Proper Embodiment: The Role of the Body in Affect and Cognition

Mog Stapleton

PhD in Philosophy

The University of Edinburgh

2011

Declaration of authorship

I, Mog Stapleton, declare that this thesis is composed by me and that all the work herein is my own, unless explicitly attributed to others. This work has not been submitted for any other degree or professional qualification.

Mog Stapleton, 31st August 2011

Parts of this thesis have been published, or are forthcoming (with the support of my supervisor) in the following articles:

Invited chapter on Emotion (translated by Marraffa) for Marraffa, M. & Paternoster, A. (Eds) (2011) *Scienze cognitive. Un'introduzione filosofica* (Cognitive Sciences. A Philosophical Introduction), Carocci, Italy.

"Feeling the Strain: Predicting the Third Dimension of Core Affect". Invited commentary on Lindquist, Wager, Kober, Bliss-Moreau, and Barrett. "The brain basis of emotion: A meta-analytic review". (Forthcoming). *Behavioral and Brain Sciences.*

In addition, sections of chapters 3 & 4 will appear in the following articles which have been accepted (subject to revisions):

"Steps to a Properly Embodied Cognitive Science". Review paper for *Cognitive Systems Research*. Accepted, May 2011 (subject to revision).

"Es are good: Cognition as Enacted, Embodied, Embedded, Affective and Extended" (with Dave Ward). In Paglieri, F. & Castelfranchi, C. (Eds). (Forthcoming). *Consciousness in interaction: The role of the natural and social environment in shaping consciousness.* John Benjamins' *Advances in Consciousness Research*. Accepted, July 2011 (subject to revision).

Abstract

Embodied cognitive science has argued that cognition is embodied principally in virtue of gross morphological and sensorimotor features. This thesis argues that cognition is also *internally embodied* in affective and fine-grained physiological features whose transformative roles remain mostly unnoticed in contemporary cognitive science. I call this 'proper embodiment'. I approach this larger subject by examining various emotion theories in philosophy and psychology. These tend to emphasise one of the many gross components of emotional processes, such as 'feeling' or 'judgement' to the detriment of the others, often leading to an artificial emotion-cognition distinction even within emotion science itself. Attempts to reconcile this by putting the gross components back together, such as Jesse Prinz's "embodied appraisal theory", are, I argue, destined to failure because the vernacular concept of emotion which is used as the explanandum is not a natural kind and is not amenable to scientific explication.

I examine Antonio Damasio's proposal that emotion is involved in paradigmatic 'cognitive' processing such as rational decision making, and argue (1) that the research he discusses does not warrant the particular hypothesis he favours, and (2) that Damasio's account, though in many ways a step in the right direction, nonetheless continues to endorse a framework which sees affect and cognition as separate (though now highly interacting) faculties. I further argue that the conflation of 'affect' and 'emotion' may be the source of some confusion in emotion theory and that affect needs to be properly distinguished from 'emotion'. I examine some dissociations in the pain literature which give us further empirical evidence that, as with the emotions, affect is a distinct component along with more cognitive elements of pain. I then argue that affect is distinctive in being grounded in homeostatic regulative activity in the body proper.

With the distinction between affect, emotion, and cognition in hand, and the associated grounding of affect in bodily activity, I then survey evidence that bodily affect is also involved in perception and in paradigmatic cognitive processes such as attention and executive function. I argue that this relation is not 'merely' casual. Instead, affect (grounded in fine-grained details of internal bodily activity) is partially constitutive of cognition, participating in cognitive processing and contributing to perceptual and cognitive phenomenology. Finally I review some work in evolutionary robotics which reaches a similar conclusion, suggesting that the particular fine details of embodiment, such as molecular signalling between both neural and somatic cells matters to cognition. I conclude that cognition is 'properly embodied' in that it is partially constituted by the many fine-grained bodily processes involved in affect (as demonstrated in the thesis) and plausibly by a wide variety of other fine-grained bodily processes that likewise tend to escape the net of contemporary cognitive science.

Table of contents

Acknowledgements

This thesis and my intellectual development over the course of my PhD has emerged as a result of a coupling between myself and several people who have helped me shape ideas and entertain new ways of thinking and who have put me on new paths of discovery. My principal supervisor Andy Clark has been key to this throughout as well as giving me the opportunity to be part of the CONTACT project. I am deeply grateful to him for his guidance, encouragement, and support over the last few years. My secondary supervisor Julian Kiverstein has also been a critical and supporting figure giving me detailed feedback on several of the chapters, I am very grateful to him. Some of the work in this thesis has been presented at doctoral group workshops and I thank all those in the DoG group for their comments and suggestions and for providing such a fantastic forum in which to try out new work.

During my research visit to Toronto in 2008, Evan Thompson was extremely generous with his time and was a brilliant source of help in forming my ideas and directing me towards research which helped me shape my continuing development and this has underpinned the shape of my research and thinking. I am very grateful to him for this. This time in Toronto also afforded me the opportunity to attend Marc Lewis' neuroscience classes which set me off on a journey of neuroscientific discovery for which I thank him. Before coming to Edinburgh I was lucky to be part of Ron Chrisley's 'Philosophy of AI and Cognitive Science' research group at the University of Sussex. This period provided a lot of opportunities for trying out ideas and broadening the scope of my research, and I thank Ron and everyone in the PAICS group for my time there.

As the thesis neared its final stages Matt Nudds generously gave his time to read and comment upon a full draft. And, Dave Ward has been outstandingly valiant in slogging through the last few weeks with me, providing detailed comments, suggestions for improvement and relentless encouragement. I thank both for their generous help and commitment; the thesis, and I, are much the better for it. I would also like to thank Suilin Lavelle who kindly read through and commented upon a couple of chapters.

I thank my parents, my sisters and (now sadly deceased) grandparents for always supporting me in my academic endeavours. I wish that my grandfather Ben Dunkey was here to see the finished product. I would like to give a special shout-out to my flat mates: Olle Blomberg, Dohyoung Kim, and Eusebio Wawaru for gallantly putting up with me over the last year - remaining patient with me even during the final crazed months. Finally, I want to thank all my friends here in Edinburgh and those far away but who have provided support over email and Facebook. I especially want to thank Alisa Mandrigin, SJ Stapleton, and Dave Ward for being my friends and keeping me going this last year. Thank you.

Acknowledgements are due to the AHRC who funded this PhD research through the ESF Consciousness in a Natural and Cultural Context. My scholarship was through the CONTACT (Consciousness in Interaction) node of this project, PI Prof. Andy Clark & Prof. Susan Hurley.

Introduction

The term 'embodied cognition' is used to cover a diverse range of theses and methodologies for research in cognitive science. They are united in claiming that the body matters, in some way, to cognition. This is in contrast to orthodox cognitive science which was essentially disembodied; the body was merely considered to be a vehicle for inputs and outputs to and from the central nervous system (or whatever provided that function in the system of interest), which was where all the information processing relevant to cognition was deemed to take place. Against this, standard embodied cognitive science (for example Brooks 1991; Clark 1997) argues that gross morphological features can play a computational role in cognition. The basic idea is that some of the computational work essential to cognition can be partially offloaded to, and realised by, bodily processes and structures (physical gestures are an oft-cited example (see Clark (2008); Goldin-Meadow (2003)) external to the central nervous system. Cognition is thus extended so that it encompasses parts of the body, and also those parts of the non-biological world that support the appropriate offloading of computations. However, this means that as far as standard embodied cognitive science is concerned, the body *qua body* does not play a special role; only the body *in virtue of its ability to be a vehicle of computations*. The result is that, although research in this paradigm is based on the role of the body in cognition, the body really isn't the important factor.

This thesis argues that the body plays a much more fundamental role in cognition than that. It is not only the morphological and sensorimotor capacities of the gross physical body that are of relevance to cognition, but also the very fine details of our embodiment, including what I will dub 'internal embodiment'. In particular, I argue that affective information, both in the form of neural information from the body proper and in the form of molecular signaling between neurons and perhaps even non-neuronal cells, contributes to cognition in a profoundly important way and cannot be simply 'factored out' if our goal is a proper understanding of what makes us the flexible and adaptive cognitive creatures we are. I thus argue that the particular details of our embodiment matter in a way that goes significantly beyond the ways normally cited in standard 'embodied cognitive science'. I call this the thesis of 'proper embodiment'.

The structure of my treatment runs as follows. The goal of the first two chapters is to look at particular theories of emotion that have been influential in philosophy of cognitive science.

There we will see that (1) even supposedly embodied theories of emotion and cognition present the role of the body in such a way that it is relatively easy to factor it out, and (2) that (partly as a result of this) there is an implicit and theoretically misguided emotion-cognition distinction inherent in theories which purport to be reconciling emotion, cognition, and the body. Chapters three, four and five, display the bodily and affective components of the above theories of emotion and cognition, and argue that affect, grounded in interoception and the fine details of our physiological functioning, is crucially involved in cognition: involved, moreover, to an extent that it is plausibly considered as partially constitutive of cognition. I conclude that cognition is embodied in a much more fundamental way than that proposed by either standard embodied cognitive science, or the embodied theories of emotion outlined in the first two chapters.

In a little more detail, the aim of chapter one is to display some of the central landscape of emotion theories in philosophy and psychology, including a more detailed look at some aspects of Jesse Prinz's (2004) embodied appraisal theory. I rehearse some brief background concerning the distinction between feeling and appraisal theories in the history of emotion psychology and argue that although Prinz advertises his theory as a "grand reconciliation" between the two, it actually falls short of providing one. This is because contemporary appraisal theories in psychology are already as, or more, 'embodied' than his model. I also argue that, in the light of recent work by Griffiths and Scarantino on the role of the kind 'emotion' in scientific psychology, Prinz's model is actually best seen as merely a *descriptive* theory rather than a proper *explication* of emotion.

Where Prinz tries to reconcile feeling, bodily changes, and cognition in an account of the emotions, another major recent model of emotion, due to Antonio Damasio (1996), argues that emotion is crucially involved in cognition. Damasio's focus, however, is upon the role of emotion in relation to the paradigmatically cognitive capacity of decision-making. In the second chapter I focus on this hypothesis which is often cited as neural evidence for embodied emotion playing a role in cognition. But I ultimately reject it as unhelpful - and in some ways fundamentally misguided - as a model of understanding the relation between affect and cognition. I suggest that the particular way in which Damasio argues that embodied emotion is necessary for decision-making perpetuates a bias which is unwarranted by the best contemporary neuroscience: a bias that persists in treating emotion and cognition as essentially separate, though interanimated, faculties. I outline several possible alternative explanations for the results from the Iowa gambling task experiments, including

perseveration and deficits in reversal learning, which do not require appealing to such separate faculties of emotion and cognition. In so doing, I both undermine a key source of support for Damasio's hypothesis, and lay the groundwork for the more satisfactory integration of emotion and cognition to be attempted in subsequent chapters.

The models discussed in chapters 1 & 2 are vulnerable to a variety of criticisms (including those of Griffiths and Scarantino outlined in chapter 1) because they assume emotion to be a unified category. I follow Griffiths and Scarantino in rejecting this and hence must seek an alternative strategy for investigating the relations between the various components of affective phenomena. In chapter 3, I show that a detailed exploration of dissociations between feeling, sensation, and behaviour in pain pathologies shows that while responses to affective stimuli incorporate valence (in this case the feeling of unpleasantness), valence may also stand alone as a component in its own right. This sets the stage for investigating bodily affect (and the corresponding experience of affective feeling) apart from the other components of affective phenomena. Through consideration of various cases, I then show that in cases where valence is dissociated from the emotional component - and so there is no representation of threat to guide a behavioural response - valence on its own may guide a particular, and different, behavioural response (for which I provide an explanation in the third part of the chapter). In the second part of the chapter, I begin to explore the neurobiology of affective states, showing how they are grounded in information from the internal body (interoception). I review research which shows that primitive structures in the brainstem and diencephalon (which subserve the regulation of the internal body) might support a basic level of consciousness and interaction with the environment. I show how these 'primordial-' or 'homeostatic emotions' incorporate motor response and, in the third part of the chapter, argue that this is what grounds the behavioural aspect of valence, comprising a "minimal appraisal mechanism".

Having thus (I hope) grounded affect in interoceptive information, chapter 4 goes on to explore several models which propose that affective information plays a crucial role both in exteroceptive perception and in cognitive processes more generally. Here I discuss emerging work in neuroscience suggesting that top-down information can influence processing all the way down, even to responses associated with early vision. With the possibility of previously processed information systematically impacting even such early perceptual processing, the stage is set for investigating the possible role of affective information in processing that was previously thought to be purely perceptual or cognitive. I first review a model of perceptual

processing which centrally involves the integration of affective information. I argue that this model provides empirical support for the phenomenological claim that affect structures our perceptual phenomenology and is plausibly a core ingredient of our mental lives more generally. Finally, I review some recent arguments from neuroscience that converge on the idea that 'emotion' processing cannot be coherently distinguished from 'cognitive' processing, which provide another (partially independent) way of showing that affective information pervades throughout all processing.

In chapter five I step back a little, to ask just what *kind* of a claim should be made about the relation between affect and cognition given the work that I have presented earlier in the thesis. I outline a range of possible problems attending the popular distinction between 'cause' and 'constitution' when dealing with putative scientific explanations of mechanisms. I then explore some possible alternative frameworks that involve other (less metaphysically loaded) notions such as explanatory inseparability and the 'difference/contribution' distinction. I argue that, given the body of work outlined in the previous chapters, affect understood in terms of afferent or efferent homeostatic information is not usefully to be distinguished from cognitive processing.

There is another form of affective information, however, which might indeed be thought to be "merely modulatory" rather than genuinely constructive of cognition: molecular signalling. But here too, the attempt to parcel out the affective contribution fails. To show this, I present work from evolutionary robotics which strongly suggests that non-neural information processing exploiting area effects resulting from the diffusion of (virtual) gases, makes an integral contribution to evolved solutions to cognitive problems. I argue that likewise, many of the components of our specific physiology cannot be isolated as mere background conditions for cognitive processing but instead play a key constructive role.

The model which I defend diverges considerably from what might be considered 'standard opinion' in mainstream cognitive science. However, I claim that it is grounded in compelling current work in neuroscience and artificial cognitive systems research that will inform the shape of much future cognitive science. If I am right, affect, grounded in the fine internal details of human physiology, lives right at the core of cognition. This delivers a 'properly embodied' model of cognition.

Chapter I

Emotion theories and the role of the body

Section 1: Theories of Emotion

Emotion theories are (unsurprisingly) typically theories of *emotions*. They try to capture what the emotions have in common and how they are different from other mental states or processes such as rational decision making. Emotions seem to have several aspects in common in virtue of which we generally think of them as forming the category 'emotion'. These aspects include the phenomenological feeling; the 'what it is like-ness' to be happy or sad. There are also the particular behaviours or dispositions towards behaviours that we experience when in the throes of an emotion. Fear, for example might encompass a disposition to avoid that which is fearful and without this avoidance disposition one might question whether you were truly 'scared'. Wanting to avoid something that is fearful of course requires that we have implicitly or explicitly judged or appraised something *as* fearful. This may in turn require beliefs about the object of emotion. Finally emotions seem to be concerned with value, we are moved by objects and situations that have value to us and our personal projects, I do not fear rejection if I do not value being included in a social group in the first place, and I do not fear being mauled by a vicious dog if I do not value my health and life.

Different theories of emotion (or rather 'the emotions') tend to emphasise one or other of these aspects of emotion at the expense of the others and sometimes identify the category 'emotion' with that privileged aspect. Feeling theories for example take emotions to be the feelings that we feel when in an emotional state. This captures the passivity that often seems inherent in emotion; we feel 'moved', and often intensely 'affected' by emotional stimuli. The most influential version of this type of theory in recent psychology is the James-Lange 'feeling theory' which identifies emotion with the feeling of bodily changes. At the other end of the spectrum are cognitive theories which identify emotions with particular judgements, judgements about value and how one is faring in the world (see for example Nussbaum, 2004). Appraisal theories sit somewhere in the middle. While appraisals can be thought of as judgements, typically in contemporary psychology they are thought of in more subpersonal

terms (see Arnold 1960; Scherer 1999). Appraisal theories are typically categorised as cognitive however, because they emphasise the importance of appraising a situation over the feeling aspects of the emotion. That is, in order to distinguish between emotions they point to differences in the appraisals that typically elicit that emotion, rather than differences in the 'feeling' states in those emotions. In this chapter I look at these theories in a little more detail and try to discern and display the (differing) roles of the body in each.

1.1 Feeling theories

William James and Karl Lange famously argued that emotion is the feeling of bodily changes. More precisely James argued that "the bodily changes follow directly the *perception* of the exciting fact, and that our feeling of the same changes as they occur *is* the emotion" (James 1884, original emphases). This model contrasts with the pre-Jamesian model of emotion which took the feeling state to be prior to the physiological responses (Scherer, 2000). According to the James-Lange model when we see a bear in the woods changes happen in our peripheral nervous systems, both autonomic and somatic. Most likely these changes will be things like upregulation of the sympathetic nervous system so the body is in a high state of arousal (with for example the heart pumping faster, and activation of processes in the somatic nervous system preparing the muscles for action).

Pre-James:
Perception of event -> emotion/feeling -> physiological and action responses

James:
Perception of event -> physiological and action responses -> emotion/feeling

The feeling of the changes that occur in our body may be better thought of as 'affect'. Indeed Scherer (2000) has argued that James conflated 'emotion' with 'affect' and that this confusion has been the primary source of the dispute between feeling theorists and cognitive and appraisal theorists. The confusion occurs because 'emotion' and 'affect' are often used interchangeably to refer to the entire category of emotions as well as a particular feeling. While emotions clearly are affective states (or processes) we must of course be careful about identifying the two; there are affective states that we might not want to think of as strictly emotions, such as pain and hunger (among others). 'Emotion' and 'affect' are also used interchangeably to refer to the feeling of bodily changes as well as to the category of

emotions. However using 'emotion' in this way risks conflating it with the category of emotions (as happened subsequent to James' usage). Similarly, 'affect' has come to refer to the physiological changes themselves rather than just the feeling of those changes.

Feeling theorists of emotion may accept that 'emotion' in this context means affect (in the sense of 'feeling of bodily changes'), and yet still argue for a feeling theory of emotion. Thus one might distinguish one emotion from another in virtue of the perception of a particular pattern of bodily changes even if the emotions themselves are not identified solely with the particular feeling state but also encompass action tendencies and cognitive or appraisal elements. Cognitivists argue that this is not the case and that the very same physiological feelings can be interpreted in different ways to yield different emotions.

1.2 Cognitivism

Feeling theories fit naturally with our common-sense understanding of emotions. But with the advent of behaviourism phenomenological feelings became unacceptable both as explanandum and explanans. The behaviourists saw emotions as conditioned responses built upon a primary stimulus-response mechanism present at birth. Cognitivism arose in part as a reaction against the behaviourist disregard for internal mental mechanisms as explanans. It became the orthodox theoretical basis for emotion theories, and psychology in general, for most of the latter half of the 20th century and remains hugely influential to this day.

In contrast to the behaviourists identification of emotions with a particular type of conditioned response, cognitivists take emotions to be representations of the stimulus, and so often identify them with evaluative judgments. Taking emotions to be evaluative judgments fits neatly with the propositional attitude psychology that dominated mid-20th century analytic philosophy of mind, as these judgments are about states of affairs in respect to the subject; just as a belief or a desire is a mental state which is about something (its content), an emotion is assumed to be a mental state which is about (represents) some real or imagined state of the world. Cognitivists thus hold that an emotional episode may involve physiological changes but that it is the attendant appraisal which allows us to discriminate one emotion from the other. Without this 'labeling' effect, the physiological changes, or feeling of those changes do not constitute (or bring about) an emotion.

The paradigmatic study that cognitivists appeal to as showing that the feelings of physiological changes are not sufficient to elicit and discriminate between emotions is due to Schachter and Singer (1962). Schachter and Singer ran an experiment in which a group of subjects were injected with epinephrine (adrenaline), which stimulates arousal of the sympathetic nervous system. The subjects were then made to wait in a room with other participants one of whom was a false participant (a confederate). The confederate acted either in a euphoric, clown-like, or irritable fashion, and the participants responses to this were monitored. The subjects reacted to the confederates' behaviour with a matching response and Schachter and Singer took this to be evidence that the different contexts caused the (same) aroused physiological state to be interpreted in different ways showing that it is the interpretation of an event which is the relevant factor in discriminating emotions rather than the physiological changes or feelings of those changes.

Schachter:
Perception of event -> physiological arousal -> cognitive explanation of arousal based on context -> emotion/feeling

This experiment has however been the subject of much controversy. One of the reasons for this is that Schachter and Singer drew their conclusions from the behaviour of the subjects and not from the self-report questionnaire that followed in which most of the participants in both conditions (including the anger inducing condition) reported that they were in a good mood (Prinz, 2004, p. 71). Schachter and Singer put this down to the subjects wanting to please the experimenters. However, they did not consider that as well adapted social creatures we have techniques of mirroring and coordinating emotional behaviour which increase our empathic communicative abilities and ease social discomfort in novel situations. Perhaps, then, the subjects of the experiment may not have actually "had" the emotion that the observations of the behaviour made it seem that they had. Rather they may have been mirroring the emotion for the purposes of social cohesion. (Of course it might be that mirroring is enough to bring about the actual emotion: studies such as those in which a subject's mouth is forced into a smile by having to bite a pencil show that the self report of the emotional state is then more positive (see for example Soussigan, 2002)). Secondly the authors did not seem to consider that the arousal caused by the dose of epinephrine may have merely primed the body for reaction to stimuli (provided by the confederate) and therefore made easier the shift into the physiognomic changes relevant for each emotion (Prinz, 2004).

If this were the case then emotions could still be constituted by the pattern of bodily changes rather than the appraisal aspect itself,

1.3 Appraisal theories

Appraisal theory began (Arnold, 1960) by emphasizing the importance of 'significance evaluations' of events to explain the elicitation of different emotions. This differed from cognitivist theories, such as the Schachter-Singer approach, as these focussed on the use of context to create an emotion rather than thinking in terms of evaluations which are significant to the emoter and which *elicit* an emotion. Arnold proposed that events were appraised with respect to a number of dimensions, such as whether those events were beneficial or harmful; the presence or absence of some object; and the relative difficulty of approaching or avoiding that object (Scherer, 1999, p.637). What can be seen by this initial characterization of the theory is that there is no claim that emotions "are" appraisals. As Scherer (1999) explains in his review of appraisal theory:

> A central tenet of appraisal theory is the claim that emotions are elicited and differentiated on the basis of a person's subjective evaluation or appraisal of the personal level significance of a situation, object, or event on a number of dimensions or criteria. (Scherer, 1999, p. 637)

Appraisal theories claim only that appraisals are elicitors of emotions. They do not claim that appraisals cannot be by-passed, and emotions induced by other means. Rather, they see appraisal theory as the best tool available for the purposes of prediction and the differentiation of emotions. In this respect appraisal theory is open to accusations of excessive cognitivism as the evaluations that these theorists appeal to as elicitors of emotion seem on the face of it as if they are deliberate and conscious and perhaps even to be identified with 'cortical' (as opposed to 'limbic') processing. This is the approach that the "meaning" strand of appraisal theories took. These appraisal theorists are characterized by their focus on the "analysis of the propositional nature of the semantic fields that underlie the use of specific emotion terms, almost in the sense of definitions" (Scherer, 1999, p. 640). Scarantino (2005) notes that this kind of cognitivism has had a great influence in the philosophy of emotion under the auspices of propositional attitude psychology, in particular in the works of Broad (1954), Kenny (1963), and Solomon (1976). These philosophers brought the problem of intentionality to the table, supporting cognitivist theories with the assertion that emotions could not be identified with feelings as feelings lack an intentional object whereas emotions are always about something (Broad 1954).

We should note however that this is very much a philosophical strand and was by no means characteristic of appraisal theories in general.

Scherer (1999) outlines three other major historical theoretical approaches to appraisal theory, the main players in which are all psychologists and not philosophers. The other theoretical strands that Scherer runs through are Criteria, Attributions, and Themes. In the Criteria approach it is postulated that a fixed set of dimensions/criteria are used in evaluating the significance of events (see for example, Arnold 1960). The Attribution approach (which mainly spanned the early-to-mid 1980's) is characterised by the focus on the nature of the causal attributions involved in emotion-antecedent appraisals. That is, they posited that one could distinguish a number of emotions such as anger, pride and shame, on the basis of either internal or external attribution of responsibility. The Themes approach attempted to link the elicitation of specific emotions to the identification of certain themes (specific patterns of goal-relatedness) of events. The key player here is Lazarus (1991) whose "core-relational themes" approach will be discussed later on (in relation to Prinz's theory).

