Dickinson, A & Clayton, N (2007). Planning for the future by western scrub-jays. *Nature*, Vol 445, no.22

Rajala AZ, Reininger KR, Lancaster KM, Populin LC (2010). Rhesus Monkeys (Macaca mulatta) Do Recognize Themselves in the Mirror: Implications for the Evolution of Self-Recognition. *PLoS ONE* Vol. 5, no. 9, e12865. doi:10.1371/journal.pone.0012865

Rakison, David (2007). Is Consciousness in its Infancy in Infancy? *Journal of Consciousness Studies*, Vol 14, Nos. 9-10, pp66-89

Rawlins, J (2005). Reflections on Mirror Systems. In Hurley & Chater, *Perspectives on Imitation, Vol 1: Mechanisms of Imitation and Imitation in Animals*. A Bradford Book, MIT Press, Cambridge Mass. & London, England, pp195-199

Reese, Elaine (2009). Development of autobiographical memory: Origins and consequences. In P. Bauer (Ed.), *Advances in Child Development and Behavior*, Vol. 37, The Netherlands: Elsevier, pp145-200.

Reiss, Diana & Marino, Lori. (2001). Mirror self-recognition in the bottlenose dolphin: A case of cognitive convergence. *Procedings of the National Academy of Science* 98:5937-5942.

Rescorla, Robert (1990). The Role of Information About the Response-Outcome Relation in Instrumental Discrimination Learning. *Journal of Experimental Psychology: Animal Behavior Processes*. Vol. 16, No. 3, pp262-270

Richards, William (1984). Self-Consciousness and Agency. *Synthese*, Vol. 61, no. 2, p149

Ristau, C A (1983). Symbols and Indication in Apes and Other Species? Comment on Savage-Rumbaugh et al. *Journal of Experimental Psychology: General*, Vol. 112, pp498-507

Rizzolatti, Giacomo (2005). The Mirror Neuron System and Imitation. In Hurley & Chater, *Perspectives on Imitation, Vol 1: Mechanisms of Imitation and Imitation in Animals*. A Bradford Book, MIT Press, Cambridge Mass. & London, England, pp 55-76

Rizzolatti, G, Fogassi, L & Gallese, V (2006). Mirrors in the Mind. *Scientific American*, Nov 2006

Roberts, William (2002). Are Animals Stuck In Time? *Psychological Bulletin.* Vol. 128, no. 3, pp473-489

Roberts, William (2006). The Questions of Temporal and Spatial Displacements in Animal Cognition. In Wasserman, & Zentall (eds.), *Comparative Cognition: Experimental Explorations of Animal Intelligence*, Oxford University Press, New York, pp145-163

Robertson, E & Köhler, S (2007). Insights from child development on the relationship between episodic and semantic memory. *Neuropsychologia*, Vol.45, pp3178–3189

Rochat, Philippe (2002). Ego Function of Early Imitation. In Meltzoff & Prinz (eds.), *The Imitative Mind: Development, Evolution, and Brain Bases.* Cambridge University Press, Cambridge, UK, pp85-97

Rochat, Philippe (2003). Five Levels of Self-Awareness as they Unfold in Early Life. *Consciousness and Cognition* Vol. 12, pp717-731

Rosenthal, David (1986). Two Concepts of Consciousness. *Philosophical Studies* Vol. 49, pp329-359

Russell, J; Alexis, D & Clayton, N (2010). Episodic future thinking in 3- to 5-year-old children: The ability to think of what will be needed from a different point of view. *Cognition*, Vol.114, no. 1, pp56-71

Ryle, Gilbert (1966/1994). Self-Knowledge. In Cassam (ed.), *Self-Knowledge*. Oxford University Press, Oxford, UK., pp19-42

Sartre, Jean-Paul (1943/2000). *Being and Nothingness*, (translated by Haxel E Barnes), Routledge, Oxon, UK

Savanah, Stephane (2006). Mirrors for Fainting? (Letter to the editor), *Scientific American Mind*, Dec2006/Jan2007

Savanah, Stephane (2012a). The Concept Possession Hypothesis of Self-Consciousness. *Consciousness and Cognition*, Vol 21, no.2 pp713-720

Savanah, Stephane (2012b). Mirror Self-Recognition and Symbol-Mindedness. *Biology and Philosophy*, doi: 10.1007/s10539-012-9318-2

Savastano, H & Miller, R (1998). Time as content in Pavlovian conditioning. *Behavioural Processes* Vol. 44, pp147–162

Schilhab, Theresa S.S, (2004) "What Mirror Self-Recognition in Nonhumans Can Tell Us About Aspects of Self", *Biology and Philosophy* Vol. 19, pp111-126.

Schulman, A, H & Kaplowitz, C (1977). Mirror-Image Response During the First Two Years of Life. *Developmental Psychobiology*, Vol. 10, no. 3, pp133-142

Schwitzgebel, Eric (2011). Belief, *The Stanford Encyclopedia of Philosophy (Winter 2011 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2011/entries/belief/>.