All these strands fed into modern appraisal theories which can be characterized as proposing that *the core process of emotional reaction is one of significance evaluation*. Appraisal theorists differ as to the number and definition of appraisal dimensions that they propose (Scherer, 1999, p. 642), less so as to which responses are elicited by the appraisals (i.e., what constitutes a full-blown emotion), and they vary as to whether emotional responses and appraisals are static or dynamic, and whether uni- or multi-level processing is involved. Since the 1980's appraisal has been understood to occur at low-levels of the nervous system (Scherer 1999), and Lazarus has argued that much of appraisal can be unconscious (Scherer 1999, p. 642). The identification of current appraisal theories with cognitivist models of emotion (such as propositional attitude psychology) is therefore unwarranted. This can be seen in particular with multi-level theories of emotion such as (Barnard, 1985; Teasdale & Barnard (1993); Teasdale 1996, 1997; Power & Dallies, 1997) which propose that the "schematic model level" is what generates emotion. This level processes patterns of "recurring themes and regularities extracted from the patterns of propositional and sensory codes synthesized in previous situations that have elicited a given emotion" (Teasdale, 1999, p. 670). It is the "implicational code" which represents these patterns which is processed and gives rise to the emotion. Thus both sensory and cognitive contributions can be "expressed, integrated, and can modulate the production of emotion" (ibid) and either sensory *or* cognitive contribution is sufficient.

Scherer argues that the "emotion-cognition controversy" in emotion theories has been a debate about the minimal cognitive prerequisites for emotion and that this dissolves if one accepts that (1) appraisal theory is generally trying only to explain full-blown emotions (and not other affective states such as sensations or moods), and (2) that appraisals can be low-level and automatic, potentially being subserved by subcortical pathways (Scherer, 1999, p. 645). Indeed we can see just how low-level these appraisals may be, and how unjustified the charge that appraisals are to be thought of as deliberative, in the following passage:

> … one can argue that we need a general, overarching term to cover the fundamental fact that it is not the objective nature of a stimulus but the organism's "evaluation" of it that determines the nature of the ensuing emotion. A completely automatic, reflexive defense reaction of the organism also constitutes an intrinsic assessment, a valuation, of the noxiousness of the stimulus (although it may not necessarily produce a full fledged emotion […]). Even if simple feature detection is involved, the outcome of the process constitutes an assessment of the significance of the detected stimulus to the organisms, given that feature detectors that have any behavioral consequences are automatically "significance detectors". (Scherer, 1999, p. 647)

As Scherer notes, such low-level appraisals of significance are clearly a quite different processes from deliberate appraisals of significance and may result in different emotions and action tendencies, however it is not clear that they aren't underpinned by shared "functional-adaptational aspects" and should therefore not both be treated as significance evaluations (appraisals). Despite some opposition to viewing appraisals as low-level, mainly from the philosophical, "meaning" branch of emotion theory, this is now pervasive in modern appraisal theory, which itself is viewed as the standard theory of emotion in psychology today.

Scherer's own preferred version of appraisal theory is the 'component model' theory of emotion which takes feelings, appraisals, bodily changes, action tendencies, and facial and vocal expressions to all be components of emotion processes (Scherer, 2005). Key to his theory is that emotions are episodes, rather than static states. These components thus all feed in to each other in the course of an emotion episode, and emotion is not to be reduced to any one component, though the emotion may be driven by (and differentiated by) one of these components, such as intrinsic appraisal (evaluating features of the stimulus in terms of genetic or learned preferences) or transactional appraisal (evaluating features of the stimulus in terms of "their conduciveness for salient needs, desires, or goals of the appraiser") (Scherer, 2005, p. 701).

Modern (appraisal and component models):
Perception and appraisal of event -> physiological, expressive and motivational changes -> reflection of these changes in a monitoring system -> changes in feeling state -> feeds in to perception and appraisal of event… [carries on recursively].

1.4 The concepts and categories of emotion and cognition

In everyday language we use the terms cognition and emotion to identify complex traits that the kind of complex creatures that cognitive scientists are interested in possess. On the face of it, it seems that it is our abilities to cognize and to emote that distinguishes us from simpler organisms such as plants and bacteria, and it seems that it is one or other of these (or both working in a symbiotic relationship) that allow us to be the flexible, adaptive creatures that we are. But are the vernacular concepts of emotion and cognition useful to the project of explanation and induction? Or, if we really want to understand what emotion and cognition really are might we have to give up the vernacular categories? Griffiths (1997) argues that the vernacular concept of emotion does not correspond to a natural kind under his general theory of natural kinds, which requires that a category must have causal homeostasis. A category has causal homeostasis if the set of correlated properties have some underlying explanation that makes it projectable (Griffiths 1997, p. 187). This underlying explanation might be in the form of a particular mechanism. If the same mechanism underlies all the members of a category then it is mechanically projectable; we can include x in the category iff it is also realized by the same mechanism.

Griffiths argues that the category of emotions to which the vernacular concept refers includes at least two subcategories; affect programs and higher cognitive emotions. Affect programs are neural and bodily changes which are "complex, coordinated, and automated" and together are taken to constitute an emotional response. Examples of these changes are expressive vocal and facial changes, musculoskeletal responses, endocrine system changes, and autonomic nervous system changes. Higher cognitive emotions on the other hand are what might be thought of as more social emotions, such as envy, guilt, jealousy and love. As there is no underlying explanation for the set of properties that are included in both affect programs and higher cognitive emotions which could allow the category 'emotion' to be projectable, Griffiths argues that it is of no use in a scientific psychology of emotion. To cling to the vernacular concept by trying to identify it with just a subset of the vernacular

category, e.g. trying to identify anger with a particular affect program, will not be adequate as it will result in losing some of the vernacular concept. Given that the vernacular concept 'emotion' does not pick out a natural kind (where a natural kind is identified as having projectable mechanisms) it is not useful for scientific explanation.

Griffiths states that "[t]he development of a scientific psychology of emotion requires emotion concepts to be refined or replaced so that the categories corresponding to emotion concepts have strong causal homeostasis. Such categories will be projectable and therefore useful for explanation and induction" (Griffiths, 1997, p.228). He requires causal homeostasis at all three levels of explanation; task description/ecological, computational and implementational (Sterelny, 1990)[1]. The ecological approach to emotions, which takes emotions to be responses which are classified "according to their presumed adaptive function" (p. 231), comes very close to the vernacular concept of emotions (or at least the philosophical conception of the vernacular concept) as it embraces the organism-environment relations which the emotion has as its function to represent or respond to. Griffiths argues however that in this approach there is only causal homeostasis at the task level and not at the computational and implementational levels. The reason it does not exhibit causal homeostasis at these levels is because the causal homeostatic mechanism of an ecological category is a set of adaptive forces, but by their very nature adaptive forces are only sensitive to properties at the task level. He gives the example of disgust which might remain a good category in cognitive ecology even if our computational and neurological models of disgust in e.g. rats cannot be extrapolated to birds "let alone to octopuses or Martians" (Griffiths, 1997, p. 234).

The ecological approach on its own is not sufficient to give us a scientific psychology of emotion. Griffiths favours instead a historical, phylogenetic account which uses the categories of homology rather than ecology. The class of affect programs which comes out of such an approach; surprise, fear, sadness, anger, disgust, and joy can be shown to have evolutionary relationships, that is they have developed historically. Because of this shared history if we take an affect program such as disgust then we find that not only is there a causal homeostatic mechanism at the task description level (for example, an averse reaction to things in the environment that harm the organism) but also at the computational and implementational levels (homologs of the underlying structures that realise disgust will be in each species).

[1] Sterelny's levels of explanation are an adaption of Marr's (1982) levels of visual processing.

Affect programs on their own do not account for all emotions however. Griffiths suggests that something that seems to be in common with all those states that we call emotions is their "passivity" by which he means that they are independent of long-term planned action (p. 230). Affect programs clearly fit this bill; whichever way you understand them (Ekman for example takes them to be literal neural programs whereas Griffiths takes them to be a coordinated set of physical (and neural) changes) they are complex, coordinated and automated. That is, they are more similar to reflex action than other sorts of information processing. Passivity is a result of the separate information storage and processing that leads to the production of the affect programs (Griffiths, p. 230), that is to say the affect program system is highly modular (in the Fodorian sense[2]). Not only can affect programs not account for all of those states that we call emotional but this feature of passivity can be seen in emotion responses caused by higher cognition as well. "Higher cognition" is to be understood as "the processes in which people use the information of the sort they verbally assent to (traditional beliefs) and the goals they can be brought to recognize (traditional desires) to guide relatively long-term action and to solve theoretical problems" (Griffiths, 1997, p.92). Higher cognitive emotions can be seen to have the feature of passivity because they are "irruptive motivations" (Griffiths, 1997, p. 243):

> These [irruptive] motivations are not derived from standing general goals by means-end reasoning. They occur in response to certain immediate circumstances. Because they occur independently of the general derivation of means to ends, they frequently disrupt longer-term plans. Loyalty leads people to keep an agreement at great cost. People are driven to revenge themselves even when it is disastrous for them to do so." (Griffiths, 1997, p. 243).

As higher cognitive emotions exhibit passivity as well, while other forms of information processing do not, and together affect programs and higher cognitive emotions seem to encompass pretty much all of the concepts and categories to which the vernacular concept of emotion(s) extends, we have good reason to think that emotions *are* either affect programs or higher cognitive emotions. Note that passivity does not count as a causal homeostatic mechanism however (and therefore does not enable the category of emotions to be a natural kind) as there does not seem to be a single mechanism which realizes this passivity and which is projectable. The result is that some of those things that we call emotions turn out to

[2] See (Fodor,1983).

be affect programs, and others turn out to be higher cognitive emotions, and these are two different categories.[3]

Scarantino (2005) extends Griffiths' project and suggests a positive way of overcoming the difficulty presented by the vague vernacular category of emotion. He argues that trying to capture a set of necessary and sufficient conditions for emotion is an ill-conceived project because the best account of emotion concepts is a cluster account. An emotion theorist who is involved in the folk emotion project should therefore seek only to give a descriptive account of the folk emotion categories. While a worthwhile project, this is not a scientific project. If we want fruitful definitions, that we can use for the purposes of explanation and prediction then we need to engage in 'explication'. Scarantino thus borrows and develops Carnap's (1950) notion of explication. To explicate a category we take the folk category but do not aim to capture all and only the meaning of that folk emotion category. Instead, the explicating emotion theorist only seeks to achieve similarity in use between the category emotion and the new explicated category (the explicatum). The primary concern is to show what useful theoretical purpose is served by the explicated category. An explication of any category is always *relative to some theoretical purpose*, so explication is by nature pluralistic. This means there may be many useful explications of "emotion". For example, Scarantino (2005) explicates 'emotion' for the theoretical purposes of scientific psychology. In his "Urgency Management System" he names his explication of emotions "umotions". And he defines them as "a special type of superordinate system which instantiates and manages an urgent management tendency by coordinating the operation of a cluster of cognitive, perceptual and motoric subsystems" (2005, p. 282). Whether or not his particular explication project works, for our purposes it suffices to be aware that it is not self-evident that "emotion" is definable in such a way as to make the referent of the term consistent in all contexts.

Giving up on the vernacular concept (apart from its use in guiding us to what is of interest to explore) has another side-effect. The vernacular concept of emotion included a sharp distinction between emotions and propositional attitudes (beliefs and desires). Contrary to what one would expect given the inclusion of 'desires' in belief-desire (or 'folk'/'philosophical') psychology, emotions here are not seen as required for reasoning.

[3] Griffiths is not wedded to there being only two categories. We should continue to isolate the categories by seeking causal homeostatic mechanisms so if it turns out that there are not causal homeostatic mechanisms at each level for the category of higher cognitive emotions then we should split it into categories that do.

Desires can be affectless cost-benefit cognitions. The sharp distinction between cognition and emotion is not one that we ought to take as our null hypothesis. Current research in the cognitive sciences which I will review in later chapters shows that emotion and cognition are intertwined and at some levels it may not be helpful to make a distinction at all, and certainly not one based on the vernacular concepts.

1.5 Summary of section 1

Feeling theories, cognitive theories, and philosophical appraisal theories all emphasise one component of emotions at the expense of the others. Modern appraisal theories, and in particular Scherer's (2005) component model manage to incorporate bodily and feeling components, even if they are not taken to be the primary eliciting and differentiating factor in an emotion episode. However, these models (unlike the feeling and cognitive theories) are not attempts to define emotion by presenting constitutive conditions. Rather they are practical models used for researching the mechanisms of elicitation of emotions in typical circumstances. Notably they do not try to account for all affective phenomena, only "full-blown" emotions, and can readily incorporate the dynamic, episodic nature of emotion processes which allows many components to be involved and interact. What can be seen by exploring these models is that the affective components of emotion, both bodily changes and feeling state, are distinct from (though they may interact with) the cognitive components, even when cognitive components are conceived of in terms of low-level processing as in modern appraisal theories. This fragmentation of the vernacular concept 'emotion' into various components suggests that there may be nothing that really answers to the vernacular concept. Rather the different philosophical theories (feeling; cognitive; the meaning strand of appraisal theory) end up giving accounts of the different fragments. One solution to this is to attempt to reconcile the components, which is what Jesse Prinz tries to do in his "embodied appraisal theory" which I explore in detail in the remainder of the present chapter. Another approach would be to follow Griffiths in rejecting reconciliation and to focus on one of the fragments exploring what role it plays in cognitive processing and phenomenology. This is the approach that I will take in the following chapters, showing that affect (bodily and experiential) is at least partially constitutive of cognition.

Section 2:
Prinz' Embodied Appraisal Theory: the "grand reconciliation"

We saw in the previous section that feeling theorists and cognitivists emphasise different components of emotion at the expense of the others. Prinz tries to reconcile these components while arguing for a theory of emotions which is embodied in that the perception of bodily changes is constitutive of an emotion, while maintaining a representational link between the bodily changes and their intentional object. This is an important new addition to embodied theories of emotion as previously embodied theories, such as Damasio (1999), could not account for the intentionality of emotions without succumbing to identifying emotions not merely with the embodied aspects but coupling it with an evaluative process (Damasio, 1999, p. 139; Prinz, 2004a, p.55).

Prinz (2004a; 2004b) defends what he takes to be the core idea of the James-Lange somatic emotions thesis; that emotions are perceptions of bodily changes. The James-Lange view is normally considered to be that a perception of an exciting fact (i.e., some emotional stimulus in the environment, such as a fearsome bear charging at you) leads to bodily changes, and it is the feeling of those bodily changes that is the emotion. Prinz modifies this so that 'feeling' is understood to be a 'perception', and this means that he is not tied to defending a purely personal level account of emotions; a perception may be subpersonal.

[A]: James-Lange view
Perception of exciting fact (e.g. scary bear) -> bodily changes -> [feeling (experience) of those changes = the emotion]

Prinz defends this core idea (a weaker claim than the above) in a modified form.

[B]: Perception of exciting fact (e.g. scary bear) -> bodily changes -> [perception (personal level or subpersonal level) of bodily changes = the emotion]

The core idea [B] that Prinz wants to defend takes emotions to just be the perception of certain patterns of bodily changes, and while there is good reason to think that this is the case given the current state of empirical evidence both for the thesis and the absence of evidence against it which Prinz reviews, there is one abiding problem which faces the feeling theorist. If we just take the modified J-L view presented in [B] it is not clear how we can account for

the intentionality of emotions. It seems intuitive that given normal conditions it is irrational and inappropriate to react to a writing desk with abject fear, or to an attacking grizzly bear with happiness. This is because we consider our emotions to be reflecting/registering/representing something about the external world; they have some sort of intentionality, and "…[i]ntentionality renders emotions amenable to rational assessment. They can be right or wrong, appropriate or inappropriate, warranted or unwarranted, rational or irrational" (Prinz, 2004a, p.54).

Prinz proposes (2004b, p53) that we add a Dretskian notion of representation to [B]. According to Dretske (1981; 1988) a mental representation is a mental state that (1) carries information, and (2) can be erroneously applied. Carrying information is understood according to Shannon & Weaver's (1963) model wherein "…a state carries information about that with which it reliably cooccurs" (ibid). So a state x can carry information about y in virtue of x being caused by y. To get the possibility of erroneous application x must have the *function* of carrying information about y; that is, x ought to have been acquired (either by learning or evolution) for the purpose of carrying information about y (though it may also carry information about/be set off by non-y's).

On its own, the Dretskian theory of representation added to [B] cannot get Prinz what he needs for the intentionality of emotions. Recall [B]:

Perception of exciting fact (e.g. scary bear) -> bodily changes -> perception of bodily changes

So the bodily changes might be thought to represent the perception of the exciting fact, and the perception of bodily changes to represent the bodily changes. Of course, it is not that simple, as the perception of the exciting fact itself ought surely also to be thought to be representing the exciting fact:

Exciting fact (e.g. scary bear) <– *represents?* <– perception of exciting fact <– *represents?* <– bodily changes< – *represents?* <– perception of bodily changes

Prinz argues that if the perception of bodily changes can not represent the bodily changes then emotions do not have intentionality as there is no survival advantage to the system knowing about those bodily changes. And, if there is no survival advantage then they would

not have been "set up to be set off" by them; that is, the perceptions of the bodily changes do not have the function of representing the bodily changes after all. Without having the function of representing the bodily changes it is not clear that they can misrepresent them, and therefore Dretske's criterion of the possibility of erroneous application is violated (Prinz 2004b p.58). Relatedly, Prinz argues that it isn't clear why it should be the case that we should reason badly if we misrepresent the changes in our bodies: "Suppose I do not know whether a certain course of action will make my blood vessels dilate or constrict? Does my ignorance lead me into recklessness? If so, it is not clear why." (Prinz 2004b, p.59).

To enable the perception of bodily changes (the emotion) to represent the relevance from the exciting fact (the danger incurred by the bear rushing at you) Prinz introduces Kenny's (1963) distinction between formal and particular objects (Prinz 2004b, p. 62). The objects of an emotion are the environmental conditions, which perturb the agent. If a grizzly bear is attacking you, the *particular object* of your emotion is the grizzly bear in question. But the formal object is the description of the relation between you and the bear in virtue of which the emotion arises. You are not scared of the grizzly because it has big teeth or giant claws, though both these might well be true of it; in an alternative situation these would not cause you fear. It is the danger that the bear poses to you which causes fear to arise in you. Danger is a relational property between the particular object and you, and it is this which causes you to be scared. If the bear posed no danger; if it were chained up for example, rather than causing fear it might cause pity. So Prinz follows Kenny in proposing that "…[a] *formal object* is the property in virtue of which an event elicits an emotion, and a *particular object* is the event itself." (Prinz 2004b, p. 62).

Prinz proposes that emotions represent their formal objects rather than their particular objects. It is easy to see that a system's ability to represent danger would confer a survival advantage to it. He suggests that the formal objects that emotions represent are more than just alternate descriptions but correspond to Lazarus's (1991) core relational themes. Core-relational themes are the organism-environment relations which hold when one is in an emotional state, such as 'danger' for fear. Prinz argues that the emotion 'fright' for example represents "facing an immediate, concrete, and overwhelming physical danger"; and 'sadness' represents "having experienced an irrevocable loss" (Prinz 2004b, p. 16, Table 1.2; Lazarus 1991, p.161, Table 3.4). The core-relational theme is a relation between the agent and the environment; it is a relational property. It encompasses a response dependent property; in the case of sadness the relational property is loss and the response dependent

property is being of value (sadness only arises as a result of the loss of things that are valued by you). But, importantly, being a loss isn't a response dependent property in its own right; it is still a loss even if you don't know about it and therefore are not given the opportunity to respond (Prinz 2004b, p.63). Where do Kenny's objects fit in on our picture of [B] then? Recall that the skeleton of Prinz's theory is [B] below:

[B]:
Perception of exciting fact (e.g. scary bear) -> bodily changes -> perception of bodily changes

We will need to add the particular object, which will be the 'exciting fact':

[C]:
Exciting fact (e.g. scary bear) -> perception of exciting fact -> bodily changes -> perception of bodily changes

But as noted before the links in this chain are not representational. So Prinz introduces a subclass of structured representations; "appearance-tracking detectors". An appearance-tracking detector can represent the essential properties of x by registering the overt features of x. I represent a coke can (whatever properties are essential to being a coke can) by registering the cylindrical shape, the pattern of red and white, the cursive font etc. The essential properties are the "real content" and the features by which we detect this real content are the "nominal content".

The links in [C] are not allowed to be representational but they may indicate *registration*. Prinz proposes that while the perception of bodily changes does not represent the bodily changes, it does register them. And it is in virtue of detecting these features (the nominal content) that the real content (the core-relational theme) is represented. So, whereas I represent the coke can in front of me in virtue of registering its shape, and coloured pattern, my emotion of fright (a perception of particular bodily changes) represents danger in virtue of registering the actual changes in bodily state (accelerated heart rate, sweaty palms etc.).

The last step is to explain how formal objects are connected to particular objects. Prinz suggests that we can understand this in terms of (mental) elicitation files (2004a, p. 55). The story goes that we are either wired so that certain changes in the environment, that are for

example dangerous, elicit a particular bodily response, or starting from a few pre-wired triggers (such as looming objects) we learn new associations between those bodily responses that are elicited and alternative environmental elicitors (in this way we can learn to be scared of, for example, creepy crawlies). Evaluative judgments that something is, e.g., dangerous, can also be elicitors and become associated with bodily responses on the basis of similarity to, or association with current triggers. So in the case of a looming object, the loomingness causes the change in bodily state in virtue of having the property of being dangerous. The elicitation file has an association between the particular looming object (loomingness) and dangerousness (the formal object):

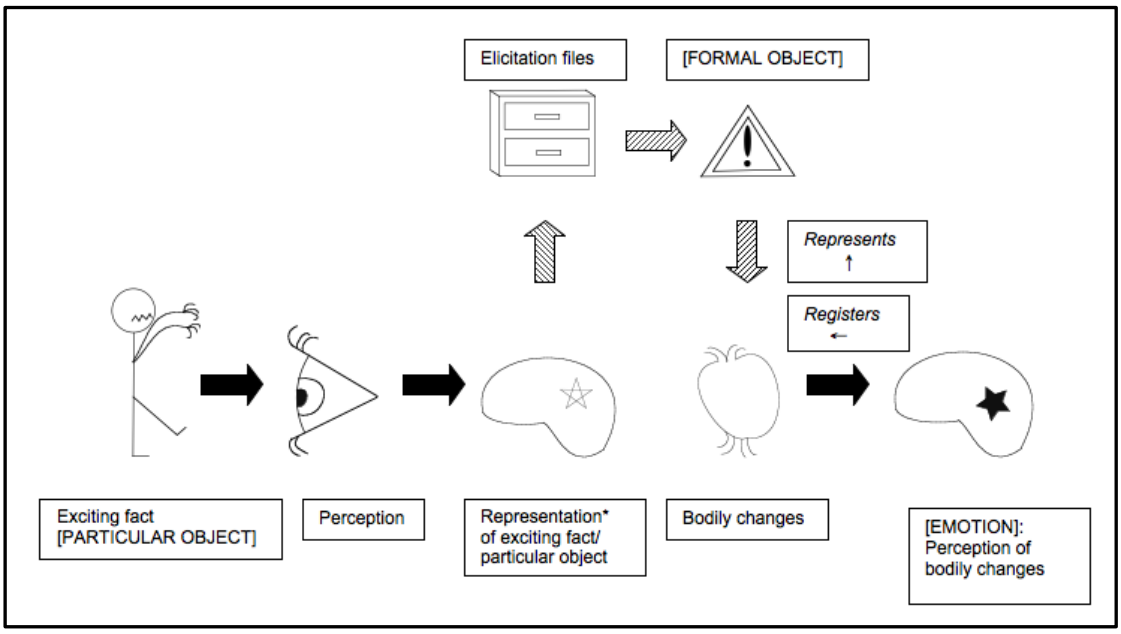
Elicitation file:

Triggers -> [Property] -> Bodily responses

e.g., Looming object -> [Danger] -> Bodily changes which protect agent from danger

So, cashed out and expanded I take the diagram below to be an accurate embellishment of [B] as Prinz defends it:

[D]: Prinz's Embodied Appraisal Theory of Emotions



[Representation* = a neural pattern, not a representation according to Dretske's criteria]

Prinz does not answer the question of how particular objects can be connected to bodily changes and hence perceptions of bodily changes, but he does not see this as a problem as

we know that there is some causal link and so any emotion theory will have to give some sort of a mechanism for this dependency; the core thesis does not depend greatly on how this dependency is cashed out.

While [B] was an clearly an embodied theory of emotions, Prinz considers [D] an embodied *appraisal* theory. He contrasts this type of appraisals with evaluative judgments. For Prinz, evaluative judgements are only part of the story and not fundamental to it; fundamentally emotions are embodied. Cognitions, defined by Prinz as including "representations that are under the control of structures in executive systems" (2004b, p.47) do not need to play a role. Prinz evades the cognitive element by defining appraisals as "…any representation of an organism-environment relation that bears on well-being…" (Prinz 2004a, p. 57). As we have seen, the organism-environment relation that bears on well-being is the formal object (corresponding to Lazarus' core-relational themes) and this may be triggered by subpersonal associations. The representation of the formal object (through the formal object reliably causing the bodily changes, and these in turn being registered by the perception of these bodily changes) is the perception of bodily changes; i.e., the emotion. As the emotion both registers bodily changes and through these represents the organism-environment relation, it is both embodied and an appraisal. Under this definition judgments can be appraisals, but not all appraisals need be judgments. As Prinz explains:

> My suggestion is that certain bodily perceptions … represent roughly the same thing that explicit evaluative judgments represent, but they do it by figuring into the right relations, not by deploying concepts or providing descriptions. Our perceptions of the body tell us about our organs and limbs but they also carry information about how we are faring. (Prinz, 2004a, pp. 57).

In Prinz's model evaluations may play a causal role but they are not constituent parts of the emotion itself; they are merely part of an elicitation file. Prinz is explicit that "the fact that elicitation files help establish the intentional content of emotions does not entail that they should be regarded as constituent parts" (2004a, p. 57). His primary reason for this seems to be that emotions are elicited by different representations on different occasions and so if these representations were considered to be constitutive of emotion then emotions would be different on each occasion, a consequence that he wishes to avoid. For Prinz, the emotion itself just has one constituent part; the perception of bodily changes. It gets its intentionality in virtue of its *representing* the formal object/organism-environment relation which it does in virtue of registering the nominal content of the representation, the bodily changes).

Evaluations, in so much as they do sometimes play a role in the model, like subpersonal appraisals, play a merely causal – not constitutive – role.

Section 3: Objections to Prinz' embodied appraisal theory

In section two I sketched Prinz's embodied appraisal theory. Here I present three objections to the theory. I argue that (1) that there is a worrying conflation between emotion and emotional experience in his model (2) that an account of the constitution of emotions, such as the one Prinz offers, is not a scientifically plausible endeavour, and (3) Prinz's theory is not in fact a "grand reconciliation" between feeling and appraisal theories. I conclude that Prinz's attempt to reconcile the feeling and cognitive components of emotion is destined to failure as a result of his conflating a descriptive project with a scientific project.

3.1 Emotions and emotional experience

Prinz does not define emotion as subjective feeling as, for example, dimensional feelings models do (see Grandjean et al. p. 485) and he has a separate account of emotional consciousness (the AIR theory, see Prinz 2004b, pp. 209-220) which is supposed to account for subjective feeling. However, Prinz cites the James-Lange necessity claim as a reason to favour defending the James-Lange thesis (what he calls 'the somatic hypothesis'):

> What kind of an emotion of fear would be left, if the feelings neither of quickened heart-beats nor of shallowed breathing, neither of trembling lips, nor of weakened limbs, neither of goose-flesh nor of visceral stirrings, were present, it is quite impossible to think. (James, 1884, 193f; cited in Prinz 2004a, p. 46)
>
> If from one terrified the accompanying bodily symptoms are removed, the pulse permitted to beat quietly, the glance to become firm, the color natural, the movements rapid and secure, the speech strong, the thoughts clear, -- what is there left of his terror? (Lange 1885, 675; cited in Prinz 2004a, p.46)

We must be careful however, not to take these quotations out of context. In the first (from James 1884) emotions are clearly identified with feelings. But what is not made clear is that 'feelings' here refers to conscious feelings; the experience of emotion. This is explicit in other passages in the James' essay, such as:

> If we suppose [the brain's] cortex to contain centres for the perception of changes in each special sense-organ, in each portion of the skin, in each muscle, each joint, and each viscus, and to contain absolutely nothing else, we still have a scheme perfectly capable of representing the process of the emotions. An object falls on a sense-organ and is apperceived by the appropriate cortical centre; or else the latter, excited in some other way, gives rise to an idea of the same object. Quick as a flash, the reflex currents pass down through their pre-ordained channels, alter the condition of muscle, skin and viscus; and these alterations, apperceived like the original object, in as many specific portions of the cortex, combine with it in consciousness and transform it from an object-simply-apprehended into an object-emotionally-felt. No new principles have to be invoked, nothing is postulated beyond the ordinary reflex circuit, and the topical centres admitted in one shape or another by all to exist. (James 1884, p. 204)

It is less clear from Lange's quote above that he identifies emotion with conscious feeling of the bodily changes, but careful reading of the text surrounding that extract shows that can this is also the case:

> We have in fact no absolute and immediate means of determining whether a sensation is of a psychical or bodily character. Furthermore, no one is able to indicate the difference between psychical and somatic feelings. Whoever speaks of a psychical impression does so indeed solely upon the basis of a theory, and not upon an immediate perception. Without doubt, the mother who sorrows over her dead child would resist, probably even become indignant, if anyone were to say to her, that what she feels, is the exhaustion and inertness of her muscles, the numbness in her bloodless skin, the lack of mental power for clear and rapid thought -- all of which is made clear by the idea of the cause of these phenomena. There is no reason, however, for her to be indignant, for her feeling is just as strong, as deep and pure, whether it springs from the one, or the other source. But it cannot exist without its bodily attributes.
>
> If from one terrified the accompanying bodily symptoms are removed, the pulse permitted to beat quietly, the glance to become firm, the color natural, the movements rapid and secure, the speech strong, the thoughts clear, -- what is there left of his terror?
>
> If we cannot rely, therefore, in this question upon the testimony of personal experience, because it is here incompetent, the matter is thereby naturally not yet explained. If the hypothesis of psychical emotions be not made necessary by subjective experience, it may nevertheless be requisite if without it one cannot perhaps understand how the bodily manifestations of the emotions come into existence. (Lange, 1885, 675)

The bodily changes for both James and Lange are thus "bodily manifestations of emotions" and to become emotions the perception of these changes needs to become linked with consciousness; emotions are thus the *feeling* of bodily changes in the way we normally

understand the term feeling, which is not the same as Damasio's usage [4]. The necessity claim that James and Lange are making, then, is that for a felt emotion to occur, that is to have an emotional experience, one needs to feel changes in the body. If you take away the conscious feeling of those bodily changes, one does not feel, for example, scared.

There are good reasons for not wanting to define emotions as feelings. If we are engaged in the descriptive project (Scarantino, 2005) we might appeal to our folk psychological notions of emotion; how we use emotion terms in order to predict others behaviour. If we could set it up somehow that a system had all the physical manifestations of anger, which will include the action tendencies associated with anger, but we were assured that he was experiencing a lovely soothing calm, euphoric feeling, it would be of no use to us to say of him that his emotion is one of calmness, or happiness. What would be of use to us is that his physical changes and action tendencies allow us to predict that he will behave in certain ways when he is confronted, and in normal life this is what we mean when we say that someone is angry. In reality it often happens that people are so focused on something else, or perhaps are pathologically low in insight about their own emotions, so that they do not realize that they are in a bad mood, or depressed until prompted by an outsider. I do not think that we would be comfortable in saying that such a person was not grumpy, or angry, or depressed just because they hadn't realized it, because they hadn't had a conscious feeling of that bodily state. Of course it could be argued that such a person did have conscious feelings of their emotion - they just hadn't had consciousness of that consciousness until prompted. But this seems to just push the problem back one level; surely then we ought to be identifying emotion with the feeling of the feeling of the bodily changes?

It is true that Prinz is quite clear that he does not want to be tied to a personal level account of emotion and that is why he changed the thesis that he wished to defend from [A] above, to [B]; that emotions are the *perceptions* of bodily changes. Indeed in his chapter on emotional consciousness he states "I join the majority in denying that emotions are mere feelings" (Prinz 2004b, p. 198). But then it is not clear how the James-Lange necessity claim (i.e. the subtraction argument – Prinz 2004a, p. 46) is supposed to help support the kind of somatic theory that Prinz is proposing. The subtraction argument – that if the feeling of the bodily components of an emotion were subtracted, there would be nothing (no 'emotion') left to feel – does no work if the bodily feelings can be subpersonal. This is not a criticism of

[4] Damasio uses the term 'feeling' to refer to first order representations of bodily changes; the consciousness of this feeling comes as a result of a second-order representation (see Damasio, 1999).

Prinz's model as such but rather a criticism of his use of the subtraction argument in motivating a somatic account. Despite his claims Prinz is not defending the James-Lange thesis; his is not a feeling theory of emotion, and camouflaging his theory as a variant on the James-Lange theory risks conflating emotional feelings (affect) with emotion as a (not necessarily conscious) perception of bodily changes[5]. This conflation may be due to the framework in which Prinz has chosen to build his theory; a traditional analytic philosophy of mind informed by contemporary cognitive science (essentially a descriptive rather than explicative project). Sometimes the conflation is merely a linguistic mistake resulting from common ways of talking about emotions. And perhaps partly it is just a result of having to explain a complex theory in stages. However, sometimes it is because he talks in terms of sufficiency, necessity, and constitution (which I will discuss more in the next subsection). This is the vocabulary of a descriptive project and not a scientific one. However, as a descriptive project it seems to be lacking as it ignores parts of what the vernacular category of emotions would include, such as action tendencies (emotional behaviour).

In sum, although Prinz's model does not itself turn on the subtraction argument, his use of it in motivating a somatic account of emotion results in his conflating affect with emotion. This then leads him to make a move which effectively expunges affect from his account; affect is not constitutive of emotion. But it is one thing to deny that emotions should be *identified* with feelings, and quite another to eliminate them from an account of emotions entirely, especially considering that Prinz is specifically trying to give an account of the vernacular category of emotions. Likewise, for Prinz, the bodily changes are not considered as constitutive of emotion even though they reliably cause the perception of themselves, which is what he identifies with emotion. Prinz's "embodied appraisal" theory, in clear contrast to the James-Lange model, is thus essentially *disembodied* and *affectless*.

3.2 Emotions and constitution

It is traditional in philosophy to try to understand the constitution of whatever is of interest and Prinz frequently talks in terms which makes clear that he is making a constitution claim about emotions. He talks about "the emotions themselves" (Prinz 2004b, p. 2), as a means of distinguishing emotions from elicitors of emotion. He also talks about the sufficiency of bodily changes for emotions, which also suggests a constitutive view. Moreover he explicitly talks of the elicitation files and their contents as not constituent parts (e.g. Prinz 2004a p.

[5] I thank Edoardo Zamuner for bringing this discrepancy to my attention (personal communication).

57). Constitutive views are troubling in cognitive science as it isn't clear that mental events can be neatly delineated from their causes. Take "the emotions themselves", for example. This assumes a linear model of psychology in which an emotion elicitor causes an emotion which itself is distinct from the elicitor. But biology just isn't this neat, the processes which underpin our abilities are highly dynamic and recurrent. One could respond that the biology may not be this neat, but that we can abstract to a level at which we can give an explanation in linear causal terms. This may be true but I do not think that Prinz is doing this. What is most potentially exciting about Prinz's model is that he is combining abstract psychological concepts with neuroscience. What he takes to be the emotion is a perception of bodily changes, which we should take to be what, in neuroscience, is called a representation, i.e., some neural pattern which arises as a result of changes elsewhere in the system (these changes include stimuli to the sensory organs). This assumes therefore that this neural representation which also has the function of representing (according to Dretske's criteria; we might call this a 'philosophical representation') core-relational themes, is clearly delineable. Can we really get any explanatory value from identifying an 'emotion' with this representation? Let us ignore the controversy over whether it can make sense to talk of brain states or processes in neuroscience (see Bechtel and Mundale 1999; Brown 2006), and let us ignore that the brain is a complex, dynamical system, and let us even ignore (for now) that emotional processing involves many recurrent loops and spirals between different areas in the brain. Even taking the most simplistic notion of a neural representation as an emotion we are left with the consequence, according to Prinz's model, of being able to point to an area (or multiple areas) of the brain and saying "that there is the emotion; that is fear!". Prinz is committed to this consequence because he is identifying 'emotion' with the *perception of bodily changes* (a registration – in the brain - of the bodily changes which were reliably caused by a formal object and so represent this formal object) and not the *feeling of bodily changes*. We should be uncomfortable with such a consequence.

Even when we understand some neural pattern to be a neural representation of a set of bodily changes and a 'philosophical' representation of danger (i.e. according to Dretske's criteria rather than just a neural map of changes), that still doesn't seem to be sufficient for it to count as an emotion. Firstly, there are no action-tendencies yet mandated. But it is plausible that (as in Lazarus' account) action tendencies are integral to emotions. Prinz's model follows Plutchik (1984) in distinguishing emotions from motivations, though his distinction is not grounded in the same contrasts as Plutchik. For Prinz, emotions are motives rather than motivations. A motive is something which provides a reason for action, whereas a

motivation is that which actually impels us to act (see Prinz 2004b, pp. 191-194). Motivations include states like fatigue, hunger, thirst and sexual drive, and they are "affectively motivated action commands" (Prinz 2004b, p. 195). Emotions can cause motivations but for Prinz "…when emotions cause motivations, those motivations never count as constitutive parts of emotions" (*Ibid*, p. 196). Although emotions and motivations often come together they don't need to; "affectively motivated" in the above definition of motivations means only that some valenced representation causes it, and while all emotions are valenced representations (valenced embodied appraisals), not all valenced representations are emotions. Valence, in Prinz's model, is just an inner reinforcer which attaches to the representation and shouts "seek!" or "avoid!".

Prinz also appeals to a semantic difference to separate motivations from emotions. Emotions as embodied appraisals are representations only of the core-relational theme. Whereas those motivations that *seem* to be embodied appraisals are representing the bodily changes; that is we can understand e.g., hunger as having the function of representing the bodily state. So rather than representing the real content (core-relational theme) in virtue of registering the nominal content (bodily changes) as is the case for emotions, motivations have bodily states as both the real and nominal contents; a motivation such as fatigue represents *only* the bodily changes and no core-relational theme.

The true shape of the issues is now becoming apparent. Prinz accounts for how emotions are linked to action tendencies at the same time as differentiating between emotions proper and other bodily representations, such as hunger and fatigue, But this story is unhelpfully complex. Prinz is presenting a constitutive account and as such emotions for him are not partially constituted by action tendencies; instead there is merely a causal relation between emotions and motivations (i.e. the 'motives'). It seems unlikely that organisms would evolve a system for representing danger in virtue of registering their bodily changes but then have to go through another step to activate the action command, especially as the bodily changes that are being registered by this representation (that is supposed to be the emotion) are ones which, according to Prinz's model, have been set up to respond to the situation in some way already. All that ought to be needed is quite literally a command (personal level or subpersonal level) to "go!" (or indeed no command at all, it is more likely to be the case that these responses are automatic and that more complex creatures – i.e. those with better developed cortices - develop the ability to *inhibit* them). But Prinz's action commands seem to be more than a simple order to go; they are commands to the body of what to do. I suggest

that this has already been done by this stage and that these affective motivations are plausibly constitutive of the emotion (if one wishes to speak in terms of constitution). To substantiate this idea I shall (in chapter 3) offer a minimal appraisal model which incorporates the motor aspect within affect itself (not the actions themselves but the processing underpinning the responses) and which can also account for the less reflective cases of 'emotional' behaviour.

Another way in which Prinz's story is too complex is the addition of valence markers. The structure of the emotion is supposed to be partially an inner positive or negative reinforcer, which then guides or prompts the action command as to whether it should be an approach or withdrawal action. These inner reinforcers are inner states that arise as a result of an association (either genetic or learned) with stimuli (rather similar to Damasio's (1996) Somatic Marker Hypothesis which I will discuss in detail in the next chapter). But what is this stimulus: Is it the particular object (the bear running at me)? Is it the perception of the particular object (the neural representation of the bear running at me)? Or, is it the bodily changes?

If it is the particular object, as we might expect given that in psychology this is what would normally be called the stimulus, then we might expect the association with inner reinforcers to happen at the stage where the stimulus is being (neurally) represented in the brain. But then the association is with the perception of the particular object and not with the perception of bodily changes. And, if the valence marker is supposed to be part of the structure of the emotion it needs to be associated with the perception of bodily changes. If it is elicited by the perception of the particular object then it is unclear that it is really part of the structure of the emotion. For the valence marker to be structurally integrated into the emotion it needs a direct route to the perception of bodily changes. This could either come from the bodily changes, or it could come through the elicitation file route. The elicitation file route would make sense as this is what associates particular objects with formal objects. However, if this is right, then the inner reinforcer becomes part of the formal object and is just *represented by* the emotion. Likewise if the stimulus is taken to be the representation of the exciting fact (rather than the exciting fact itself). Prinz is quite clear that the formal object is not a constitutive part of the emotion so if the inner reinforcer is part of the structure of the formal object, it is not part of the structure of the emotion.

We are left only with the possibility of the stimulus being the bodily changes. It is not unreasonable to imagine associations between particular bodily changes and inner reinforcers; in fact, this is surely what homeostasis is. I will argue later in the thesis that affect has precisely this dual face of bodily changes and valence. However it is not clear that Prinz would be happy with this; It seems that Prinz really does want the stimulus to be external. After all, if we were to take the suggestion above of the bodily changes being associated with inner reinforcers, then rather than the valence marker being separate but structurally integrated into the emotion, the perception of the bodily changes would just be valenced. It need not have any association extra as the perception of bodily changes itself is an association. If we were then to introduce the notion of action-oriented representations (see Clark, 1997; Mandik, 2005), we could also cut out the extra step which Prinz wants to make for action commands, and thus have a much neater, more biologically plausible account. Note that the kind of account sketched briefly above need no longer be a constitutive account; that is, emotions no longer need to be thought about in terms of having an internal structure which needs to be accounted for in terms of constitution[6]. Prinz's ambition to provide a constitutive account of the vernacular category of the emotions is destined to failure because it simply cannot include all the components that the vernacular category encompasses. And, by rejecting those components as 'merely' causal and not constitutive of emotion he ends up with a constitutive account of something which does not in fact correspond to the vernacular category after all (though it may account for part of it). In particular, he ends up with an account of emotion which is, I argue, non-affective.

The solution would be to either (1) reject the vernacular category of emotions as the explanandum (as Griffiths and Scarantino suggest), or (2) to reject attempts to provide a constitutive account of emotion (like modern day appraisal theorists) instead accepting that emotions are dynamic, episodic, and involve multiple components and processes. Rather than separating the cognitive, bodily, valence, and experience components from what is constitutive to 'the emotion' as Prinz does (recall that for him none of these are considered to be constitutive of emotion; emotions are constituted only by the perception of bodily changes and the representation of the formal object) these approaches integrate them and thus explicate the various emotions with the components of the vernacular category (though the category itself may not be amenable to such explication). The lesson I take from this is that it is a more scientifically viable project to seek the components of an explanandum and

[6] I think that Scarantino (2005) does this with his "umotions" project, which he gives "pushmi-pullyu representations.

investigate their interaction than to force a constitutive account from a top-down (vernacular) concept of the explanandum. It is for this reason that, in the following chapters, I will separate out the affective (bodily and feeling) components and investigate their role in cognition.

3.3 Embodiment and appraisals: rejecting the "grand reconciliation"

Prinz claims his theory is a "grand reconciliation between the appraisal tradition and the tradition inaugurated by James and Lange" (2004b, p. 68). I submit that no grand reconciliation has there been achieved. Firstly, as I discussed earlier in the chapter, the feeling theory as inaugurated by James and Lange focused on the necessity of the *experience* of bodily changes for an emotion and not the bodily changes themselves or even some neural percept of those changes. Prinz is therefore rejecting the key element of the feeling theories; that emotion is the feeling of bodily changes[7]. At the level of neurobiology Damasio's research is highly relevant along with the research of other affective neuroscientists such as LeDoux (1996) and Panksepp (2004). But like them, he does not present a theory of emotion that can be understood to rival the appraisal tradition, rather his results feed into modern appraisal theories. Secondly, and most importantly, Prinz targets only cognitive appraisal theory, i.e., the specialized version of appraisal theory which takes appraisals to be deliberative judgements. The alternative he suggests, a subpersonal account of appraisals (2004a, p. 57), just *is* what modern day appraisal theorists take to be appraisals. Though everyone is clear that one *can* appraise a situation deliberatively, it is taken as read in psychology that typically appraisals are fast, subpersonal processes which allow us to respond to situations. As I outlined in the first section, modern appraisal theory does not reject the importance of bodily responses; they play an important part in what appraisal theorists understand as emotions. Nor do they ignore the importance of the neural pathways that run between sensing the physiological changes and effecting action. Prinz has therefore not so much reconciled feeling and appraisal theories but rejected cognitive appraisal theories in favour of a more modern understanding of appraisal. No reconciliation is needed between feeling theories and modern appraisal theories as the latter explicitly incorporate bodily changes and feelings as components of emotions.

[7] Damasio (1999) does take emotion to be a neural representation of bodily changes, and in some ways it is his lead that Prinz is following. But Damasio's position (as we shall see in chapter 2) is itself insufficiently philosophically nuanced.

What, then, is the role of the body in appraisal theories? As in Prinz's model perceptions of physiological changes are involved; they contribute to the appraisal as they feed in to the sensory contributions outlined above. But what makes appraisal theories such as Scherer's *more embodied* than Prinz is that they understand emotion not as a static state but as an episode which is characterized by "continuously occurring changes in the underlying appraisal and reaction processes" (Scherer, 1999, p. 648). These reaction processes are made up of a number of components including subjective feeling, physiological responses, motor expression, and action tendencies (*ibid*).

Prinz's grand reconciliation was supposed to be between feeling theories and appraisal theories. But, as we saw, no such reconciliation was actually required. Current appraisal theorists acknowledge the embodiment of emotion, even though for methodological purposes they focus on conscious elicitation of emotions and measuring emotional response using self-report. Furthermore, the role of significance evaluation that Prinz so heavily relies on (from Lazarus 1991) is still the focus of appraisal theories. As Scherer puts it "present-day appraisal theories can be considered the culminating formalization of two centuries of philosophical notions that have always insisted on significance evaluation as the core process of emotional reactions" (Scherer, 1999, p. 654).

Propositional attitude theories of emotion were philosophical incarnations of the original (cognitive/deliberative) appraisal theories in psychology, and I suggest that Prinz's theory is a philosophical incarnation of current appraisal theories. It is not a grand reconciliation because none was needed. What it does do, however, is give a story as to how emotions can be intentional, given that they are embodied, and this *is* needed to bring psychological theories of emotion into philosophical discourse. Such a story however does not require giving a constitutive account of the vernacular category by forcing the components together as if emotions were static representations.

3.4 Summary

My key objection to Prinz's model is that by having emotion involve only the perception of bodily changes and the representation of the formal object as its actual constituents Prinz loses the very elements that were supposed to motivate the model in the first place, and which would justify talk of a "grand reconciliation" involving the feeling component and the actual bodily changes. This flows from what I see to be a more fundamental problem with

Prinz's model; that it is straddling descriptive and explicative methodologies. Descriptive theories take the folk emotion categories and try to understand what makes emotions a category, while explicating theories try to show that there is an understanding of these categories which is useful to a particular theoretical purpose, though this understanding may not be isomorphic to the folk category. Prinz is struggling to maintain the folk category of emotions and its requirements, such as representations and intentionality, while stretching his model to give a scientific understanding of emotions at the same time. While this is an admirable ambition, it is fundamentally mistaken because it is mixing two projects that need to be kept separate. The representational story might (if you are sympathetic to that kind of approach) afford an explanation at the folk level. That is, it might help us understand the folk theory of emotion. But the neural account suggests a different understanding. By mixing it with the folk level we lose the opportunity to deploy our new understanding to truly explicate emotional phenomena. The result is a model that looks as though it is scientifically motivated but on closer inspection is not.

This chapter has explored the role of feelings, bodily changes, and cognition in several of the main accounts of the emotions. I considered in detail an account due to Jesse Prinz, representing a recent and philosophically nuanced attempt at a bodily account of emotions. I next examine another major recent model of emotion, due to Antonio Damasio, who proposes that the feeling of bodily changes is crucially involved in cognition. In chapter two I will investigate this proposal in detail before then presenting (in the core of the thesis) an alternative model of the role of the body and affect in cognition.

Chapter II

The 'Somatic Marker Hypothesis', criticisms, and alternatives

In the previous chapter I rejected Prinz's attempt to reconcile the cognitive and affective components of emotion. The role of this chapter is to examine another model of the interaction between cognition and emotion: the Somatic Marker Hypothesis (henceforth SMH). The SMH is a presented as a model of the role of emotion in cognition, in particular the role of emotion in real-world decision making. Decision making is a paradigmatically cognitive ability, and while it may seem trivial that emotion can affect decision-making in a negative way, i.e., making us bad decision makers, Damasio has famously argued that *emotion is crucially involved in good decision making*. While I agree that affect is central to cognitive processes, including decision making, in this chapter I reject this particular hypothesis of their interaction. In section one, I argue that the somatic marker hypothesis is not validated by the particular experimental setup used by Damasio and colleagues (the Iowa Gambling Task). In section two I present alternative explanations that can account for the data.