Seager, William (2001). The Constructed and the Secret Self. in Brook & DeVidi, (Eds.) *Self-Reference and Self-Awareness*, John Benjamins Publishing Company, Amsterdam/Philadelphia, pp247-268

Sellars, Wilfrid (1956/1997). *Empiricism and the Philosophy of Mind.* Harvard University Press, Cambridge Mass, USA & London, UK.

Shillito, D J; Gallup, G G & Beck, B B (1999). Factors Affecting Mirror Behaviour in

Western Lowland Gorillas. *Animal Behaviour*, Vol. 57, pp999-1004

Shoemaker, Sydney (1968/2001). Self-Reference and Self-Awareness. In Brook & DeVidi, (Eds.) *Self-Reference and Self-Awareness*, John Benjamins Publishing Company, Amsterdam/Philadelphia, pp81-93

Shoemaker, Sydney (1986/1994). Introspection and the Self. In Cassam (ed.), *Self-Knowledge*. Oxford University Press, Oxford, UK., pp118-139

Siegel, Susanna, "The Contents of Perception", *The Stanford Encyclopedia of Philosophy (Spring 2011 Edition)*, Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/spr2011/entries/perception-contents/

Sloman, Steven (1996). The Empirical Case for Two Systems of Reasoning. *Psychological Bulletin*, Vol. 119, no. 1, pp3-22

Smith, E E & Medin, D L (1981). *Categories and Concepts*. Harvard University Press, Cambridge, Massachusetts and London England.

Smith, J David (2005). Studies of Uncertainty Monitoring and Metacognition in Animals and Humans. In Terrace & Metcalfe, (eds.) *The Missing Link in Cognition; origins of Self-Reflective Consciousness*. Oxford University Press, Oxford, UK, pp242-271

Southgate, V; Senju, A & Csibra, G (2007). Action Anticipation Through Attribution of False Belief by 2-Year-Olds. *Psychological Science*, Vol. 18, no. 7, pp587-592

Spelke, Elizabeth (1994). Initial knowledge: six suggestions. *Cognition*, Vol. 50, pp431-445

Stalnaker, Robert (1998). What Might Non-conceptual Content Be? In Villanueva, E (ed.) *Concepts (Philosophical Issues Series Volume 9)*, Ridgeview Pub Co, Atascadero, California, USA.

Stechler, Gerald (1982). The Dawn of Awareness. *Psychoanalytic Inquiry*, Vol. 1, pp503-532

Sterelny, Kim (2003). *Thought in a Hostile World*. Blackwell Publishing, Malden, MA, USA

Sterelny, Kim (2006). Folk Logic and Animal Rationality. In Hurley & Nudds, (eds.) *Rational Animals?* Oxford University Press, London, England, pp294-311

Sterelny, Kim (2012). *The Evolved Apprentice.* A Bradford Book, MIT Press, Cambridge Mass. & London UK.

Strawson, P F (1994). The First Person – and Others. In Cassam (ed.), *Self-Knowledge*. Oxford University Press, Oxford, UK., pp210-215

Suddendorf, T & Busby, J (2005). Making decisions with the future in mind: Developmental and comparative identification of mental time travel. *Learning and Motivation*, Vol. 36, pp110–125

Suddendorf, T & Corballis M (1997). Mental time travel and the evolution of the human

mind. *Genetic, Social & General Psychology Monographs;* May97, Vol. 123, no. 2, p133

Suddendorf, T & Corballis M (2008). Episodic Memory and Mental Time Travel. In Dere; Easton, Nadel & Huston, *Handbook of Episodic Memory*, Elsevier, Amsterdam, The Netherlands, pp31-42

Sutton, John (2010a). Memory. *The Stanford Encyclopedia of Philosophy (Summer 2010 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2010/entries/memory/>.

Swartz, K B & Evans, S (1994). Social and Cognitive Factors in Chimpanzee and Gorilla Mirror Behaviour and Self-Recognition. In Parker, Mitchell & Boccia, (eds.) *Self-Awareness in Animals and Humans: Developmental Perspectives*. New York: Cambridge University Press, pp189-206

Sweatts, J. David (2010). *Mechanisms of Memory (2nd edition)*. Academic Press (Elsevier), London, Burlington and San Diego

Szpunar, Karl; Watson, Jason & McDemott, Kathleen (2007). Neural Substrates of Envisioning the Future. *Proceedings of The National Academy of Sciences*, vol 104, no.2, pp642-647

Tarsitano, M. S., & Andrew, R. (1999). Scanning and route selection in the jumping spider. *Portia labiata*. *Animal Behaviour,* Vol. 58, pp255–265.