Whilst I greatly respect Damasio's work in affective neuroscience, I object to this particular hypothesis (the SMH) because it implicitly continues to endorse a framework which sees affect and cognition as separate faculties. While the SMH encourages us to think about emotion and cognition as highly interacting in typical (good) functioning, the way the hypothesis is presented encourages us to continue to think in terms of interacting 'modules' or 'faculties' and further that cognition can function without this interaction even if suboptimally. In subsequent chapters I will argue for a much deeper relation between affect and cognition.

Section 1: The Somatic Marker Hypothesis and Criticisms

1.1 The Somatic Marker Hypothesis

The 'Somatic Marker Hypothesis' (SMH) is the hypothesis that paradigmatic cognitive processes, such as decision-making, are not purely 'cognitive' in nature but also involve emotional components. The idea is that processes like decision-making do not consist merely

in emotionless cost-benefit analysis but rather make use of representations of bodily changes in guiding appropriate behaviour in virtue of links between areas such as the association cortices and areas which subserve autonomic functions or the representations of these bodily changes. These links which 'mark' cognitive processing with bodily (somatic) changes are thus called "somatic markers" and Damasio and colleagues hypothesise that the area which subserves the somatic markers is the ventromedial prefrontal cortex (VMPFC).

The SMH was developed in response to observations in neurological patients with focal damage in the frontal lobe (Damasio, 1996, p. 1413). In particular, it was developed to explain why it is that patients with lesions in the ventromedial prefrontal cortex seem to be impaired in their personal lives while achieving adequate or superior results in laboratory tests of their intelligence and rationality. They retain IQ, and performance in standard neuropsychological laboratory tests for learning, and knowledge retention for both factual knowledge and skills, logical problem solving, attention, and working memory. Damasio explains:

> Before the onset of brain damage the patients may be described as intelligent, creative and successful; but after damage occurs the patients develop a pattern of abnormal decision making which is most notable in personal and social matters. Specifically, patients have difficulty planning their work day; difficulty planning their future over immediate, medium and long ranges and difficulty choosing suitable friends, partners and activities. The plans they organize, the persons they elect to join, or the activities they undertake often lead to financial losses, losses in social standing and losses to family and friends. The choices these patients make are no longer personally advantageous, socially inadequate and are demonstrably different from the choices the patients were known to have made in the premorbid period. (1996, p. 1413)
>
> The disturbance shown by this particular class of patients cannot be explained in terms of defects in (a) pertinent knowledge; (b) intellectual ability; (c) language; (d) basic working memory; or (e) basic attention. (1996, p.1414)

Damasio noticed however that these patients seem less able to express emotion and to experience feelings in some situations where one would expect them to (Damasio 1996, p.1414). Damasio therefore hypothesised that the lesions affected normal emotion processing and thus the normal interaction between emotion (impaired in the pre-frontal patients) and cognition (not-impaired in the pre-frontal patients) which enables successful everyday functioning.

The somatic marker hypothesis proposes that in normal functioning the ventromedial areas of the prefrontal cortex function as linking areas between primary and secondary association

cortices and autonomic structures such as the amygdala and hippocampus. These autonomic areas are responsible for sending instructions to the body causing the bodily changes that, according to SMH, we feel as emotions when represented in the brain. These linkages provide the substrate for an emotional memory of a situation or parts of a situation. There is however a dissociation between complex stimuli which require cortical processing and basic stimuli which can be processed subcortically. If there is an event which induces primary emotional responses, then no extra cortical processing is required for the immediate emotional response of preparing the body for fight or flight. Information about looming, for example, is sent straight to the thalamus which sends the information straight to the amygdala, hypothalamus and brainstem nuclei. The processing in these areas has the result that changes in the viscera, vascular bed, endocrine system and non-specific neurotransmitter systems are effected; and thus fight or flight is initiated.

The association areas have projections to the VMPFC and when activity in those areas rises above a certain threshold information is passed to the VMPFC. Likewise there are projections from the central autonomic effectors to the VMPFC so when the bodily response is effected information about this is passed to the VMPFC. This means that the VMPFC comes to subserve associations between certain situations, or components of those situations, and the emotional responses that went with them. And, because the VMPFC also has projections *towards* the central autonomic effectors, activating these learned associations (by means of the primary or secondary association cortices sending information to these 'convergence zones' in the VMPFC) will result in information being passed to the central autonomic system resulting in a bodily reaction.

In short, if a situation or a component of that situation has previously been paired with a highly emotional state it is likely that an association between that component and that emotional state has been formed in the VMPFC. Reactivating the neural pattern that subserves the association in the VMPFC (the somatic marker) can be done by triggering areas in the association cortices that categorized parts of the original situation. This will result in the bits of the VMPFC that subserve the association (the somatic markers) firing and thus triggering bodily changes as a result of its projections to the central autonomic effectors. It should be added that Damasio considers these linking stations in the VMPFC to be dispositional memories. He explains:

> The linkages are 'dispositional' in the sense that they do not hold the representation of the facts or of the emotional state explicitly, but hold rather the potential to reactivate

> an emotion by acting on the appropriate cortical or subcortical structures... (1996, p. 1414)

Damasio calls these associations in the VMPFC 'somatic markers' because they mark (or 'tag') highly salient information with a bodily reaction. It should be noted however that it is not the case that a full bodily reaction is always caused. Under Damasio's model of emotion, emotion is not identified with the bodily changes themselves but rather the representation of these bodily changes in the somatosensory cortices. As a result one can have an emotion in virtue of the central autonomic effectors being activated, the body subsequently changing, and these changes being represented in the somatosensory cortices. Or, one can have an emotion by skipping the activation of the central autonomic effectors and bodily changes and just go direct to activating representations in the somatosensory cortices. Damasio calls this latter track the "as if loop". So, if the convergence zones in the VMPFC that record the links between components of situations and somatic states (i.e. the somatic markers) project to the somatosensory cortices rather than the central autonomic effectors then the actual bodily changes won't happen but there will still be an emotion (whether or not this emotion is felt depends, according to Damasio, on whether these representations in the somatosensory cortex are themselves represented but we need not go into this here). Having the "as if" emotion may lead to certain physiognomic changes such as raised heart rate, increased galvanic skin response etc. but the full suite of somatic changes need not be effected for the emotion to be present.

So to summarise the process by which somatic markers are activated here is an outline of the steps (based on Damasio, 1996, p.1415):

(1) Situation arises for which some factual aspect has been previously categorized..
(2) dispositions activated in higher-order association cortices
(3) ...leads to recall of pertinently associated facts - experienced in imagetic form
(4) (nearly) simultaneously: related ventro-medial prefrontal linkages also activated
(5) as a consequence the emotional disposition apparatus is activated (e.g. in the amygdala)
(6) result of combined activations = approximate reconstruction of a previously learned factual-emotional set

Let us return to the patients with lesions in their VMPFC. These people mainly appear to have deficits in what we might consider as personal and social situations rather than traditional "objective" tests of intelligence and reasoning. How does this fit in with the

somatic marker hypothesis? Damasio's idea is that personal and social situations are frequently linked to punishment and reward. Punishment and reward are of course linked to pain and pleasure and these in turn to bioregulatory phenomena which are represented in the somatosensory system (1996, p. 1416). For Damasio, emotion is the representation of some homeostatic regulatory processes and is very much intertwined with pain and pleasure, whether or not we consider these to literally be emotions, as they are both bioregulatory and represented in the somatosensory system, and are thus somatic according to Damasio's usage of the term.

1.2 The 'Iowa Gambling Task' as empirical evidence for the Somatic Marker Hypothesis

Damasio's hypothesis is that normal responses to situations which are linked to the body in this way, i.e. responses that are linked to punishment and reward even if not overtly, are mediated by the linking stations in the VMPFC, that is, the somatic markers. If these linking stations are rendered inactive due to a lesion then the person still experiences the situation and the memory and can use the rest of their brain to figure out what to do, but they have no internal bodily response to the situation and this impairs their decision-making. We know that people with lesions in the VMPFC (a) at times take longer to come to decisions, and (b) make decisions that lead to negative consequences. Damasio and colleagues argue that this because the quick and easy method of being made aware of the future negative consequences by associating past similar experiences with negative bodily responses is not available to them and that this is shown in a laboratory simulation of real-life decision making; the Iowa Gambling Task.

The Iowa Gambling Task (henceforth IGT) was created to simulate the conditions by which people learn to play a card game to their advantage; in normal circumstances players only have available limited knowledge of contingencies and are subject to rewards and punishments (Bechara et al. 1994). The gambling task is set up to be a simplified version of card games, retaining these important features. It runs as follows. Four decks of cards are laid out: A, B, C, and D, and the subject is given a $2000 loan of play money. The subject must select one card at a time from any of the four decks until they are told to stop (they are not told how many selections must be made before the game is stopped, but the experimenters stop it after 100 selections), switching decks whenever, and as many times, as they want. The aim of the game is to maximize the profit on their loan.

When the subject turns a card over they always receive some money. The amount that they receive varies according to which deck it was taken from and is not announced until after the card is turned. Cards taken from decks A and B are rewarded with $100, and cards selected from decks C and D are rewarded with $50. Occasionally however a card is selected that gets the standard reward but in addition a punishment of being fined a certain amount of money. How much the subject is fined depends on which deck the card is selected from, and the position of the card in that deck. The fines from decks A and B are higher than from decks C and D, and the frequency of fines varies such that in decks A and C punishment is more frequent than in decks B and D, but these fines are smaller than the fines levied on decks B and D. What this means is that in the long run even though selecting just from deck A would yield more fines comparable to selecting from just deck B the same amount of money would be lost, and the same goes for selecting deck C over deck D. The result of the way the fining is set up is that by selecting cards from just A and B the subject would earn more money than if they selected from decks C and D, but they would also lose more money so that they would actually end up in debt (these are therefore referred to as 'bad' decks), whereas if the subject chooses from decks C and D (the 'good' decks) they will end up with a small profit as although each particular winning is much smaller so is the amount that they get fined.

Damasio and his team first used two different types of control subjects; normal subjects, and patients with brain damage but whose damage was not in the frontal lobes (Bechara et al. 1994). Both these groups followed the same steps. First they sampled cards from all the decks for a while. Then gradually they came to play more from the 'good' decks, that is decks C and D. And, then about half way through they adopted the strategy of playing mainly or only from the good decks and stuck with this strategy. These subjects therefore ended up making money on the task. In contrast the patients with damage in their VMPFC started like the controls by sampling all the decks for a while *but then played predominantly from the bad decks* (decks A and B). Despite continuing to lose money they would not amend this strategy and would have to borrow money to continue playing. It should be noted that the task is such that it isn't possible to calculate the net gains and losses generated from each deck as you play, Damasio and colleagues know this because they tried to get players with superior IQ to do this and they failed. That the task is achievable seems to be down to 'sensing' either overtly or covertly which decks are risky and which decks are profitable.

The results of this experiment show that VMPFC patients do not seem to learn from bad experience and thus change their behavioural strategy so as to avoid punishment or increase reward. This fits with the anecdotal evidence of their behaviour in real life situations. Damasio and his colleagues then repeated the experiment but this time recording the skin conductance responses (SCRs) of the participants (Bechara et al. 1996). While all the participants had SCRs that responded to loss of money and that did not respond to selecting a card from a good deck (C and D), the SCRs of the control subjects elevated in anticipation of selecting from the bad decks (A and B). This happened even before the control subjects settled on a strategy, so it seems that this is not linked to being *aware* of the danger of the bad decks. In contrast the VMPFC patients did not have elevated SCRs in anticipation of selecting from the bad decks. Damasio and colleagues conclude that the results of this experiment support the hypothesis that the link between the decision making apparatus and emotion is what is helping control subjects to 'sense' which decks are risky or not and thus come to a helpful strategy

I want to draw attention to the fact that what Damasio refers to as 'emotion' differs from the vernacular concept of emotion. For Damasio, an emotion is the representation of bodily changes in the somatosensory cortex. It would be better here to refer to this as affect, or as the representation of affective/internal bodily changes. We can thus see that the hypothesis is not so much that emotion is central to decision making, but that affect/internal bodily changes play a role in healthy decision making. As such Damasio is getting to the core of what interests me in this thesis; the role of the body in affect and cognition. However, this particular model of the relation between affect and cognition is one that I ultimately want to reject. In the rest of this chapter I will go through some criticisms and alternatives to the SMH which lead me to conclude that the SMH is neither the only nor the preferable way that we can explain the behaviour of VMPFC patients in the IGT. This is important because while the rest of this thesis will be devoted to grounding affect in bodily processes and telling a story of how it is likely to be partially constitutive of cognition, my story will place affect in a much more fundamental position in respect to cognition than the SMH does. The SMH proposes that affect is linked to the cognitive machinery in healthy functioning, and thus implicitly endorses a picture of cognitive function which holds the affective and cognitive machinery as separate (though interacting) faculties. I think that this underplays the role of affect in cognition and that recent work in the cognitive sciences can support a story in which affect plays a deeper role. For now however let us return to Damasio's hypothesis and consider some criticisms and alternatives.

1.3 Two somatic marker hypotheses

Colombetti (2008) is critical of the conclusions that Damasio and his colleagues draw from the experiments on VMPFC patients using the Iowa Gambling Task (IGT). She explains that there is a conflation in much of the literature between lack of decision-making and bad decision-making. This conflation is apparent both in descriptions of the behaviour of these patients in real life situations and in the discussions of the somatic marker hypothesis. That is, patients are described both as indecisive, requiring external factors to push them towards a decision, *and* as behaving impulsively with no regard to the long term consequences of their decision. Likewise, sometimes the somatic markers are proposed to be embodied preferences needed for *choosing among options*, and at other times they are proposed to be embodied preferences needed to *consider long-term outcomes of available options* (Colombetti, 2008, p. 53).

This ambiguity in what somatic markers are means that the somatic marker hypothesis itself is ambiguous and it is thus not clear what it is supposed to predict. Colombetti suggests that there are in fact two separate somatic marker hypotheses; the general and the specific. The general somatic marker hypothesis (SMH-G) is the claim that somatic markers "embody preferences that allow decision-makers to choose among options" (2008, p. 57). This hypothesis thus predicts that where somatic markers are missing (for example, if they are located in the VMPFC and this area is lesioned) then the patient will be unable to choose between options and so the behavioural signs of this will include procrastination, inactivity, the decision maker not having preferences, and becoming caught in cycles of endless reasoning (Colombetti, 2008, p. 57). In short, there will be no decision making.

In contrast to the SMH-G the specific somatic marker hypothesis (SMH-S) claims that somatic markers provide a "farsightedness" which is "necessary to consider and evaluate the long-term consequences of an option" (Colombetti, 2008, p. 57). This hypothesis therefore predicts quite different behaviour than the SMH-G. If the SMH-S is correct then in cases where the VMPFC is lesioned and the somatic markers put out of action there is no reason to suppose that decisions won't be made. These decisions will however be characterized by their impulsivity and "short-sightedness", that is, the patient would be concerned only with short-term outcomes.

Colombetti argues that whichever of these versions of the hypothesis that we take (SMH-G or SMH-S) we do not see evidence to support them with the results of the Iowa Gambling task experiments. Regarding SMH-G, the IGT appears to show that VMPFC subjects *do* have preferences and can act on them, which is entirely contrary to what the SMH-G predicts. Colombetti notes, "…if anything [the performance of VMPFC and amygdala patients during IGT] suggests that somatic markers are not necessary to choose among options and to implement a decision." (2008, p. 66). The VMPFC patients *are* decision-makers, they are just *bad* decision-makers!

On the other hand, the IGT provides no evidence for SMH-S as the task is set up such that maximizing profit doesn't require being able to consider future outcomes beyond the immediate step of which card to turn right now. At no stage during the experiment is the subject required to consider what the preferable long-term outcome must be and therefore take steps to achieve this that contradict the short-term aims (Maia & McClelland, 2004; Colombetti, 2008). Thus it may well be that VMPFC patients are short-sighted in regard to decision-making but the IGT neither shows this to be the case nor that it is not the case; it is methodologically unsuitable to test this hypothesis.

Colombetti's evidence for the methodological unsuitability of the IGT for testing SMH-S (that VMPFC patients are short-sighted when it comes to decision-making and thus more likely to select short-term gain irrespective of the danger this poses for their long-term gain) is that if one were to imagine how a subject who really was short-sighted would behave, that is a subject who was only interested in maximizing the gain on the next card, it looks as though they would behave like the control subjects rather than the VMPFC patients. Colombetti writes:

> Such a subject, using past rewards and penalties as a guide to what to expect on the next card, might very well favour A and B at first, and, once A and B start to yield more losses than gains, reverse this preference, exactly as normal subjects do (as Tables 2 and 3 show, in the original IGT the 'bad decks' are objectively advantageous in the early trials; deck B, in particular, yields the first loss only when subjects hit card 9). There is no reason to think that a subject concerned with maximizing long-term gain, and a subject concerned with maximizing expected gain on the next card only, would behave differently (Colombetti, 2008, pp. 60-63).

If Colombetti is right, then what is making the VMPFC patients behave differently from the control subjects is not that they are playing a short-term game rather than a long-term game; everyone is playing a short-term game because it is the same as the long-term game in this

set up. Something else must be happening. We know that the VMPFC patients are not choosing only from the first two decks merely because they are the first two decks since they, like the control subjects, start off by sampling all of the decks before then moving to predominantly selecting from decks A & B. If it is not the case that they are selecting at chance then this suggests that there must be a strategy being used. And, as Colombetti points out the strategy of selecting from the higher decks is a good strategy at the beginning of the task; for example, it is not until the 9th selection from deck B that the first loss is incurred (2008, p. 60). This suggests that the VMPFC patients are capable of *choosing a strategy* but the different behaviours of VMPFC and control subjects could arise from the VMPFC patients having a difficulty in *switching strategies*.

As discussed in the previous section, Damasio and colleagues place great weight on the role of anticipatory skin conductance responses (aSCRs) which occur in the control subjects but which are not exhibited by the VMPFC patients when they select from decks A and B. While they attribute this to anticipation of the loss that they may incur, Colombetti points out that Tomb et al., (2002) show that the anticipatory SCRs occur in normal subjects prior to anticipated wins of a substantial amount, so it is possible that the control subjects are anticipating the increased *wins* rather than the increased losses. In which case the original hypothesis that VMPFC patients are concerned only with immediate gain would be contraindicated given that they have no anticipatory SCR to that gain, and thus, according to the hypothesis no mechanism for selecting the deck which will award them the gain.

This leaves us with the following questions: (1) What is the mechanism that underpins the VMPFC patients' ability to develop the strategy of choosing from decks A and B in the first place? And, (2) why do VMPFC patients not switch strategies once it is clear that their initial strategy no longer can help them win, in the way that the control subjects do?

If it is not the case that aSCRs provide a warning signal for loss, but rather they indicate the possibility of winning a large amount, then these could be the mechanism by which the controls settle on the strategy of choosing from decks A and B until this is no longer a winning strategy. But as the VMPFC patients do not exhibit aSCRs this cannot be the mechanism by which they come upon this strategy. Given that the behaviour of both groups in the beginning is the same it seems reasonable to suppose - in the absence of other evidence - that the mechanisms underlying this behaviour are the same in both groups. In which case aSCRs may play no role at all in the choice of this strategy.

That the deficit in VMPFC patients may be to do with a difficulty in switching strategies once one is selected would fit in with accounts of their behaviour in non-laboratory conditions. Firstly it is not at odds with descriptions of these patients having difficulty coming to a decision in the first place. VMPFC patients exhibit this difficulty but as has been noted they still do come to decisions all the time, even if those decisions are bad ones. It seems reasonable to assume that these decisions are come to as a result of external factors, though which external factors play the greatest role in this is left open. We know for example that the patients often go against the advice of their friends and family, so it is not just that they are suggestible; they seem to be suggestible in certain contexts, or perhaps particularly suggestible to certain things. Regardless, being in an experimental environment with a set task and clear instructions as to what is expected of them, they are able to make decisions, and it seems reasonable to interpret these constraints as external pressures[8].

This might explain why the VMPFC patients engage in the task and come to a strategy in the first place. Once they are involved in the task however and have started accruing winnings the external pressure of continuing to win may not suffice to change the strategy. We are not told whether it is distressing for the patients to lose; whether they exhibit negative emotions as a result of losing. If they do not then this would indicate there isn't much pressure for them to win. Technically they have already succumbed to the external pressure and therefore, if I am right in thinking that this is the key to their behaviour in the task, they have no motivation to change their strategy unless (1) not changing the strategy causes enough distress that this becomes sufficient motivation, or (2) the task is interrupted once selecting from decks A and B becomes disadvantageous and instructions are once again given to the subject that what is required of them is to choose such that they maximize their winnings. By doing this, then the VMPFC patients may well adjust their strategy.

If this is right then (contra Colombetti) this gives a way for SMH-G to be right after all (though she is still right that the IGT is unsuitable for validating the hypothesis). Recall, SMH-G was that somatic markers embody preferences used in decision making, and therefore it predicted that a lack of somatic markers would mean an inability to choose.

[8] This might be testable by placing the VMPFC patients in a room with the experimental set up but no instructions and no hint as to what was expected of them. If they engaged in the task and performed in the same way it would suggest that external pressures were not key to their making decisions in this task. If however they did not engage with the task or played arbitrarily then this would provide evidence in favour of this hypothesis.

While Colombetti is right in pointing out that VMPFC patients do in fact make decisions, albeit bad ones, this does not contradict SMH-G. All SMH-G predicts is – everything else being equal – a lack of somatic markers would result in a person being unable to come to a decision. However I suggest that in most situations in life everything else is not equal. There are constant external pressures on everyone to behave in certain ways. Think, for example, of young children. Much of the time we would not consider them to be coming to a purely internally-motivated decision to do what they do. Mostly, constraints and expectations are made clear and as they change the changes in expectations need to be made explicit; what may be acceptable behaviour in a playground is not acceptable in the supermarket, it may be not so much that small children choose to behave in differently in these contexts but that the contextual information and reinforcement that they are given in each provide some constraints as to what behaviours are available for that context.

Should we think of this still as decision-making? The primary difference seems to be that adults internalize these pressures and have the capacity to have their behaviour automatically guided by them. We might stipulate that those who cannot internalize such pressures, such as young children or the neurologically impaired, fail to count as making decisions. In that case, though we should acknowledge the constraining definition of decision-making and acknowledge that just because someone lacks this capacity, this does not mean that the person will not act in such a way as to appear to be making decisions much of the time. And, such a person, without having access to internal pressures, in the absence of external pressures will exhibit exactly the kind of indecisiveness that the VMPFC patients are described as having in everyday life[9].

It might seem that the impulsivity that is described in VMPFC patients is contradictory to their deficits in making choices. This depends on exactly what behaviours are being described as impulsive. It might be that the behaviours that follow from the patients succumbing to external pressures are considered impulsive, because the patients have not thought through the consequences and thus come to a decision based on cost-benefit analysis (or using the somatic-marker shortcut for this). Alternatively, it might be that the actions that are 'chosen' are deemed impulsive as a value-judgement because they are often different to those that friends and family would have chosen. Lack of attention span for example can

[9] A nice example of this indecisiveness in everyday life is given in Damasio (1994) who describes a patient's attempt to decide on a date and time for his next appointment, and in the absence of clear reasons to choose a particular slot continues to deliberate for over half an hour until the secretary has to help him make a choice.

look like impulsivity. We know that the PFC is involved in being able to temper impulsivity – this ability grows as the white matter to this area develops over adolescence. So it makes sense that damage to this area would result in impulsivity, even if impulsivity is not just a different description of the way that VMPFC patients decide, i.e. using external pressures rather than internal.

So lets play Colombetti's game and imagine what a VMPFC patient would look like if the SMH-G was right:

(1) they would have difficulty making decisions if there were no overt pressures on them towards one decision or another
(2) they would be guided by whatever the strongest external pressure on them was (whatever that strength may consist in), and thus choose options that we would deem inappropriate or suboptimal (if indeed we should call this 'choosing' at all).
(3) this behaviour/method of choosing may well appear "impulsive"

All of this is exactly what is described. Thus, the SMH-G doesn't need to predict indecision, it may predict a deficit in switching strategy instead. This deficit in strategy-switching will present as perseveration, as we see in the VMPFC patients.

1.4 Perseveration in the IGT

Rather than considering VMPFC patients to be impulsive, or making bad decisions, I think that it would be useful to see them as perseverating on the task (perhaps as a result of a deficit in strategy switching as suggested above) and so compare their behaviour with classic perseverative behaviour. The 'A not-B error' task is a famous task in developmental psychology developed by Piaget (1963), which he used to show that infants of 10 months and under show a lack of a grasp on object permanence. The task runs as follows. The infant can see two hiding locations; A and B. The experimenter hides a toy in location A (in a well or under a cup), a pause is given and then the infant is given the opportunity to reach for location A or B. The infants consistently reach for location A showing that they understand that the experimenter hid the toy there. After repeating this several times the experimenter changes strategy and hides the toy in location B. As with the previous trials a pause is given and then the infant is given the opportunity to reach for either location A or B. Rather than reaching for location B however, the infant tends to reach again towards location A despite

having just seen the experimenter put the toy in location B. That is, they perseverate, continuing to use the strategy that they used before; choosing A and not-B (thus the task is either called the 'A not-B error' or the 'perseveration task'). As infants 10 months old and younger perseverate in this task but infants 12 months and over do not Piaget concludes that 10-12 months is the age at which infants develop a schema for object permanence. [10]

Whether or not it is right that the infants' behaviour in this task indicates a development of a schema for object perception, it remains unclear what the mechanisms are for this shift in strategy that we see between 10 month olds and 12 month olds. However, Smith and Thelen (2003) show that, by small changes to the experimental set up they could affect the choice of the infants so that they no longer consistently perseverate. Under Smith and Thelen's dynamic field model this is because rather than there being a single cause of the infants reaching towards A or B, there are many changes in the whole infant system that self-organise into a behavioural expression. When the experimenter directs the infant's attention towards location A by hiding the toy there this increases activation at A in the dynamic field (where the dynamic field is a description of activation changes in the infant). If this activation crosses a certain threshold then it becomes likely that when given the opportunity the infant will reach for that location. Upon reaching for a particular location a 'memory' of that activation is created and this in turn becomes an input at the next trial. This makes it more likely that, if everything remains equal, the infant will once more reach to that location. Every time the infant does this stronger 'memories' of the action are formed; that is, the field shifts so it is easier and easier for it to slip into activation state A. Thus, the explanation for the A not-B error is that while, if the infant is given the opportunity to reach for B straight after seeing the experimenter put the toy in location B, the activation levels still going in will push the infant to reach towards B, when a pause is given this allows the activation levels to settle back to the state which favours reaching towards A.

By (1) changing the timing of the delay, (2) making the covers of the hiding spots, or the hiding itself, more attention grabbing, and (3) changing the amount of prior reaches to A, the experimenters have managed to manipulate when the error occurs. And, of particular interest to fans of embodied cognition, the appearance (or disappearance) of the error can also be manipulated by making the reaches on the B trial different from the reaches on the A trial by changing the posture of the infant, for example, by changing the infants position from sitting

[10] Note that the delay in this task between the experimenter hiding the toy and the infant being given the cue to reach is crucial, without this delay infants do not perseverate; they reach towards the correct location of the toy.

(on the A trial) to standing (on the B trial – through the hiding, pause and search). According to Smith and Thelen's model this is because the differences "decrease the influence of the A trial memories on the activations in the field" (2003, p. 346). This change allowed 8-10 month olds to match the behaviour of 12 month olds on the task. A similar experiment was done with different wrist weights on the infants in each trial, and again the error was disrupted and the 8-10 month old infants behaved as the 12 month olds. Smith and Thelen say of these results:

> These results suggest that the relevant memories are in the language of the body and close to the sensory surface. In addition, they underscore the highly decentralized nature of error: the relevant causes include the covers on the table, the hiding event, the delay, the past activity of the infant and the feel of the body of the infant. (Smith & Thelen, 2003, p. 346)

It is the latter suggestion that is of interest to us just now; the "highly decentralized nature of error". It is this lesson of decentralization that, more than anything, needs to be taken into account by Damasio and his colleagues when seeking a hypothesis to explain the behavioural differences in patients with lesions in their VMPFC. Let me summarise my view of the relation between the VMPFC's behaviour in the IGT, and the 8-10 month olds behaviour in the A not-B error task. I have argued in this chapter that SMH-G predicts not mere indecisiveness (as Colombetti proposes) but rather a difficulty in strategy switching, which will be expressed in some situations as perseveration, even when the individual knows that a different strategy would be more successful. In the case of the A not-B error task the infants are made clearly aware that the toy has been hidden at location B yet they still proceed to reach for location A. In the IGT the patients are rationally competent to judge that selecting from decks C and D is safer and yet they still proceed to select from decks A and B.

In the A not-B error task the perseveration can be manipulated by changing certain dimensions of the dynamic field, such as by changing the infant's posture. I suggest that although we don't yet know which dimensions to manipulate, the perseveration of VMPFC patients in the IGT will be similarly manipulable, and that this will show that error in this case is also (like in the A not-B error) highly decentralized, and thus not purely due to the somatic markers. We will see evidence of this in section two of this chapter in which I present alternatives to the SMH. I will argue that the somatic marker hypothesis (whether it means to or not) oversimplifies the mechanisms which underlie decisions, emotions, and behaviour. This oversimplification arises from the theory's situation in a tradition in which

uni/mono-causality is the pervasive explanatory strategy and a tendency to think of cognition and affect/emotion as separate faculties.

1.5 Summary of section 1

I have argued that the Somatic Marker Hypothesis, interpreted as the claim that somatic markers embody preferences that allow decision makers to choose among options (rather than providing "farsightedness") which Colombetti refers to as SMH-G, is not necessarily wrong as it predicts deficits in strategy switching (not simply indecisiveness), which I have shown is not contradicted by the results of the IGT. However the SMH-G does seem to be *incomplete* as:

(1) aSCRs are evident in normals prior to selecting winning decks and not in VMPFC patients which, given my argument that VMPFC patients can and do decide on the strategy of picking the winning decks before the switch is needed, suggests that if aSCR's are a mechanism then they are not the only mechanism by which we decide or choose a strategy. This in itself does not mean that aSCRs do not perform this function in normal subjects, but it does count in favour of the multi-causal explanatory strategies such as those advocated by Smith and Thelen.

(2) The somatic marker hypothesis frames the role of the somatic marker as the link to emotion or emotion memory that helps reasoning or decision making because it constrains the space of choices. It is not clear that just linking to autonomic processing is a link to emotion, nor is processing in the cortex necessarily non-emotional (specific examples of this will be discussed in chapter four). Damasio is of course aware of this, indeed his stance on emotions is that they are cortical representations of certain somatic changes, and can therefore be activated not only receiving information about bodily changes but also by activating those cortical representations themselves. However we can still see the vestiges of a paradigm which holds that affect and cognition are separate but interacting faculties throughout the hypothesis: the somatic markers are hypothesized to be in the PFC and thus just act as a "link" to affective areas, rather that the PFC or the affective processing being constitutive of the decision-making apparatus. And, as I mentioned above, the role of the representation of bodily changes (via the link that is the somatic marker) is to constrain the cognitive possibilities. Thus, while Damasio is arguing that emotion (affect) is necessary for normal cognitive processes such as decision-making, he is arguing within and perpetuating

further the bias that emotion (affect) and cognition are separate faculties and have different areas dedicated to them.

Section 2: Alternatives to the Somatic Marker Hypothesis

In the first section of this chapter I argued that while the SMH is not necessarily wrong, it is incomplete and may be an unhelpful way to think about the relation between affect and cognition. This gives us reason to be sceptical of the SMH. In this section, I will consider an alternative model which does not rely on the artificial distinction between affect and cognition. This is not to say that affect is not involved in decision-making; I will argue in chapters four and five that affect is integrated in all cognition. Rather, affect need not be so coarsely appealed to as feeding into decision-making in the way that the somatic marker hypothesis envisages.

2.1 Reversal learning: An alternative to the SMH

Maia & McClelland (2004; 2005) argue that there is no evidence in favour of the Somatic Marker Hypothesis (henceforth SMH). They argue that there is good evidence to suggest that the behavioural impairments exhibited by VMPFC patients are a result of deficits in Reversal Learning (RL). RL is the ability to switch from one approach to a task to another following a change in reinforcement. A case of RL would be if I was consistently praised (or rewarded in some other way) for choosing to attend classes and then once I had learnt this association, the praise was awarded only for *not* going to classes. Having no deficits in RL I would quickly learn that to obtain praise I must desist from attending class and (providing that no greater reinforcement, such as a lust for knowledge, encouraged me to stay) I would stop attending class, thus switching my strategy.

RL may underpin the impairments of VMPFC patients in the Iowa Gambling Task (IGT), which Bechara and Damasio consistently use to support the SMH. If it were the case that somatic markers were not involved but the patients had a deficit in RL then, if the IGT were recreated just as in Bechara and Damasio's experiments, just the same data would be collected. That is to say, the results that Bechara and Damasio get from the IGT task on VMPFC patients can be just as well interpreted using the hypothesis of impairment of RL, because all we see in the original experiments is that the VMPFC patients settle on one

strategy (choosing the highest paying decks – which also happen to be the decks with the highest fines) and stick to it.

Maia & McClelland (M&M, 2004) argue that this alternative interpretation of the IGT is supported by a significant amount of other research. Rolls et al (2004) showed that patients with ventral frontal damage could report when contingencies in a task had changed and yet fail to adapt behaviour, just as we see in VMPFC patients in the IGT. This, M&M claim, is consistent with studies of reversal learning in animals with ventral frontal damage (Rolls et al. 1994; Rolls 1999; Schoenbaum et al. 2002; cited by M&M 2004, p. 16080). More specifically Fellows and Farah (2003; 2004) have shown this in VMPFC patients – thus Bechera and Damasio cannot claim that the results of Rolls et al. and the animal research were due to lesions of a slightly different area in the PFC. Fellows and Farah (henceforth F&F) show that VMPFC patients show normal acquisition of learning, but impaired reversal (F&F, 2003). In their (2004) paper F&F present a shuffled version of the IGT, in which everything is kept the same as in the original IGT except that instead of there being an initial advantage to the first 2 decks (recall these decks initially paid out large wins and only later resulted in large fines), there is no initial advantage. If the SMH were correct this ought not to make a difference as the same amount and value of cards is being chosen, so the VMPFC patients ought to still favour the first two decks. But F&F show that without the initial reinforcement of those decks, by means of them giving huge winnings and no fines for the first nine cards or so, the behaviour of the VMPFC patients is indistinguishable from the controls (see M&M, 2004, p. 16080). This result is compatible with the hypothesis that the VMPFC patients have a deficit in RL as in the shuffled task there is no need for them to do any reversal learning as they do not learn the "incorrect" pattern in the first place.

Finally M&M argue that F&F (2004) and Rolls et al. (1994) show that the deficit in reversal learning of these patients correlates with the level of impairment in daily functioning (M&M, 2004, p. 16080). This is particularly important as impairment in daily functioning was what the SMH was designed to explain; the IGT was created to be a laboratory experiment that as closely as possible simulated the natural decision making requirements faced by VMPFC patients (and of course neurotypical people) in their day to day lives.

M&M (2005) provide more evidence in favour of the RL hypothesis. They cite Fellows and Farah's (2005) study in which F&F present VMPFC patients with a new version of the shuffled IGT. In this version the decks are switched round so that the first two decks are the

“good” decks rather than the “bad” decks. Thus, even though the amount of reward and punishment that each deck provided stayed the same as in the original IGT (and in their first shuffled IGT) there is no longer an initial advantage to the first two decks as there was in the original IGT, but rather the initial advantage goes to the second two decks (the ‘bad’ decks). That is to say in the new version of the shuffled task the first two decks earn the subject less money but also fine them less money and so enable a greater overall gain. Again, as with their initial shuffled IGT, in this case the VMPFC patients performed similarly to controls. This result is problematic for the SMH which predicts that even in a shuffled version of the IGT the VMPFC patients would have difficulties and would end up selecting mainly from the bad decks. However it supports the RL hypothesis as yet again, without the VMPFC patients being required to perform reversal learning they perform no worse than neurotypical subjects. This suggests that the RL hypothesis is not only an alternative hypothesis to SMH but is a preferable hypothesis.

The Reversal Learning Hypothesis is not necessarily inconsistent with the Somatic Marker Hypothesis; for example it may be the case that RL involves somatic markers. However, M&M (2005) argue that this is not so. Their reasoning is that the VMPFC not only has projections to the autonomic areas (which the SMH relies on) but it also has direct projections to the striatum (caudate nucleus and putamen). Given that the somatic loop that is the main underpinning of the SMH would be noisy and inefficient and that projections from the VMPFC to the striatum could directly guide action selection (M&M 2005, citing Rolls, 1999), the SMH seems superfluous to requirement. Additional evidence from Rolls is that when the regions of the striatum that receive projections from the VMPFC are lesioned, the same deficits occur in RL that are apparent as a result of lesions in the VMPFC. There is thus no need for postulating somatic markers.

Given that the VMPFC seems to generate both autonomic responses and signals that guide behaviour using different routes, these ought to be doubly dissociable within the VMPFC (M&M, 2005, p. 163). While we don’t have enough evidence to conclude this for the VMPFC, M&M explain that such a dissociation has been found in the amygdala, the central nucleus of which projects to those areas that underpin autonomic and reflexive responses. The basolateral amygdala, on the other hand, projects to areas underpinning the control of instrumental behaviour. This is just the kind of dissociability they expect to see in the VMPFC as well. However, even the double dissociability of the amygdala is in itself a problem for the SMH as “if bodily states were important in guiding instrumental behaviour,

interfering with the generation of those states should affect instrumental behaviour" (M&M, 2005, p. 163) and yet the double dissociability of these structures means that lesioning the part of the amygdala which subserves the autonomic responses does not affect the functioning of the part which subserves the control of instrumental behaviour.

2.2 Behavioural strategies in decision making

The advantage that the SMH had over the RL hypothesis on its own is that RL by itself does not fully explain decision making. In the SMH emotion (affect), realized by the activation of the autonomic areas, gave rise to the valence and motivation required to activate a behavioural response. Motivation may however be accounted for if we think in terms of behavioural strategies. M&M suggest that a simpler way of understanding behavioural strategies than appealing to the SMH would be to understand them in terms of *exploitation* and *exploration* (M&M, 2005, p. 163). The idea behind these strategies is that decision-making in an uncertain environment will involve a switch between exploitation and exploration. Reliance on just one of these behavioural strategies will be suboptimal; although exploiting a situation (sticking with what seems like a good situation) might seem a good strategy, in fact optimal behaviour requires some exploration as well - as Daw et al. (2006) explain: "…exploration is often critical for organisms to discover how best to harvest resources such as food and water." (Daw et al. 2006, p. 876). The information gathered by exploratory behaviour enables the agent to then exploit it.

The hypothesis about exploitation, exploration and decision-making is that exploitation is the "default" behaviour. Decision-making is the capacity to voluntarily switch from the exploitative tendency to an exploratory mode of behaviour. The relation of RL to this should be already clear; for a system to be able to reverse learn in the first place it must desist from exploiting its environment and learn a new contingency. Likewise without the capacity for RL a system that explored the environment would not learn from this exploration and then be able to exploit the new information.

Daw et al. (2006) provide evidence for the exploitation/exploration hypothesis of decision making by using a model drawn from machine learning and using it to inform a gambling experiment which they created in order to attempt to be able to identify particular areas of the brain involved in each strategy. They note that research has suggested that exploitation (also known as "appetitive choice") is underpinned by a dopaminergic, striatal, medial,

prefrontal network. However the neural substrates of exploration are less clear. Their experiment is therefore an attempt to disentangle the substrates of exploration from those of exploitation.

The task they used was a 'Four-armed-Bandit' task. This was essentially a simulation of 4 slot machines any of which the participant was allowed to choose from. The participants won points from the machines, which were later exchangeable for money. The slot machines had different mean payoffs and these changed randomly and independently from trial to trial, so at each trial the participant had to learn again which would be the best machine(s) to win the most money.

The experimental set up that Daw et al. use is particularly interesting because the neural activation underlying exploration and exploitation are difficult to dissociate. They therefore decided to exploit the tight coupling between computational modeling, behavioural analysis and functional neuroimaging. The experimenters used the participants strategies for the trials on the 'Four-armed-Bandit' task to see how they compared to 3 well-known Reinforcement Learning strategies: (1) e-greedy, in which mostly the best option is chosen but occasionally there is a random action; (2) softmax, in which the decision to explore and which suboptimal action to take is determined probabilistically on the basis of expected values; and (3) softmax with uncertainty bonus, which is the same as softmax but a reward is also given for choosing actions that have uncertain consequences but that will be informative (Daw et al. 2006, p. 876). The softmax bonus is considered to be the optimal strategy for reinforcement learning tasks in which the optimal solution is computationally intractable.

When they compared the fit of these models to the subject's behavioural choices they saw strong evidence for value sensitive (softmax) over undirected (e-greedy) behaviour. However, surprisingly, they did not see significant evidence to justify an extra parameter for uncertainty bonus. Once they had decided that softmax was the model with the best fit, they used it to: "generate regressors containing value predictions, prediction errors and choice probabilities for each subject on each trial." (Daw et al. 2006, p. 876). They used this methodology to identify which brain activity reflected which kind of action chosen; exploratory or exploitative. The idea behind this was that *if the actual choice was predicted by their model then the choice was exploitative*. If however the actual choice was predicted by the model to have a lower value, then that would be taken as evidence that it was an exploratory choice.

They did not find any areas that had higher activity for exploitative decisions rather than exploratory decisions, which would support the hypothesis that exploitation is the default mode of behaviour. However, some areas were more active when the participant made what they have assumed (by the above) to be exploratory decisions. These were the right anterior frontopolar cortex and the anterior intraparietal sulcus (differentially to exploitative decisions for which regions of the striatum and VMPFC were active). The right anterior frontopolar cortex is associated with high-level control, and so Daw et al. suggest that "the signal we observed in anterior frontopolar cortex could reflect a control mechanism facilitating the switching of behavioural strategies between exploratory and exploitative modes" (p. 878) differentially to exploitative decisions for which regions of the striatum and VMPFC are active.

While it is not so clear exactly what role the anterior intraparietal sulcus is playing in this model, it is known to be implicated in decision making in both humans and primates, and they note that different subregions of this area have different output modalities, so it seems likely that while the right anterior frontopolar cortex may underlie the capacity to switch strategy to exploration, the anterior intraparietal sulcus may be involved in generating the exploratory behaviour. Interestingly they note that the anterior border of the sulcus, which is close to the exploration-related activation, is associated with grasping and manual manipulation. This seems apt given that the first forays into exploration that we do as infants predominantly entails grasping and manipulating objects with the hands.[11]

2.3 Contextual and episodic control in decision-making

So far, the research I have outlined suggests that reversal learning is needed for decision-making, and that decision-making may be borne out by the behavioural modes of exploitation and exploration. We have a new way of understanding valence in terms of value sensitive behaviour but as it no longer entails motivation we seem to be missing at least one step in a model of decision-making. Motivation is required for guiding the selection of behaviour. Without it, it is unclear why the switch from the default behavioural strategy of exploitation to the exploratory strategy would occur. And, under the model presented so far, decision-making just is this capacity to voluntarily switch strategies.

[11] This is of course very likely much before the capacity to voluntarily switch between exploitative and exploratory strategies would be in place.

Kouneiher et al. (2009) investigated the neural correlates of episodic and contextual control in decision-making and the influence of episodic and contextual motivational cues on these areas. These cognitive stages of decision making come from Koechlin et al.'s (2003; 2007) cascade model of control. Contextual control is the use of a current cue for selecting task-appropriate behaviour, whereas episodic control is the use of a cue in the past which determines, for a period of time (an episode) how cues will be interpreted (Egner, 2009, p. 821). Egner explains the distinction between contextual and episodic control with the example of eating in a restaurant with someone. If both of you are eating, when the waiter brings you your food the cue of your friend still without his dinner causes you to override the urge to tuck in, and you wait until he receives his food too before you eat. Your friend's presence is the contextual cue that stops you from starting to eat straight away. If however, your friend had informed you that he was not going to be eating with you and was just keeping you company, his presence would not provide this contextual cue. That he is not going to partake in the meal provides a way in which all other cues for that period of time will be interpreted. It is an episodic cue; shifting the space of possible cues within that episode. Egner explains, "both contextual and episodic control signals allow us to transcend habitual stimulus-response associations, but they do so in different temporal frames and they are arranged hierarchically: episodic control affects contextual control, but not vice versa." (Egner, 2009, p. 821)

Kouneiher et al. (2009) use an information-theoretic informed fMRI study to investigate the neural basis of the neural control signals and how they are integrated with motivation signals. Koechlin et al. (2003) had shown that "the mid-lateral PFC was only activated by episodic control, whereas posterior lateral PFC activation increased with both episodic and contextual control demands" (Egner, 2009, p. 822). Kouneiher et al. (2009) replicated this data with the addition of manipulations to the experiments used so that half of the trials were associated with motivational cues (a cue that signalled that a correct response would result in being awarded money and an incorrect response would result in a fine). They managed to manipulate contextual and episodic motivational cues by making some of the blocks of trials that were subject to reward and punishment high incentive and others low incentive so that these provided motivational cues that the subjects were either in a sustained high or low stake environment, or an environment which was just undergoing mere transient increase or decrease in stakes (Egner, 2009, p. 822). Half of the trials had the motivational cue which signalled earning or losing money, the other trials did not have this cue and the subjects

neither earned nor lost money in these. Of the trials which were given the motivational cue some of the blocks were low-incentive so that the gains and losses were only small, and some blocks were high-incentive with the gains and losses being substantial. The idea behind this set up was that the extra trials provide contextual motivational cues which signal an increase in stakes from the trials in which there was no reward or punishment. So the fMRI scan would pick up which areas were lighting up in response to the trial being a motivational one rather than one without any motivational cues. This would indicate which area was active when someone is registering a motivational cue. Within the motivational trials there are episodic motivational cues which indicate either that they are in an environment with high stakes or low stakes. The difference in these trials will show up in the fMRI scan and so we are able to see which areas are underpinning the episodic (as opposed to the contextual) cues. Egner explains:

> Kouneiher et al. employed functional magnetic resonance imaging (fMRI) to assess how regions of the human lateral and medial prefrontal cortex (PFC) mediate the relationship between cognitive demands and motivational considerations. Their results suggest that motivational incentives engage areas in medial PFC that in turn energize regions of the lateral PFC, which are involved in selecting task-appropriate behavior. Furthermore, both medial and lateral areas appear to obey an anterior-to-posterior (rostro-caudal) functional organization, according to whether representations of cognitive and motivational task parameters stem from temporally distal or proximate cues. (Egner, 2009, p. 821)

Their results showed that parallel areas in the medial and lateral prefrontal cortex may be responsible for motivation and selection of behaviour. The areas responsible for motivation are both in the medial prefrontal cortex (near the midline of the PFC) parallel to the midlateral PFC and posterior lateral PFC involved in selection of behaviour (episodic and contextual respectively). The area involved in episodic motivation showed up as the dorsal anterior cingulate cortex, while the one involved in contextual motivation showed up as the pre-supplementary motor area. The former retain "incentive values" of past events and energise or inhibit the lateral prefrontal resources which guide action selection based on past events. While the latter "evaluate immediate contextual incentives for action and energize (or inhibit) lateral prefrontal resources that guide action selection according to immediate contextual signals" (Kouneiher et al., (2009) p. 944):

> In this system, functional interactions from medial to lateral regions convey motivational incentives rather than control demands and regulate the relative influence of immediate and past information in the cascade of top-down selection processes operating in the LPC. (Kouneiher et al., (2009), p. 944).

The connectivity analysis between the areas shows that episodic control has primacy when the midlateral PFC is active as it is associated with top-down excitation of posterior lateral PFC (Egner, 2009, p. 822). This means that the episodic control shapes the space for the contextual activation. However there need not always be episodic activation, there can be contextual activation on its own. Episodic activation is used only when the agent needs to be using past information to guide action, and this is not always needed; current contextual cues can often suffice.

In summary, Kouneiher et al.'s research shows that both areas of lateral PFC are involved in task appropriate control. The posterior part guides task appropriateness by current context, whereas the anterior part guides task appropriateness by referring to a temporally distal cue to guide task appropriate behaviours and when this is active this can guide how contextual cues are then interpreted. Each control area is energized by a parallel area in the medial PFC; excitation of the lateral areas by the medial areas provides the needed motivational incentives for task selection. Conversely the medial areas can also inhibit the lateral areas. The medial areas not only evaluate incentives for action, and energise the lateral areas, but the dACC (the medial area that energises episodic control) retains incentive values for past events (Kouneiher et al. p. 944).

Although Kouneiher et al. and Egner do not mention VMPFC deficits, as their research is done on neurotypical subjects, there looks to be a connection here. If the dACC retains incentive values for past events that helps it to energise the episodic control centre, which allows the agent to override the contextual responses and project appropriate task control over an episode then, if the dACC or the links between the dACC and the lateral prefrontal regions were damaged, we would see deficits in episodic task control such that past events, even if learned, would not inform action selection. This is of course just the type of deficit we see in VMPFC patients; they do not seem to be able to make judgements that require them to think about a whole episode rather than just respond to contextual cues.