Tarsitano, M. S., & Jackson, R. R. (1997). Araneophagic jumping spiders discriminate between detour routes that do and do not lead to prey. *Animal Behaviour,* Vol. 53, pp257–266.

Thompson, R K R & Contie, C L (1994). Further Reflections on Mirror Usage by Pigeons: Lessons From Winnie-The-Pooh and Pinnochio Too. In Parker, Mitchell & Boccia, (Eds.) *Self-Awareness in Animals and Humans: Developmental Perspectives*. New York: Cambridge University Press, pp392-409

Tomasello, M & Carpenter, M (2005). Intention Reading and Imitative Learning. In Hurley & Chater, *Perspectives on Imitation, Vol 2: Imitation, Human Development, and Culture*. A Bradford Book, MIT Press, Cambridge Mass. & London, England, pp133-148

Tulving, Endel (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), Organisation of memory. New York: Academic Press

Tulving, Endel (1985). Memory and Consciousness. *Canadian Psychology*, Vol. 26, no. 1

Tulving, Endel (2005). Episodic memory and Autonoesis: Uniquely Human? In Terrace & Metcalfe, (eds.) *The Missing Link in Cognition; origins of Self-Reflective Consciousness*. Oxford University Press, Oxford, UK, pp3-56

Tulving, Endel (2007). Are there 256 different kinds of memory? *The Foundations of Remembering*. Psychology Press, East Sussex, UK, pp39-52

Tulving, Endel & Kim, ASN (2009). Autonoetic Consciousness. In Bayne, T; Cleeremans, A; & Wilken, P (Eds). *The Oxford Companion to Consciousness* (pp. 96-98). Oxford University Press

Tye, Michael (1995). *Ten Problems of Consciousness*. A Bradford Book, MIT Press, Cambridge, MASS & London UK

Van Gulick, Robert (1988). A Functionalist Plea for Self-Consciousness. *The Philosophical Review*, Vol. 97, no. 2

de Veer, Monique & van den Bos, Ruud, (1999). A Critical Review of Methodology and Interpretation of Mirror Self-Recognition Research in Nonhuman Primates. *Animal Behaviour*, Vol.58, pp459-468

Voelkl, B & Huber, L (2000). True imitation in marmosets. *Animal Behaviour*, Vol. 60, pp195-202.

Vonk, Jennifer & Povinelli, Daniel (2006). Similarity and Difference in the Conceptual Systems of Primates: The Unobservability Hypothesis. In Wasserman & Zentall (eds.), *Comparative Cognition: Experimental Explorations of Animal Intelligence*, Oxford University Press, New York, pp363-387

deVries, Willem (2011). Wilfrid Sellars. *The Stanford Encyclopedia of Philosophy (Fall 2011 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2011/entries/sellars/>

Walsh, Wendi A & Banaji, Mahzarin R (1997). The Collective Self. In Snodgrass & Thompson, *The Self Across Psychology*, Annals of the New York Academy of Sciences, Volume 818, New York, pp193-215

Weiler, J; Suchan, B & Daum, I (2010). When the future becomes the past: Differences in brain activation patterns for episodic memory and episodic future thinking. *Behavioural Brain Research*, Vol. 212, no. 2, pp196-203

Wellman, HM; Cross, D & Watson, J (2001). Meta-Analysis of Theory-of-Mind Development: The Truth About False Belief. *Child Development*, Vol 72, no. 3, pp655-684

White, J; Southgate, E; Thomson, J & Brenner, S (1986). The Structure of the Nervous System of the Nematode Caenorhabditis elegans. *Philosophical Transctions of the Royal Society of London, B (Biological Sciences).* Vol. 314, no.1165.

Whiten, A; Horner, V & Marchall-Pescini, S (2005). Selective Imitation in Child and Chimpanzee: A Window on the Construcal of Others' Actions. In Hurley & Chater (eds.), *Perspectives on Imitation, Vol 1: Mechanisms of Imitation and Imitation in Animals*. A Bradford Book, MIT Press, Cambridge Mass. & London, England, pp 263-283

Wright, Wayne, (2003). McDowell, demonstrative concepts, and non-conceptual representational content. *Disputatio*, Vol 14. pp37–51.

Zentall, Thomas (2001). The case for a cognitive approach to animal learning and behaviour. *Behavioural Processes* Vol. 54, pp65–78

Zentall, Thomas (2005). Animals may not be stuck in time. *Learning and Motivation*, Vol. 36, pp208–225

Zentall, Thomas (2008). Representing past and future events. In Dere, Easton, Nadel, & Huston, *Handbook of Episodic Memory*, Elsevier, Amsterdam, The Netherlands, pp217-237

Zentall, T & Akins, C (2001). Imitation In Animals: Evidence, Function and Mechanisms. In R G Cook (Ed.), *Avian Visual Cognition* [On-line]: www.pigeon.psy.tufts.edu/avc/zentall/