This way of understanding the deficits makes the accusation of VMPFC patients as short-sighted that I outlined earlier (and was specifically acounted for by Colombetti in the SMH-S) more understandable. The VMPFC is a separate area than the dACC which Kouneiher et al. propose is involved in energizing the episodic control area. However, they are anatomically very close by each other and are both involved in a number of circuits. It is not unlikely that (i) damage to the VMPFC consistently also damages the dACC, or at least that

(ii) damage to the VMPFC disrupts circuits crucial to the ability of the dACC to energise the mid lateral PFC. So, I suggest that the behaviours of VMPFC patients can be plausibly explained accordingly by appealing to deficits in episodic control, I will develop this argument further in the next section.

2.4 VMPFC deficits and mental time-travel

In a very similar vein to the research outlined above, Gerrans (2007) proposes that we replace the SMH with the "mental time travel hypothesis" (henceforth MTT). Mental time travel is "the ability to retrieve past episodes and imagine future ones and integrate the results with other forms of knowledge as part of planning" (Gerrans, 2007, p. 469). Gerrans suggests that Damasio and colleagues focus on semantic representations rather than episodic representations (that representations of semantic knowledge (knowledge about facts) are marked with emotion, and these guide decision making). This kind of cognition however makes us reliant on contextual cues; it would mean we were only able to make decisions by integrating environmental stimuli and factual knowledge about that stimuli. This would mean that we were behaviourally inflexible in that our behaviour would be constrained to reacting to current stimuli rather than following anything that we would ordinarily consider to be a 'decision'.

Gerrans argues that planning/decision making requires the ability to voluntarily create and recreate experiences. The ability to voluntarily control the *re*-creation of experiences depends on voluntary access to episodic memory rather than declarative memory; the ability to retrieve past episodes as opposed to factual knowledge. The ability to voluntarily control the *creation* of experience is the ability to voluntarily "imagine ourselves living out future scenarios, rehearsing different possibilities", which like episodic memory is about the creation (re-creation in the case of memory) of experiences rather than facts. Known as 'prospection', it is a "future-directed analogue of episodic memory" (Gerrans, 2007, p. 469). Gerrans cites Klein and Loftus et al. 2002 and Wallis and Miller 2003 as providing evidence that imagination uses the same systems as episodic memory; "the activation of relevant perceptual sensory and emotional systems in the absence of an environmental stimulus" (Gerrans, 2007, p. 469). And, as Gerrans goes on to say "[t]he consequences for planning are enormous. Mental time travel gives humans an enormous database of situations and responses to them which can be safely rehearsed off line" (Gerrans, 2007, p. 469. This stands

in contrast to the mere ability to "perform cost benefit analyses by manipulating probabilities expressed as propositions" (Gerrans, 2007, p. 471).

On Gerrans' model mental time travel is the ability to voluntarily access the database of episodes either to re-activate them - as they are - or to integrate them with other forms of knowledge and prospect over what could be or might have been. Damage to the episodic database therefore would cause deficits in mental time-travel. What is most relevant to us, however, is that on this model deficits in mental time travel could also be a result of damage to the systems for the voluntary access of the episodic database, rather than the database itself. Voluntary control is an executive function, known to require the frontal systems; and these enable the retrieval of episodic information (Gerrans, 2007, p 474). Damage to these frontal systems therefore impairs the ability to "re-experience a study episode in sufficient detail to recollect contextual information about that episode" despite the ability to "report about the factual contents of the same episode" Wheeler et al. 1997, p. 342, cited by Gerrans, 2007, p. 474).

Gerrans' hypothesis is that damage to the VMPFC impairs the ability to voluntarily access the episodic database and thus to "use past experience to genuinely inhabit the future" (Gerrans, 2007, p. 471); thus VMPFC patients, while able to access semantic knowledge, are unable to perform mental time travel. On both Damasio and Gerrans' models the subject doing the IGT imagines what will happen in the future based on past experience. On Damasio's model, when the participant contemplates one of the decks of cards (which could be 'good' or 'bad' depending on whether it yields high or low gains and losses) somatic markers associate a valence learnt by past experience to that possibility, thus guiding choice. Damage to the somatic marker apparatus thus leaves the participant without access to the guidance from previous learning. In a manner of speaking, this constitutes imagining what will happen in the future based on past experience. It is however far removed from what Gerrans considers to be mental time travel; the ability to "genuinely inhabit the future". On his model an account of what goes on in the IGT would run as follows:

> She imagines what would happen in the future based on her past experience. Thus, in the IGT, faced with the choice from deck A, she recalls previous experiences with that deck. In so doing of course she activates the relevant emotional associations which then become available for imaginative rehearsal of the consequences of choices i.e. she imagines choosing deck A, losing a large amount of money and generates an aversive affective response either in the process or as a consequence of the process of constructing the relevant implicit or explicit imagery. (Gerrans, 2007, p. 471)

Clearly, though much of the same neural machinery is involved, this is a much more complex hypothesis than that envisioned by Damasio and colleagues, yet it would just as well account for why damage to the those brain areas which support voluntary access to episodic information and thus allow imaginative prospection, result in the behaviours that VMPFC patients exhibit in the IGT. In addition it is entirely consistent with the fact that VMPFC patients have no deficits in declarative reasoning; does not rely on SCR data which has been so controversial recently in regard to IGT participants; and relies upon the known capacity of the frontal cortex for voluntary, executive control rather than the hypothesized existence of somatic markers.

It should be clear now that MTT predicts that VMPFC patients would perform poorly on a mental time travel task. While experiments which investigate mental time travel capacities in VMPFC patients have yet to be done, the performance of these patients in the IGT which, at least on the way presented above, requires some mental time travel, is consistent with this prediction. The other prediction which Gerrans suggests the MTT generates is that "patients with a deficit in mental time travel would perform poorly on the IGT irrespective of their ability to generate SCRs to disadvantageous decks" (Gerrans, 2007, p. 471). If this is borne out, it would be very problematic for the SMH as it is inherent in the hypothesis that the generation of SCRs in response to disadvantageous decks is the mechanism (or at the very least an expression of the mechanism) of healthy decision making. Thus, if someone is able to generate SCRs to disadvantageous decks and yet still performs poorly on the IGT it looks as though at the very least SCRs cannot be the only mechanism for the decision making tested in the IGT. Gerrans cites an experiment (Gutbrod et al. 2006) involving amnesiac patients with an intact capacity for SCRs who perform the IGT. While the impairments of amnesiac patients are quite different to that of VMPFC patients the capacity for mental time travel is impaired in both cases; in the amnesiacs' case due to damage to the episodic database rather than to the voluntary retrieval executive functions that are impaired in the VMPFC patients. In Gutbrod's study:

> …9 of 11 patients performed at chance and did not show differential anticipatory SCRs to advantageous and disadvantageous decks. Furthermore the magnitude of anticipatory SCRs did not correlate with behavioural performance, leading to the conclusion that "acquisition of a behavioural preference—be it for advantageous or disadvantageous choices—depends on the memory of previous reinforcements encountered in the task, a capacity requiring explicit memory." (Gutbrod et al. 2006, p. 1315). (Gerrans, 2007, p. 471)

In addition Gerrans cites a (Heims et al., 2004) study in which patients with Peripheral Autonomic Failure (who thus do not generate SCRs) participate in the IGT and are able to learn the punishment schedules. Again, while these lines of evidence do not falsify the SMH, they suggest strongly that an alternative, which does not rely so heavily on SCRs and which accounts for the importance of explicit learning in the IGT should be preferred.

The MTT is a hopeful contending hypothesis; it seems to be consistent with other lines of investigation into decision making such as the cascade model discussed in the previous section. Voluntary recall of episodic memory and prospection are key to Gerrans' model of mental time travel. In the cascade model (Koechlin et al., 2003) there were two parallel streams of processing for contextual and episodic control. The dACC energized/motivated the mid-lateral PFC which (according to the model) underpins episodic control, and the preSMA energized the posterior lateral PFC, which underpins contextual control. On this model the episodic control signals enable us to "transcend habitual stimulus-response associations" (Egner, 2009, p. 821) over a longer temporal frame than contextual control signals enable. I suggest that this difference in temporal frame is relevant here because if mental time travel is about re-inhabiting an experience or prospection over an experience then this requires a longer temporal frame than declarative reasoning; just as it takes time to have an experience in the first place, it takes time to re-inhabit it or to imagine it. Although on Koechlin's model the episodic control areas are envisaged to control behaviour – physical actions – it seems likely that at least part of the same mechanisms are used for the episodic control of *mental* actions. The control of a mental action that has an extended temporal frame is surely the foundation of the kinds of imagination and mental time travel emphasised by Gerrans. Of course for full mental time travel, as useful for planning and decision-making, access is also needed to the episodic database; circuits that underpin episodic memory, and this will include larger networks than those discussed so far, for example the hippocampus is known to be crucial for episodic memory (see for example Burgess et al., 2002).

If this is right and the control of mental actions over an episode is needed for mental time travel, and mental time travel is needed for successful outcomes in the IGT, then it follows that damage to part of the machinery for episodic control would result in a deficit in mental time travel and poor performance in the IGT. Recent literature on the neural deficits associated with people with autism shows that the dACC is implicated in autism (see Thakkar et al. 2008; Dichter et al. 2009; Minshew & Keller, 2010). People with autism are known to have deficits in imagination. In addition, autism has long been associated with

deficits in emotion, though it has often been difficult to pin down exactly what those deficits amount to. It is often suggested that they have particular problems with 'putting themselves in someone else's shoes', which clearly requires just the imaginative mental time travel that this model would suggest they would have difficulties with if there were problems with the dACC and thus lack of energizing the episodic control areas. Lind & Bowler (2010) have shown just these sorts of difficulty with episodic memory and episodic future thinking in adults with autism.

2.5 Summary of section 2

The model that I suggest here based on Gerrans (2007) and Koechlin (2003) is that, just as Gerrans says, decision-making and planning require mental time travel. Gerrans suggested that for mental time-travel two things need to be in order: (1) the episodic database; and (2) the voluntary access to the episodic database. In VMPFC patients he hypothesized that the mechanisms underlying the episodic database were faulty, as opposed to amnesiac patients whose main deficits were with the episodic database itself. I have just suggested, however, that there is a third way that mental-time travel can be impaired. Based on Koechlin's (2003) cascade model of cognitive control, for the voluntary control of episodic action the areas underpinning this must be energized. I suggest that deficits to the area that energises episodic control would also result in impairments in mental time travel, and thus impairments in decision-making that requires imagination such as is required in the IGT. The evidence for this third possibility of impairment in time travel is the deficits in dACC in people with autism and their corresponding deficits in episodic memory and episodic future thinking. While the IGT has not been tested on people with autism, or others with dACC deficits, the model I am suggesting predicts that they will perform in the same way as those with VMPFC deficits.

2.6 Summary of chapter

The somatic marker hypothesis is the hypothesis that decision-making involves affect in virtue of markers in the ventromedial prefrontal cortex acting as links between cognitive areas and areas which subserve autonomic changes or the representation of such bodily changes. I have argued (contra Colombetti) that the general version of this hypothesis (which takes somatic markers to embody preferences which enable decision-makers to choose among options: SMH-G) is not necessarily wrong as it does not merely predict

indecisiveness. Rather, it may predict a deficit in strategy switching, which presents as perseveration in the task. This interpretation fits with the research that suggests that VMPFC patients have a deficit in reversal learning and suggests that patients find themselves locked into the default exploitative behavioural strategy. In typical functioning the switch between exploitative and exploratory behavioural modes may be underpinned by the energising of areas subserving episodic (rather than contextual) control. If part of this episodic apparatus were damaged we would see an inability to make decisions which require acting on episodic rather than contextual cues, resulting in "short-sighted" behaviour of the sort that VMPFC patients are described as exhibiting (and which Colombetti targets in the special version of the hypothesis: SMH-S). We can thus see that both behavioural aspects of VMPFC patients which the somatic marker hypotheses (general and special) try to explain can be accounted for in such a model. Furthermore, this model provides the mechanisms which could underpin Gerrans' hypothesis that the VMPFC patients' behaviours are due to deficits in mental time travel (episodic past and episodic future thinking). While such deficits may well also exhibit as deficits in emotion, as past and future emotional 'memories' will not be able to play any role in contextual driven behaviours, this may just be a side-effect of deficits in episodic functioning (in the energisation of these areas) rather than being indicative of disrupted emotion links being the mechanism for the behavioural (decision-making) deficits as the somatic marker hypothesis proposes. This is not to say, however that valence and bodily affect play no role in these processes, just that they do not play the particular role of 'marking' thoughts in the way envisaged in the somatic marker hypothesis. So, while Damasio and colleagues are right that affect biases decision-making, their particular hypothesis (1) is not validated by the Iowa Gambling Task, and (2) endorses a modularisation of emotion and cognition which is now being draw in into question by recent work in neuroscience (which I will discuss in chapter 4). I will argue that affective information is involved throughout cognitive processing and that the shape that this affective substructure takes constrains the possibilities for cognition, showing that cognition is partially constituted by affect without positing somatic markers or relying on the IGT.

In the next chapter I will examine further the relations between valence, emotion, and behaviour by investigating the dissociations evident in pain pathologies and show how affective states can be grounded in internal bodily processes. In the rest of the thesis I will present an alternative account of the role of affect in cognition.

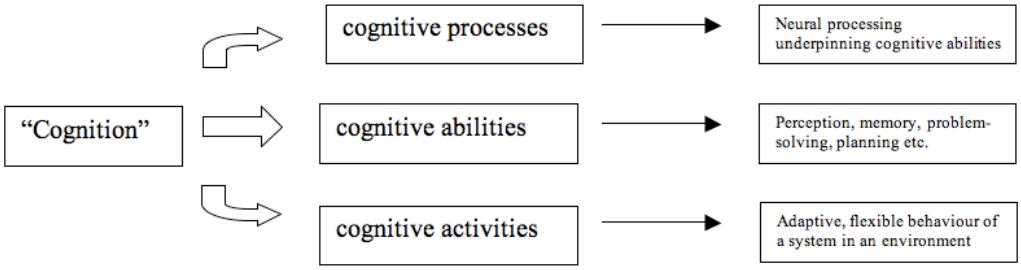
Chapter III

Grounding Affect in Interoception

Introduction

In the previous two chapters I highlighted a common conflation between 'emotion', 'affect' and the experience of affect. In both the models I discussed (embodied appraisal & somatic marker hypothesis) these affective/emotional aspects of our psychological and physiological functioning were presented as distinct (though highly inter-engaging) components from 'cognitive' functioning. In this chapter I will consider how the more basic affective phenomenon of pain fragments into components, and use this to guide us in investigating how the fundamental affective components are grounded in the afferent wing of homeostatic regulation of the body (interoception) and its corresponding behaviours. I will argue that this supports grounding a very basic kind of cognition, a 'minimal appraisal' in affective functioning. I will then argue in chapter four that interoception and the experience of affect are involved in perception and some paradigmatic cognitive processes. Before proceeding with my investigation of the role of affect in cognition however, it would benefit us to clarify the usages of these terms a little.

'Cognition' is traditionally used to refer to both cognitive processes and cognitive abilities. 'Cognitive processes' may refer to the information processing/functional/computational description level processes which are implemented by the neural processes that underlie the cognitive abilities. Or it may be used to refer directly to the neural processes that underlie the cognitive abilities themselves. Cognitive abilities are animal level abilities, such as perception, memory, planning etc. These are animal level because they are what *we* do in an environment as we are living rather than just processes that are going on in our brain, though clearly the neural processes which underpin these abilities are essential. It is the cognitive *abilities* that experimental psychologists typically measure in their tests and experiments, while cognitive neuroscientists investigate the neural correlates of (the processing underpinning) these abilities. Let us refer to cognitive abilities and cognitive processes as cognitive with a small 'c'. I will reserve "Cognition" with a capital 'C' for referring to the adaptive and flexible behaviour of a system in an environment, that is its activities rather than its abilities, which will subsume both cognitive processes and abilities.



Emotion and affect are no less complex than 'cognition'. Researchers and lay people alike use the terms differently in different contexts, and it will do us well to consider these usages before considering the relation between them and cognition. 'Affect' is typically used either as the broad category which includes emotions and moods, or as referring to the more specifically phenomenal (experiential) aspect of emotion or moods. 'Emotion' can also be used in both these ways when it is referred to in the singular. As noted in chapter one, this can cause considerable confusion.

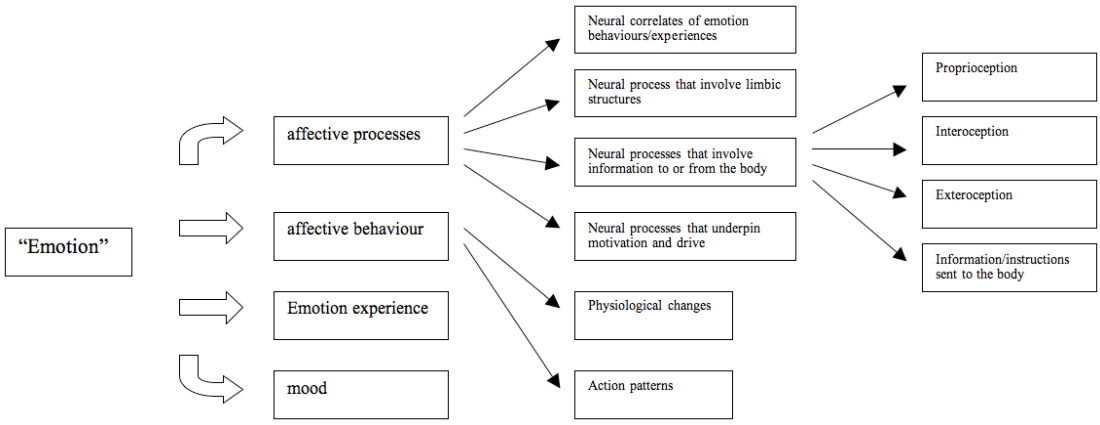
'Emotion' can also be used to refer to affective processes in the brain; neural processes that underpin emotional behaviours. It is often also used in respect to a primarily subcortical region of the brain which has come to be known as the 'Limbic System', sometimes considered to be the 'emotion brain' (see MacLean, 1973). In addition to referring to those neural processes that correlate with emotional behaviours or experiences, affective (or emotional) processing can also refer to neural processes that involve neural structures that are considered to be paradigmatically emotional, i.e., those in the limbic system, for example, the amygdala. At a slightly broader level, processing that concerns sending information to, or receiving information from, the autonomic and endocrine systems might be considered to be affective (see for example, Panksepp 2005), as might any processing that underlies motivation and drive in the system. Note that although Panksepp does not identify all affects with emotion, this is because he uses a narrower conception of emotion than the one I am using here; the conception he uses is related to action dynamics. It is clear, however, that despite his use of "emotional affects" to designate just those experiences linked to action dynamics, that in fact all of the affective phenomena which he discusses are also emotional:

> In my view, emotional feelings represent only one category of affects that brains experience. Emotional affects appear to be closely linked to certain prototypical types of action readiness (e.g., rage, fear, desire, lust, distress, nurturance,

> playfulness) that may derive their characteristic experiential feels from brain operating systems that orchestrate such instinctual responses. Other affects, constituting the pleasures and displeasures of sensations (e.g., enticing and disgusting stimuli) and bodily homeostatic and background feelings (e.g., hunger and exhilaration), reflect how life-supportive and life-detracting stimuli create neuro-phenomenological changes that help index neuro-metabolic states of well-being. (Panksepp, 2005, p. 162).

Panksepp is focusing here on affective experience, but I think that the categorisation also holds for the neural process underlying these experiences; that the bodily homeostatic processes (i.e., those to do with the internal milieu/environment) are generally identified as "emotional". As Lane and Nadel state in their book *The Cognitive Neuroscience of Emotion*, ''what distinguishes emotion from cognition may be its 'embodiment', in that the autonomic, neuroendocrine, and musculo-skeletal concomitants of emotional responses distinguish them from cognitive processes'' (Lane & Nadel, 2007, p. 7; cited in Panksepp, 2005, p.8).

At the next level up 'emotion' can be used to refer to affective behaviour which is usually a pattern of physiological changes, and/or some specific action patterns. 'Emotion' is also used to refer to the specific phenomenal experiences that often go along with these patterns of physiological changes, often combined with particular thought patterns. And finally, it can refer to moods, or someone's propensity to exhibit their physiological reactions in day-to-day life.



Having noted the ambiguities of these terms, I will now discuss an investigation into dissociations that arise as a result of pain pathologies and which I argue shows that affect is distinct from emotion. I will then explore the neurobiological basis of affect and how this grounds affective experience, arguing along the way that these provide a basis for a kind of minimal cognition.

Section 1: Pain as a case study to explore the distinction between affect and emotion

In his book *Feeling Pain and Being in Pain*, (Grahek, 2007) argues that pain is not a unitary phenomenon but has 3 components: sensory discriminative; emotional-cognitive; and behavioural. The result of this is that it turns out that 'feeling pain' and 'being in pain' are not the same. If we think of pain as a basic and homogenous experience then the distinction between feeling pain and being in pain is unintelligible but, Grahek argues, the distinction can be made intelligible by analysing the data on pathological dissociations. The dissociations which Grahek discusses are Pain asymbolia, which he refers to as "pain without painfulness"; and another – unnamed - which comes out as "painfulness without pain"; and several other conditions in which one of the three aspects of pain is missing, such as congenital analgesia, threat hypersymbolia, leprosy, surgery under curare, pre-frontal lobotomies, cingulotomies, and the influence of morphine.

1.1 'Pain without painfulness' and 'painfulness without pain'

In pain asymbolia (pain without painfulness) the sensory-discriminative aspect of pain is still present, which means that those with pain asymbolia still feel a sensation particular to pain, but it doesn't hurt (there is no painfulness; they feel pain but are not "in pain"). Grahek argues that the reason they are not "in pain", despite feeling the sensation peculiar to pain, is because the emotional-cognitive and behavioural aspects are missing. Thus, the patient attaches no meaning to the sensation; no threat or danger, there is no unpleasantness to the sensation, and the patient has no disposition to avoid the sensation.

The dissociation between pain and painfulness is a double dissociation. Just as pain asymbolia is pain without painfulness, Grahek gives evidence from (Ploner, Freund, & Schnitzler, 1999) that there can be painfulness without pain (Grahek, 2007, pp. 108-110). This is a really curious phenomenon in which the patient had selective lesions in the right primary and secondary somatosensory cortices, with the result that the patient had hypoaesthesia (a reduction in sensitivity to touch) of their left foot, leg, and face, and anesthesia (complete absence of sensitivity to touch) of their left hand and arm. Using a cutaneous laser stimulation on the arm the experimenters could not elicit pain even up to an intensity of 600 mJ, where 200 mJ was sufficient to elicit pain sensations on the right side. However from intensities upward of 350 mJ the patient described their experience as

"clearly unpleasant", despite not really being able to feel the stimulus. The authors of the study described the patient as having "pain affect without pain sensation" (Grahek, 2007, p. 108). Using Grahek's terminology we might say that despite losing the sensory-discriminative aspect of pain due to the lesions in the somatosensory cortex, the patient retained the emotional-cognitive and behavioural aspects of pain, at least in so far as the patient found the feeling highly unpleasant and wished to avoid it. In addition the patient had no real tactile sensations; he could only locate the noxious stimulus as "somewhere between fingertips and shoulder" (Ploner et al., 1999, p. 213) citied in (Grahek, 2007, pp. 108-109). To clarify the dissociations described by Grahek's review, I formulate them in tabular format below:

Pain asymbolia (pain without painfulness):

Sensory-discriminative	Emotional-cognitive	behavioural
√	X	X

Painfulness without pain:

Sensory-discriminative	Emotional-cognitive	behavioural
X	√	√

1.2 Pain without a behavioural response

There are other pain dissociations that can help clarify the subtle differences in pain and painfulness experience as a result of one (or more) of the components of the typically unitary experience of pain being missing. For example, in the 1940's surgery was undertaken while supposedly anaesthetising patients with curare. It turned out however that curare was neither an anaesthetic nor an analgesic but merely a paralytic (and for the most part a retrospective amnesiac) meaning that during the operations patients were able to feel the pain and find it highly unpleasant (even though they may not have remembered the pain afterwards) but were unable to do anything about it or show their discomfort. Technically, this example will give us ticks in all three boxes: sensory-discriminative, emotional-cognitive, and behavioural as even though the patients were unable to put a behavioural response into action there was certainly motivation or a disposition to do so had they not been paralysed. Patients who have undergone a lobotomy or cingulotomy, however, retain the sensory-discriminative component while seemingly completely losing the emotional-cognitive and behavioural component. Grahek relates case studies of patients suffering chronic pain who had

lobotomies or cingulotomies to try to lessen the pain (Grahek, 2007, pp. 126-137). Behaviourally it seemed as if these patients no longer had any pain as they ceased to complain about it and were able to move in ways that they were not able to before, when they had been disabled by their terrible pain. However on closer examination it was discovered that they did actually still have pain; when they were asked about it directly they would admit it, and, when closely observed making movements that would have been hugely painful prior to the lobotomy, physiological signs of pain, such as eyes watering, could be seen. But what these patients did lose was their aversive reaction to the on-going pain that they had experienced before (new acute pain would still be processed as normal). What this means is that the sensation was still present but it was no longer the object of fear or anxiety and ceased to signify threat or danger. That the pain no longer had these emotional-cognitive components should not, I suggest, be taken to say that the pain is no longer felt as unpleasant; the valence does not seem to be absent. A more plausible description of the case seems to be that the unpleasantness doesn't *mean* anything, and does not drive the patient to avoid the sensation, either because (1) the valence is the same as before but disconnected from the emotional-cognitive component, or possibly (2) because the valence itself is in these cases too 'nebulous' and thus not able to drive any particular action. Later in the section we will consider some cases that appear to count in favour of (1).

We saw above, with the case of pain asymbolia, that we should not think of valence as entwined in the sensory-discriminative component of pain. If the above description of the dissociations in the lobotomy patients is right then this suggests that an extra dissociation[12] should be added in our components: *valence* should also not be considered entwined in the emotional-cognitive component. So I suggest that the situation with lobotomised patients in on-going pain is as follows:

Lobotomised patients:

Valence	Sensory-discriminative	Emotional-cognitive	behavioural
√	√	X	X

[12] Grahek does discuss valence as separate from the other components but does not seem to categorise it as a component in itself. I think that it makes our analysis clearer to add it to the list of dissociations of components.

1.3 Feeling the object of pain without feeling pain

Congenital analgesia might be easily mistaken for pain asymbolia but Grahek argues that there is a subtle difference. Where those with pain asymbolia feel pain but not the painfulness of the pain sensation, those with congenital analgesia feel tactile sensations just like typical people and pain asymbolics, but there is no pain quality to these sensations. So the claim is that a typical person, upon being stabbed in the shin with a needle will feel the prick of the needle entering the shin as follows: tactile sensation + pain sensation + painfulness (it will hurt). A pain asymbolic will feel the pricking of the needle entering his shin, and the pain sensation, but no painfulness alongside these sensations. The congenital analgesic, on the other hand, will feel *only* the sensation of the pricking of the needle entering his skin, with no pain sensation and no painfulness; he neither *feels* pain nor is *in* pain.

Congenital Analgesia:

Tactile sensation	Sensory-discriminative	Emotional-cognitive	behavioural
√	X	X	X

1.4 Feeling pain without an object of pain

In threat hypersymbolia painfulness can be elicited in a patient by purely visual stimuli; in the case of the patient who (Grahek, 2007, pp. 16-17) describes the mere act of the patient seeing someone approach his arm would set off a response whereby he grimaced, tried to avoid the stimulus, and most intriguingly of all, experienced burning pain. This is a case of having all three components of pain which Grahek focuses on (sensory-discriminative; emotional-cognitive; and behavioural) but these being dissociated from tactile sensations which usually accompany pain experience, so much so that they are often identified with the sensory component of pain (Grahek, 2007, pp. 104-105). Of course, in this case the pain is still described as "burning" and this might be considered to be a tactile sensation. Even so, the dissociation from actual tactile stimulation is intriguing.

Threat hypersymbolia:

Tactile sensation	Sensory-discriminative	Emotional-cognitive	behavioural

X(?)	√	√	√

1.5 Separating valence from the emotional-cognitive component

While Grahek focuses on pain having three components; sensory-discriminative; emotional-cognitive; and behavioural, the discussion above suggests that in order to really understand the subtleties of the components of typical pain we should add in tactile sensation and unpleasant valence, both of which Grahek discusses in his case studies but does not include in the components of pain. It might be thought that valence falls under the auspices of "emotional-cognitive" but the example of the lobotomy patients suggests that this is not the case and that we should use the "emotional-cognitive" label only to refer to the meaning that is given to a sensation. The emotional-cognitive aspect of pain is whatever makes that sensation the object of fear or anxiety. So let us look at the components of the dissociation phenomena with tactile sensation and unpleasant valence added in:

Dissociation phenomenon	Unpleasant valence (painfulness)	Tactile sensation	Sensory-discriminative (pain)	Emotional-cognitive	Behavioural
Pain asymbolia	X	√	√	X	X
Painfulness without pain	√	X	X	√	√
Lobotomised patients	√	√	√	X	X
Congenital analgesia	X	√	X	X	X
Threat hypersymbolia	√	X(?)	√	√	√

What emerges from viewing the phenomena in this format is firstly that while pain asymbolia and painfulness without pain seem to support a double dissociation between feeling pain and being in pain, and congenital analgesia and threat hypersymbolia likewise seem to support a double dissociation between feeling pain and feeling the object of pain. There is none however, for the lobotomised patients. What would such a case look like? If someone felt no tactile sensation, no pain sensation, and no unpleasant valence and yet still

felt threatened and was disposed to behave as one normally does in painful or threatening situations? If there were to be such a case it would surely fit under some psychiatric pathology, but it may even be that such as case is unintelligible. This apparent unintelligibility gives us reason to believe that the emotional-cognitive and behavioural components must have an object or reference. In the other phenomena this is provided by (1) the unpleasant valence (in painfulness without pain) or (2) either unpleasant valence or the sensation of pain (in threat hypersymbolia). In this respect even though one can have unpleasant valence without the emotional-cognitive and behavioural components it may be that valence is fundamentally intertwined with the emotional-cognitive component in that the emotional-cognitive component cannot be present without valence being present; one cannot experience threat without unpleasantness. The claim is thus, that even though valence may be present without the emotional-cognitive and behavioural components, it is nevertheless core to the emotional-cognitive component where that is present. I will discuss this in further detail in section three where I propose that valence can function as a minimal appraisal mechanism.

What also emerges from viewing these dissociations in this format is that the emotional-cognitive and behavioural components seem to always be present or absent together; there is no case here discussed in which the emotional-cognitive component is present and the behavioural component is absent, or vice versa. That the behavioural component requires the emotional-cognitive component is understandable, after all why would someone have the motivation to avoid a threat if they had no experience of something as threatening (or at least processing it as threatening in some way, even if this is at a subpersonal level)? But need the converse hold true? Is it possible for someone to have the experience or meaning of threat or danger without having the motivation to avoid this threat, or is the motivation conceptually integrated into the meaning of threat?

Grahek's work on pain thus raises three questions that are particularly interesting in regard to affect. The first two questions – as described above - arise directly out of viewing the dissociations in the format above:

(1) How is valence related to the emotional-cognitive component?

(2) How are the emotional-cognitive and behavioural components related?

The third question arises from the first two and underpins these, so I will address this in the next section:

(3) What is the role of unpleasant valence in typical pain situations?

1.6 The role of valence in typical pain

In cases such as congenital analgesia it is clear that the lack of pain and painfulness results in severe disability. Not only are there the obvious dangers of not being aware when the external body is in danger and so the likelihood of burning or cutting oneself severely without one noticing is radically increased but even behaviours that we consider benign, such as standing or lying in one position too long, can severely traumatise the body and result in physical disability. In these cases the patient has no way of knowing when something is harming them unless they notice it visually, which may well be far too late to stop severe damage. One might conclude that the lack of availability of information about the damage is the key factor in these situations and thus if the patients were to be given some way of knowing that something was causing them physical harm the emotional-cognitive and behavioural aspects of pain would kick in despite there being no valence or sensory-discriminative components. Grahek shows that this is not the case however, by examining the outcome of a project for designing pain substitution systems "A Practical Substitute for Pain" (Brand & Yancey, 1997), designed for leprosy patients who were no longer able to feel pain in their extremities and consequently in danger of damaging their limbs as they had no source of information about occurring damage.

Leprosy causes the nociceptive fibres to die which means that leprosy patients have neither pain sensation nor painfulness (unpleasant valence). However because the bacilli (mycobacterium leprae) prefer cooler areas they accumulate around the far limbs such as hands and feet and warmer areas such as armpits tend to be less affected. Grahek explains that the project failed because – seemingly – they could not seem to get the emotional-cognitive component to activate. The first attempt at a pain substitution system used electronic sensors integrated into gloves and socks for the leprosy patients to wear and a hearing aid type piece which emitted a low level hum when the sensors received safe pressures and a more intrusive buzz when they detected pressures that could be slight danger and a piercing sound when they detected actual damage. However the experimenters found that the patients would just override the system when they wanted by turning the effectors off. This seems somehow absurd. Why would someone override a system designed to warn them of danger? It appears that patients would rationalise their actions to themselves by attributing to the gloves that they were always sending out false signals. Of particular

interest to us here is that when the researchers then decided to ditch the "substitute" part of the project and wire the sensors directly up to the parts of the body unaffected by leprosy, such as the armpit, such that a dangerous stimulus would actually cause (real) pain in the area the sensors were wired up to, the emotional-cognitive and behavioural elements seemed to still be absent (at least in relation to what the pain signals were supposed to be 'about'). The patients would feel the pain but when inconvenient they would just disconnect the sensors from the pain effectors and reconnect them once the task was done.

It should be noted here that in a typical organic pain system even if our pain sensors did send out false signals (in that they do not indicate any actual threat to the body), as indeed they do in some of the pathologies that Grahek discusses such as threat hypersymbolia, patients are not able to override them. Rather the emotional-cognitive and behavioural components are very clearly present, resulting in the patient being motivated to avoid the "threat". Overridability thus seems to be playing a critical role in connecting valence to the emotional-cognitive and behavioural components which seem to be constitutive of our natural concept of pain.

Of course there still is a behavioural response here; the patients are motivated to disconnect from the system which is causing the unpleasant valence. What distinguishes this behavioural response to that which is typical in pain situations is that the behavioural response is not tied to the threat but rather to the valence. Despite real pain being used, the emotional-cognitive component did not activate and guide the patients by representing threat. If threat *had* been being represented by the pain (even though it was shifted in location) then even though the patients would have the opportunity to disconnect the wires, one would assume that they would not because the awareness of the danger would guide them away from doing so (in all but emergency situations). We can take from this that to kick the emotional-cognitive component of pain into action requires more than just being informed of threat – that is, mere availability of information about the threat is not the key factor in engaging the emotional-cognitive and behavioural components in pain. Rather, the information must be in a form which has intrinsic imperative. It seems plausible that this is related to the overridability of information rather than an issue about where the pain is. For example, someone might argue that a pain in the armpit only informs the individual of danger to the armpit and so is not capable of informing the individual of danger to the hands. However, this should not matter; if there was at least one place in the body informing the subject of danger then irrespective of what part of the body has the sensations, at least some

danger should be being coded for, and thus emotional-cognitive and behavioural components would be present. This was evident to some extent in regard to behaviour at least, in the following way: Although the motivation to avoid the threat was not present due to there not really being a threat properly represented (in virtue of the fact that the emotional-cognitive component was absent) there was clearly some behavioural component that was activated. We know this because the patients were certainly motivated to avoid the pain in the armpit – and did so by disconnecting the pain effector. So, we might say that there is some sort of behavioural component present but here it is in virtue of the unpleasant valence, and a valence motivated behaviour appears to be quite different to an emotional-cognitive behaviour where the behaviour is motivated by an appraised threat rather than discomfort (even though when all is working well these aspects are deeply entwined). So this gives us a possible dissociation between the emotional-cognitive component and the behavioural component that did not seem available given the dissociations discussed earlier. Again returning to the importance of overridability of the valence dimension of pain, I suggest that this discrepancy (between pain directed behaviour and threat directed behaviour) might be overcome if the set up could be non-overridable, plausibly developing by means of many associations over time, in a direct association between what is happening at the hand and the pain in the armpit so that gradually the behavioural component would be directed first and foremost to the hand rather than the armpit.

There are thus two types of pain behaviour: typical pain behaviour which is motivated by the emotional-cognitive component, such that its intentional object is 'threat' or 'danger'; and dissociated pain behaviour, which is motivated by the unpleasant valence, and whose intentional object would seem to be the unpleasantness of the valence. This means that – contrary to what one might initially presume – at least in some pathological cases negative valence does not represent threat. Of course, as (Schenk, 2010) argues in respect to the dual visual stream literature, we must be careful not to draw too strong a conclusion about typical functioning from pathological dissociations; just because dissociations occur in pathologies does not mean that this reflects typical functioning which may well be integrated. It may be the case then that when connected up 'correctly' valence does represent threat in as much as the object of behaviour is then to eliminate the threat. This would seem to fit with the functioning of very simple animals (perhaps mice or even simpler animals) which, though we might assume that they do not have sufficiently complex cognitive capacities to appraise a situation as threatening in the way that we typically think of appraisal (in terms of judgments or deliberation, either conscious or subconscious), still react to stressful and

painful stimuli by trying to avoid it. It is plausible that in such cases it is valence which is representing the danger in the environment, which in more complex systems like ourselves is represented either by the emotional-cognitive component (as threat) or by this and the valence component combined. In the leprosy cases described above however, not only is the threat absent, which is to be expected if the emotional-cognitive component is absent, but so is any representation of danger at all. The valence seems to be tied neither to the threateningness – as would be expected in typical functioning - nor to the danger in the environment itself – as would be expected from the mouse model above. The behaviour in this situation makes no ecological sense; the object of the behaviour is the elimination of the valence itself rather than eliminating either threat (the representation of danger through the emotional-cognitive component), or danger itself. One possibility could be that in simple systems such as mice, the very simplicity that allows the negative valence to be directly representing the danger (without the intermediate step of threat) is what allows the danger to be the object of behaviour (we might think of this as a very minimal type of appraisal – we will see a possible mechanism for this in section three, where I investigate the intrinsic motor aspect of valence). Perhaps it is a result of the complexity of our cognitive system, a system that provides much more flexibility in the form of allowing us to represent danger through threat and thus have this as the object of behaviour rather than danger itself, that the elimination of this functionality means that we do not necessarily revert (phylogenetically speaking) to representing danger through valence, but fail to represent the danger altogether while still retaining the negative valence. The result of this is that neither threat nor danger are available to be the object of behaviour and thus valence stands as the only available object. So ordinarily we respond to painfulness *because* it is unpleasant but - because the threat and the valence are linked up correctly - *through* eliminating the threat. But, in situations where the valence does not represent the threat we respond only by eliminating/avoiding the source of pain. While behaviourally motivating in this regard the pain does not seem to be informing the organism about danger in an ecologically useful way. While this may plausibly be rectified by building up associations between pain and danger it would seem from the analysis above that this would require the non-overridability of the pain signals.

1.7 Summary of section 1

The analysis of the components of pain in this section allowed us to draw a distinction between three components of pain that are of particular interest to this investigation of affect

and the body; valence, emotional-cognitive, and behavioural. While these three components are clearly related I have tried to show that they are distinct, and that while valence is a core aspect of the emotional-cognitive component, it is not only a sub-component of this but a component in its own right as it is plausibly dissociable from the emotional-cognitive component. Let us review the relations that I have brought out between these components:

(1) *How is valence related to the emotional-cognitive component?*

Valence in this context refers to the unpleasantness of pain. I have shown that this unpleasantness is not merely a subcomponent of the emotional-cognitive component, which underpins the threateningess of the situation, but that they are dissociable. The example of the failure of the pain substitution project shows that one can have an unpleasant valence without the emotional-cognitive component being active (this also seemed to be the case in lobotomized patients) but we have no examples of the emotional-cognitive component being present without unpleasant valence so for now it seems sensible to suspect that the emotional-cognitive component is dependent upon valence. It looks like typically valence and the emotional-cognitive component are intertwined such that while the emotional-cognitive component (in the form of threat) represents the danger in the environment valence represents the threat. It may be the case that valence can independently represent danger and thus connect to the behavioural component (as would be ecologically useful in simple animals), such that we should consider that valence is driving the behaviour, and thus that valence is – in a minimal sense – an appraisal and in that respect can act in the same capacity as the emotional-cognitive component. However, it is not the case that - where the system is set up such that the emotional-cognitive component underpins appraisal - in the absence of the emotional-cognitive component then valence will take over this job and underpin appraisal as it may in simpler systems. It thus seems to be the case that in the human cognitive system valence and the emotional-cognitive component are importantly intertwined and function together to appraise and serve to represent the object of behaviour.

(2) *How are the emotional-cognitive and behavioural components related?*

The example of the pain substitution project shows that there can be a behavioural component in the absence of an emotional-cognitive component. However, in such a case the behavioural component seems dependent upon the valence component, and importantly, the behavioural component in this case seems to take quite a different shape (i.e., motivating avoidance of the pain signals rather than any danger in the environment). We might surmise from this that the behavioural component typical of pain in humans (and which is

ecologically useful for us) is dependent on the emotional-cognitive component. I have suggested that in the human cognitive system the emotional-cognitive component adds flexible adaptivity to the cognitive set-up such that the behavioural component takes this (for example, 'threat') as its object rather than either the valence or the danger present in the actual environment. What comes out of this analysis that is of particular interest here, is that while the emotional-cognitive component and the behavioural component are intertwined in good functioning, valence seems to also be intertwined with behaviour. So, an organism need not be capable of appraising something as threatening in order to respond behaviourally. Unpleasantness is thus sufficient for a behavioural response which allows us to think in terms of a very minimal kind of appraisal, which I will ground later in the chapter in the motoric component of behavioural homeostasis.

(3) What is the role of unpleasant valence in typical pain situations.

Unpleasant valence seems to provide more than a signal of danger; it seems to be intrinsically tied to the behavioural component in a way that other signals of danger are not (and thus other signals of danger such as buzzing or flashing lights in the cases of the pain substitution systems can be overridden). However, when it is the only object of the behavioural modification, as is the case when the emotional-cognitive component is missing, then the behavioural motivations are only to avoid the unpleasant valence and not the threat (as no threat has been represented). Thus we can see that unpleasant valence is necessary for an ecologically useful pain system but, in complex systems like ourselves, it also needs the emotional-cognitive component for the behavioural motivations to be reliably directed to the correct object.

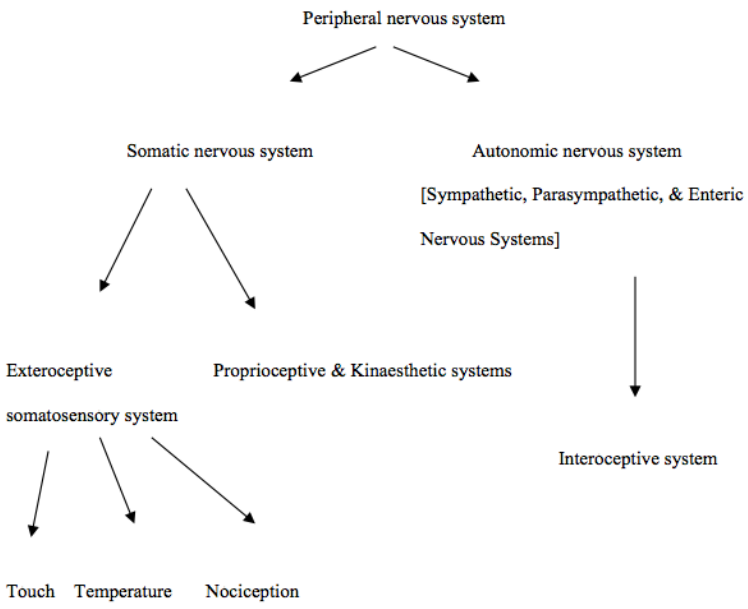
Affective phenomena such as emotions and pain, are thus comprised of various components, one of which is valence, and this, as we have seen, can stand alone in motivating a behavioural response. I now turn to investigating the neurobiology of affect, to see how valence is grounded in internal bodily changes, and examine the relation between valence and behaviour.

Section 2: The neurobiology of affect

In this section I show how affect is grounded in the afferent wing of homeostatic regulation and how it can support a basic kind of cognition. I will explore ways of thinking about affective phenomena as driving 'homeostatic behaviour' and provide the basis for grounding the experiential aspects of affect in the mechanisms underlying this which will be the subject matter of section 3.

2.1 Interoception, pain and touch

The sensory system has traditionally, since (Sherrington, 1906), been primarily categorised into two systems; the exteroceptive and the interoceptive. Exteroception refers to the perception of external sensations; sensations whose direct causes originate outwith the body. In addition to the 5 classic sense modalities of vision, smell, taste, touch, and hearing, there are the "somatic senses" which include: pain, temperature, itch, proprioception, and vestibular balance (Kandel, Schwartz, & Jessell, 2000). These are all typically considered exteroceptive, however sometimes proprioception is categorised independently and as including kinaesthesia which is the sense of the movement of body parts rather than just the sense of the position of the limbs.



The afferent sensory systems and their relation to the nervous systems

Interoception is distinct from the other senses which take the body as their object (proprioception and kinaesthesia). It is typically used to refer to the afferent sensory information from the autonomic nervous system such as the heart muscle, other smooth muscle (as opposed to skeletal muscle which is included in the somatic nervous system), and the exocrine glands (i.e., sweat glands, saliva glands, stomach, liver & pancreas). The interoceptive system is like the exteroceptive system in terms of the different sensory receptors present. In the skin for example, there are 4 types of cutaneous mechanoreceptor, and whether a stimulus is felt as touching an object or as a vibration/tingle depends on whether all four receptors are activated or whether only a subset are. Itches are transduced by means of chemoreceptors, which means that they are stimulated by chemicals in the area of the receptor. Temperature is transduced by thermoreceptors. And pain receptors (nociceptors) come in all three types of receptor: mechano-; chemo-; and thermoreceptors.

As in the exteroceptive system, the interoceptive system encompasses mechanoreceptors so some interoceptive states can localise the source and type of stimulus (pricking, burning, stretching etc.), and chemoreceptors, which do not really give much information about the source or type of stimulus as they work on the principles of diffuseness and threshold. That is, the molecules that chemoreceptors are sensitive to are diffuse rather than sharply localised as a mechanical stimulus would be, and so long as the density of molecular stimuli were above the threshold at which that receptor fires they do not distinguish between locales. Smell and taste are both implemented by types of chemoreceptor and so the presence of this type of sensory receptor is not peculiar to interoception. However, in most cases the exteroceptive somatosensory system has the distinction of having well discriminated feelings (Craig, 2003a); temperature, itchiness and pain are all not only very clearly localised but clearly point to a particular sensation. The interoceptive sensations of vasomotor activity, hunger, thirst, and other internal sensations are less distinctive; not only are they not well localised, but it can take some training to be able to accurately identify them as distinct from one another.

2.2 Interoception grounds minimal consciousness

Maps of this interoceptive information in the brain stem can be shown to support a very basic level of feeling state (see Damasio, 2010)[13]. While we ordinarily think of feelings and other experiences as arising from maps in the cortex rather than sub-cortical structures,

[13] Note that this research is distinct from the somatic marker hypothesis discussed in chapter two.

nuclei in the medulla, pons, and midbrain also contain maps. Rather than pertaining to the exteroceptive senses, these maps are predominantly representations of the interoceptive senses; mapping the state of the internal body in order to regulate it and maintain homeostasis. The medulla is the area of brainstem directly above the spinal cord, and the nucleus tractus solitarius (NTS; also known as the solitary tract) runs along the length of this and up into the lower pons. The NTS deals with afferent information from the body. In particular "… visceral sensory information from different afferent nerves interact in the nucleus of the solitary tract to produce a single visceral sensory map of the body" (Kandel et al., 2000, p. 883). Likewise, the parabrachial nucleus in the pons (directly above the medulla and below the midbrain) also receives afferent information about the visceral body:

> The nucleus tractus solitarius and the parabrachial nucleus receive a full complement of signals describing the state of the internal milieu in the entire body. Nothing escapes them. There are signals from the spinal cord and trigeminal nucleus, and even signals from "naked" brain regions such as the nearby postrema, that are devoid of the protective blood-brain barrier and whose neurons respond directly to molecules traveling in the bloodstream. The signals compose a comprehensive picture of the internal milieu and viscera, and that picture happens to be the prime component of our feeling states. (Damasio, 2010, p. 78).

The nucleus of the solitary tract and the parabrachial nucleus are richly connected to each other and the visceral sensory maps are key to regulating homeostatic (life regulation) processes, to keeping essential variables in the organism within the bounds of viability (i.e. the range of possible perturbations with which the organism can survive or thrive and not succumb to entropy[14]). They are also the first providers of whole body maps to the central nervous system. Although these areas (in coordination with the reticular activating system) are involved in awake/sleep activation on their own they are not sufficient for even a very basic kind of conscious awareness. Damasio argues however that their connections to the periaqueductal gray and superior colliculus do give rise to a basic feeling state. The periaqueductal gray (PAG) appears to be the basis of recognisably affective behaviour, and is:

> […] the originator of a large range of emotional responses related to defence, aggression, and coping with pain. Laughter and crying, expressions of disgust or fear, as well as the responses of freezing or running in situations of fear are all triggered from the PAG. (Damasio, 2010, p. 80)

[14] I take this concept of viability boundaries from Di Paolo's (2005) work on adaptivity.

But for these affective states to become conscious, Damasio argues that a connection with the superior colliculus is necessary. The superior colliculus (SC) is a primarily visual structure, receiving information both directly from the retina and the visual cortex ("Medical Neurosciences 731: Superior Colliculus," 2006). However, in addition to these maps of the visual world, the SC also contains "topographical maps of auditory and somatic information, the latter hailing from the spinal cord as well as the hypothalamus" (Damasio, 2010, p. 84). Of particular interest is that these maps may all be integrated:

> The three varieties of maps—visual, auditory, and somatic—are in a spatial register. This means that they are stacked in such a precise way that the information available in one map for, say, vision, corresponds to the information on another map that is related to hearing or body state. There is no other place in the brain where information available from vision, hearing, and multiple aspects of body states is so literally superposed, offering the prospect of efficient integration. The integration is made all the more significant by the fact that its results can gain access to the motor system (via the nearby structures in the periaqueductal gray as well as the cerebral cortex). (Damasio, 2010, p. 84)

Damasio identifies minds with such maps (representations), which is not necessarily a controversial position at least amongst neuroscientists. What *is* controversial about his position is that he believes that it is not just the maps in the cortex which are "mind-competent" but also the maps occurring in the nucleus tractus solitarius, parabrachial nucleus, periaqueductal grey, and the superior colliculus. He argues that these areas in the brain stem are "essential to mind-making" (Damasio, 2010, p. 73)[15]:

> [The nucleus tractus solitarius and parabrachial nucleus]… do not produce mere virtual maps of the body; they produce *felt* body states. And if pain and pleasure *feel* like something, these are the structures we first have to thank, along with the motor structures with which they incessantly loop back to the body, namely those of the periaqueductal gray nuclei. (Damasio, 2010, p. 77).

For Damasio, however, minds can be unconscious; the maps themselves are not sufficient for conscious primordial feelings. For conscious primordial feelings Damasio appeals to the connections with the superior colliculus. The SC not only provides us exteroceptive and interoceptive maps, as explained above, and possibly their integration, but Damasio suggests that the "powerful integration of signals" within the SC and its rich connections to brain regions for guiding movement result in images of the "in-register" maps of the superior colliculus, i.e., maps of maps – second-order maps (Damasio, 2010, p. 85). He does not think

[15] Areas which Damasio does not consider essential to mind-making (or mind-competent) are the spinal cord (p.73); cerebellum; and hippocampus (p. 74).

that these images of maps are as rich as those in the cerebral cortex, but for Damasio second-order maps are the foundation of consciousness (see Damasio, 1999). In addition, the superior colliculus produces electrical oscillations in the gamma range:

> …a phenomenon that has been linked to synchronic activation of neurons and that has been proposed by the neurophysiologist Wolf Singer to be a correlate of coherent perception, possibly even of consciousness. To date, the superior colliculus is the only brain region outside the cerebral cortex known to exhibit gamma-range oscillations. (Damasio, 2010, p. 86).

2.3 Cognition without a cortex

Damasio suggests that the superior colliculus (along with the lower brain stem structures) can, even without the cortex, ground "a very vague mind, gathering sketchy information about the world" but which may still be useful - as happens in blindsight[16]. And, that when the cortices are missing from birth as is the case in hydranencephalic children it may play an even larger contribution to mind (Damasio, 2010, p. 85). Hydranencephaly is an intriguing condition which shows us how much humans are capable of without the support of the cortex. In hydranencephalis the cortex is absent (at the very most fragments remain), replaced instead with cerebrospinal fluid. It is a distinct condition from *hydrocephaly* in which the cortex is present but has been severely compressed by the expansion of the cerebrospinal fluid in the ventricles such that in some cases all that is left is a thin layer of cortex lining the skull (in such cases the presence of the cortex can be missed and the patient diagnosed with hydranencephaly, but hydrocephalics can have quite high level (and sometimes normal) functioning as the cortex is present as normal - it is just greatly compressed). In hydranencephaly, in addition to the brain stem structures the cerebellum and thalamus may be present, as may a functioning hypothalamic-pituitary-adrenal (HPA) axis although in some cases the HPA functioning is abnormal (McAbee, Chan, & Erde, 2000).

Despite the (almost complete) absence of cortex in hydranencephalics these children can survive until late into their teens (McAbee et al., 2000) and given the right (loving and stimulating) environment do not remain permanently vegetative (as they are typically diagnosed) but can be responsive to particular individuals, music, and even visual stimuli such as favourite toys despite the total lack of visual cortex (Shewmon, Holmes, & Byrne,

[16] Blindsight is a condition wherein an individual has no experiential aspect of vision and yet retains some vision functionality, showing for example an ability to avoid obstacles at an above chance level (Goodale & Milner, 1992; Milner & Goodale, 1995).

1999). Most intriguing is that one of the subjects in (Shewmon et al., 1999)'s study could even 'scoot' around the house visually avoiding collisions:

> At age 2 years the neuroophthalmologist documented that subject 1 was fairly mobile when supine, pushing himself around in a circle with his legs. According to his adoptive mother, he could tell whether the sliding glass door to the sun porch was open and, if so, scoot through to enjoy the warmth and sunshine. Author PAB witnessed him scoot around the house, visually avoiding collision with walls and furniture. (Shewmon et al., 1999, p. 365) [17].

It seems clear that these children have a basic consciousness even if not full reflective consciousness (though it should be noted that subject 1 was fascinated by his own reflection in the mirror (Shewmon et al., 1999, p. 366), which suggests that he may pass the mirror test of self-consciousness). As the authors note, if the subjects in this paper were animals exhibiting the same level of adaptive interaction with the environment, the attribution of 'experience', in particular of pain and suffering, would be uncontroversial. Of course, that consciousness and the cognitive abilities evident in these children may emerge from the brainstem and diencephalon does not mean that these properties are typically mediated by these structures. Indeed the authors note, "[t]hat subcortical mediation of consciousness has been described so far only in congenital brain malformations suggests that developmental plasticity may play a role" (Shewmon et al., 1999, p. 371).

The kind of plasticity which this implicates is unusual. 'Horizontal' plasticity is when cortical areas that do not normally subserve a particular cortical function change such that they are able to subserve that function, or conversely when subcortical areas that do not normally subserve a particular subcortical function change so that they can subserve that function. 'Vertical' plasticity, on the other hand, is "subcortical plasticity for supposedly cortical functions" and is likely to be less robust than horizontal plasticity because "the potential for compensatory reorganization ought to be largely related to the degree of microstructural similarity between sites at issue" (Shewmon et al., 1999, p. 371). Despite being less robust, vertical plasticity - given a lack of occipital cortex from early in gestation - was clearly sufficient for the subcortical nuclei to arrange optimally for functional vision

[17] "Hydranencephaly was confirmed by CT, which showed absence of cerebral tissue rostral to the thalamus, except for small mesial temporal-lobe remnants. A thin crescent of tissue extended from the left middle fossa along the posterolateral aspect of a large midline cyst with fluid of lower density than the main supratentorial fluid[…]. EEGs showed no electrocerebral activity over the entire head except for some 50 to 60 mV theta plus low-amplitude beta in the left parietal region, corresponding to the tissue on CT scan; some tracings also revealed epileptiform discharges in the same area." (Shewmon et al., 1999, p. 364).

(*ibid*).[18] As the authors note, there are phylogenetic precedents for this kind of subcortical mediation of behaviours and perceptual functions which are traditionally considered to be 'cortical'. Firstly, falcons, owls, toads, and grass frogs are all capable of binocular depth perception despite having little or no visual cortex. And, secondly, habituation, learning, and discriminative conditioning have all been observed in decorticate animals (Shewmon et al., 1999, p. 372).

The example of hydranencephalic children shows that (1) we should not underestimate the functions of the nuclei in the brain stem; these nuclei which are primarily responsible for subserving life-regulation processes may not merely be subserving 'awakeness' but also some sort of awareness as well. We also learn that (2) given the right conditions vertical plasticity can occur such that some 'cognitive' functions (including perception) can be – if a little roughly – subserved by subcortical areas. And, (3) the brain stem and diencephalon (thalamus and hypothalamus) are sufficient for exteroception and interoception. The abilities of these children may seem very basic to us, but they include the regulation of the body itself and (at a very simple level in the child who could 'scoot' into the warmth of the sun) even intelligent interaction with an environment. Even the children who did not have the capacity to move around were able to clearly react to affectively negative stimuli such as pain or uncomfortable situations such as interacting with strangers or listening to music they didn't like.

That the cortex is not required for a basic level of consciousness and cognition should not be altogether surprising. It has long been known that decorticated animals retain significant functioning. Panksepp has further argued that this subcortically enabled consciousness (which he calls "affective consciousness[19]") can actually result in improved behaviour in some learning tasks. While normal rats get 'befuddled' in these tasks, due to being overwhelmed, "[d]ecorticate animals being generally disinhibited, shuttle readily, probably with no awareness of the benefits their tendencies for over-activity are producing for them" (Panksepp et al., 2007). So it seems that the cortex, whilst enabling greater flexibility and

[18] Note that 'functional vision' does not mean that visual experience would be identical to ours; it is doubtful that the quality of detail would be very high, but it is impressive that any of the functions of vision could be implemented in such a structurally different region from the visual cortex.

[19] Panksepp distinguishes between what he thinks of as an 'affective consciousness' and a 'cognitive consciousness', where affective consciousness consists of raw emotional feelings, or "primary-process affects" which "can exist without any cognitive awareness of those feelings" (Panksepp, Fuchs, Garcia, & Lesiak, 2007). We should not be tempted by his use of the terms 'affective' and 'cognitive' here, to think that an affective consciousness is just the ability to passively experience sensations, while 'cognitive' consciousness is the important aspect for behaviour, interaction, or typically termed 'cognitive' processes.

adaptivity, and perhaps being necessary for reflective awareness and meta-cognition, is not necessary for basic cognition and affect. The decorticate rats, like the hydranencephalic children discussed above are not purely reflexive creatures or living robots. They are less flexible and adaptive than healthy conspecifics but in addition to the basic cognitive capacities that they exhibit, they also exhibit a fundamentally affective existence.

2.4 Interoception and 'primordial emotions'

That subcortical, affective "consciousness" can suffice for basic cognitive capacities fits well with the hypothesis of (Denton, McKinley, Farrell, & Egan, 2009) that "primordial emotions are the subjective element of the instincts which are the genetically programmed behavior patterns which contrive homeostasis" (Denton et al., 2009, p. 500). Denton et al. propose a three-layered pyramidal model of emotions/consciousness where primordial emotions, grounded in interoceptors, sit at the bottom; distance receptor evoked emotions, grounded in exteroceptors, in the middle; and aesthetic emotions at the top. So, while primordial (interoceptive) emotions are feelings and behavioural patterns which are concerned with the integrity of the organism (thirst, hunger for air, hunger for food, pain and hunger for specific minerals) distance (exteroceptive) emotions are predominantly concerned with what they call "situational perception" and are the result of an interaction (rage, fear, hate, envy, happiness, playfulness, affection, anxiety, depression, disgust) (Denton et al., 2009, p. 501). The distinction between primordial emotions, and distance receptor evoked emotions thus seems to be fundamentally based in their being grounded in interoception and exteroception respectively.

In both interoceptive and exteroceptive emotions, Denton et al. take a considerable amount of the mechanisms to be hardwired, but they do not assume from this that the emotions themselves are unconscious. On the contrary they think that even primordial emotions "emerged as consciousness during evolution because they are apt for the survival of the organism" (Denton et al., 2009, p. 501):

> …the primordial emotions, the subjective component of the instinct, is the result of very powerful selection pressure, and has high survival value in signalling that the existence of the organism is threatened. As a subjective amalgam of the imperious sensations and the compelling intention it is highly likely that a powerful selection pressure would have favoured its phylogenetic emergence. This pivot of its biological relevance rests irrevocably on the fact that it is conscious, the imperious sensation being causative of compelling and apt intention to ameliorate or resolve the life threatening situation. (Denton et al., 2009, p. 505)

While the previous work discussed was focussed primarily on consciousness arising out of brainstem and diencephalic structures, Denton et al's 'primordial emotions' integrate these with limbic structures such as the anterior cingulate cortex (ACC) and Insula. For the purposes of my argument this disparity is not a problem; I am arguing that affect is grounded in interoception and structures such as the ACC and insula are involved in interoception at certain levels of processing. The examples of hydranencephalic children show however that those structures do not seem to be necessary for a basic kind of interoception that allows the body to regulate itself, and importantly also for some (albeit limited) directed behaviour and experience. What I want to draw from all of the research discussed above is the link between these basic capacities and biological value and relevance.

The above sets up the main claim that I want to make about the role of interoception in this chapter. Interoception is key to this biological relevance; it is the mechanism by which the organism can regulate itself and thus keep itself within the bounds of viability. Maintenance of homeostasis of the body in typical circumstances, however, clearly requires more than just internal changes. Unless an organism is being supplied with food and water and safety, keeping itself within the bounds of viability requires directed action which in turn requires some sort of exteroception. I will flesh this out in more detail in the coming sections. What was so interesting about the hydranencephalic cases was that the structures that have been considered to only subserve interoceptive and regulative functions can also subserve exteroceptive functions and directed action. This suggests a much tighter integration between interoception and exteroception than has traditionally been considered to be the case.

2.5 Touch, temperature, and pain as "homeostatic emotions"

We can see the above ideas clarified in respect to what Craig calls the "homeostatic emotions". These include "temperature, itch, distension, muscle ache, hunger, thirst, 'air hunger' and sensual touch" (Craig, 2003a, p. 304). Homeostatic emotions are double edged; they are feelings (sensations) but they also have an inherently motivational aspect; they drive what Craig calls 'homeostatic behaviour'. Homeostatic behaviour is necessary when the autonomic systems on their own are not able to keep the body within the bounds of homeostatic viability. Homeostasis as typically understood in the biological and psychological sciences (i.e. as elucidated by (Cannon, 1939)) is "a dynamic and on-going process comprising many integrated mechanisms that maintain an optimal balance in the

physiological condition of the body, for the purpose of survival" (Craig, 2003a, p. 303). So, whilst the autonomic functions which maintain salt, energy, water and oxygen levels are fundamental to homeostasis, homeostasis is not limited to the first-order regulation of these. Rather, "Cannon recognized that a change in any one condition usually affects several measures and elicits integrated, hierarchically organized homeostatic responses that restore optimal balance" (Craig, 2003a, p. 303) and these "hierarchically organised homeostatic repsonses" involve the regulation of the whole system. Craig's position is that part of these homeostatic responses is behaviour driven by these homeostatic needs. A body depleted in energy needs to find energy soon; 'hunger' is the drive to acquire energy, and - depending on the urgency of energy acquisition - can be more or less intense, more or less motivating. The sensation of hunger is the experience of this drive. Craig argues that the valence of the sensation (the pleasantness or unpleasantness) is a correlate of the motivation itself such that at the extreme of unpleasantness, "the discomfort grows until it becomes an intractable motivation – even though it is not 'painful', you must respond if you are to survive" (Craig, 2003a, p. 304).

Interoception is the afferent wing of homeostasis; "the sense of the physiological condition of the body" (Craig, 2003a, p. 303), so we can start to see how interoception, homeostasis and affect are related. While this may be clear in regard to temperature, Craig argues that pain is also a homeostatic emotion, and thus that the categories of interoception and exteroception - which traditionally categorise pain as exteroceptive (as a perception of external sensations) – need to be re-conceptualised.

Historically, pain was often taken to be a subjective phenomenon. Current conventional science, however, takes the view that pain and temperature sensation are submodalities of cutaneous sensation which is exteroceptive (Craig, 2003a, p. 303); they are sensitivity to stimuli originating outwith the organism. Like other exteroceptive sensations they are well-discriminated feelings (unlike visceral feelings). Neuroscientifically this conventional view is considered to be that "[…] pain is represented centrally by convergent somatosensory activity conveyed by wide-dynamic-range cells in the deep dorsal horn of the spinal cord to a modifiable pattern detector in the somatosensory thalamus and cortices" (*Ibid*). However, if this view were correct then damage to, or stimulation of, the somatosensory cortices should affect pain, which Craig argues, observations show is not the case (*Ibid*). Rather, he argues that converging evidence supports a view of pain as a homeostatic emotion (like temperature, hunger, thirst etc.). The evidence for this comes from cat and monkey studies in

which researchers have identified a common pathway for the substrates of discrete sensations including pain, temperature, itch, muscle ache and sensory touch which indicates that "specific activity representing these modalities is conveyed first of all to homeostatic response regions in the spinal cord and the brainstem". (Craig, 2003a, p. 304). Furthermore Craig argues that this research shows that primates have evolved a forebrain system from the hierarchical homeostatic system and that "this provides a discrete cortical image of the afferent representation of the physiological condition of the body (which we term interoception), along with direct activation of limbic motor cortex" and further that the data indicate that "in humans pain is an emotion that reflects specific primary homeostatic afferent activity. (Craig, 2003a, p. 304)"

Here in detail are the reasons for thinking that temperature and pain are homeostatic emotions and thus interoceptive (from (Craig, 2003a, 2003b)):

(1) There is a common pathway for fine homeostatic afferents (originating in lamina I neurons in the dorsal horn of the spinal cord and trigeminal nucleus[20].
(2) In all mammals this pathway conveys information to homeostatic response regions in the spinal cord and brainstem. (e.g. parabrachial nucleus (PB) and solitary tract (NTS)).
(3) In all mammals information from the parabrachial nucleus (PB) projects (a) to the medial dorsal nucleus (MDvc) of the thalamus and from there to the anterior cingulate cortex/limbic motor cortex (ACC); and (b) to the ventral medial nucleus of the thalamus (VMb) and from there to the insula/limbic sensory cortex.
(4) In primates, in addition to the above pathways, there is also a direct lamina I pathway (the lateral spinothalamic tract) which bypasses the brainstem areas and projects directly to the posterior ventral medial nucleus of the thalamus (VMpo) which is small in macaque monkeys but large in humans.
(5) In primates the VMpo (like the VMb in all mammals) projects to the interoceptive cortex in the dorsal margin of the insula[21].

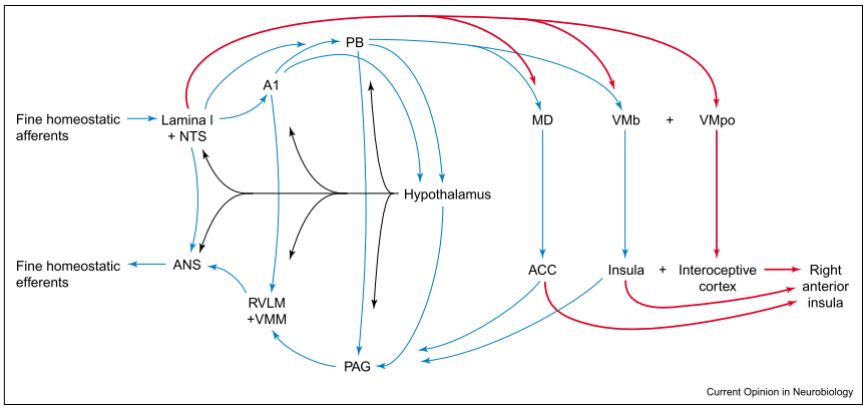
---

[20] These receive input from "[…] Aδ nociceptors (first, sharp pain), C-fiber nociceptors (second, burning pain), Aδ cooling-specific thermoreceptors (cool), C-fiber warming- specific receptors (warmth), ultra-slow histamine-selective C-fibers (itch), tactile C-fibers (sensual touch), and Aδ and C mechano- and metabo-receptors in muscles and joints (muscle exercise, burn and cramp)" (Craig, 2003b, p. 501).

[21] "Converging functional imaging studies in monkeys and humans reveal that interoceptive cortex is activated in a graded manner by noxious stimuli (pain), temperature […], itch, muscle exercise, cardiorespiratory activation, hunger, thirst, and sensual touch […]. This distinct cortical area is well-demarcated by in situ labeling for receptors of corticotropin releasing factor […], consistent with a

(6) In humans the resulting representations in the interoceptive cortex (dorsal margin of the insula) are re-represented in the middle insula and again in the right anterior insula, which is typically activated in imaging studies of human emotions and (in addition) based on evidence from FMRI studies on subjective ratings of cooling stimuli appear to be associated with subjective feelings (see Craig, Chen, Bandy, & Reiman, 2000).

(7) In primates, in addition to the lamina I pathway to the VMpo, there is also a direct lamina I pathway to the medial dorsal nucleus (MDvc) which then projects to the ACC (limbic motor cortex).



Two pathways for homeostatic afferents in primates (red & blue) and non-primates (blue only) (Diagram from Craig, 2003b).

The direct lamina 1 pathway to the thalamus and from there to - on the one hand - the interoceptive cortex, the re-representations of which information in the anterior insula appears to be associated with subjective feeling, and - on the other hand - to the limbic motor cortex (ACC) which subserves motivational drive, explains how homeostatic emotions have the dual aspect of sensation and motivation. While it is the case that in all mammals the ACC (the limbic motor cortex) and insula are activated as part of the homeostatic emotion, what is specific to primates is (1) the addition of a direct pathway which bypasses the brainstem regions, (2) the pathway through the VMpo to the interoceptive cortex, and (3) the re-representations of information from the limbic motor cortex, and interoceptive cortex in the right anterior insula. Craig draws from this the conclusion that, despite the links to ACC and

---

major role in homeostasis as limbic sensory cortex. Lesion, stimulation and evoked potential studies confirm the role of this primary cortical region in pain and temperature sensation and in autonomic function […]. A corollary Vmpo projection to area 3a in sensorimotor cortex may relate cutaneous pain to (exteroceptive) somatic motor activity […]" (Craig, 2003b, p. 502).

insula non-primates do not feel in the same way that we do. In primates the VMpo projection to the interoceptive cortex also has a corollary projection to area 3a in the sensorimotor cortex, which is known to have projections to primary motor cortex and to be involved in pain processing.

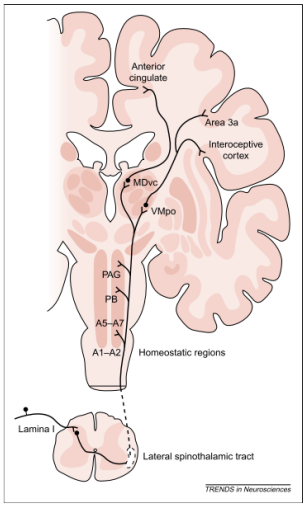


Illustration of the direct Lamina 1 pathway in primates (Diagram from Craig, 2003a).

Craig's argument that pain is a homeostatic emotion gives him a good neuroanatomical basis for recategorising pain, along with temperature and sensual touch, as interoceptive rather than exteroceptive. More theoretically, the intrinsic motivational aspect of pain also serves to distinguish pain from the exteroceptive senses which, though they can be *action guiding*, do not seem to be intrinsically motivating in the same way as the interoceptive senses (this motor component of interoception grounds the dual face of affect and may provide a basis for thinking of affect in terms of a minimal appraisal mechanism which I will explore in section three). Further to recategorising pain as interoceptive rather than exteroceptive, Auvray, Myin, & Spence (2010) argue that the lack of three specific spatial characteristics (transitivity, reversible exploration, and interposition) in the experience of pain also serves to distinguish pain from the external senses which afford information about a certain type of objecthood through these characteristics. It follows from this that pain should be thought of neither as an exteroceptive perception (i.e. perception of an object external to the body) nor an internal perception (i.e. like exteroception but with the object internal to the body). Whatever the objects of pain are, they are not manipulable in the ways that the objects of exteroception are (i.e. the characteristics which confer 'objecthood' upon them). They share this characteristic with the interoceptive senses.

It seems clear that temperature and light touch also lack the above spatial characteristics which would confer objecthood upon them, lending more weight to Craig's recategorisation of all three from exteroception to interoception. Further to this, a recent review by (Boulais & Misery, 2008) shows that skin cells (keratinocytes, melanocytes, langerhans cells, and merkel cells) also play a sensory role in skin surface perception in touch and pain, passing information along through chemical means to the nerve fibers. In addition, these cells link the surface sensory system to the immune system resulting in a neuro-immuno-cutaneous system, with Langerhans cells responding to elevation of temperature by increasing their tendency to bind to antigens, and subsequently migrating from the epidermis to the lymph nodes to initiate protective immunity (Boulais & Misery, 2008, p. 124)[22]. Recall the hallmarks of interoceptive perception were that it depends primarily on chemical, diffuse methods of information transfer, as well as being directly linked to homeostatic regulative activity. When this is taken into account it seems to make much more sense for skin surface perception to also be considered interoceptive rather than exteroceptive.

In this section we have seen that information about the internal body is the afferent wing of an overarching homeostatic loop. The integration of this information supports basic psychological functioning and 'homeostatic behaviour'. We have seen that this behaviour is not so much a *result* of the interoceptive information, but rather is an integral part of the afferent homeostatic pathway in both primates and non-primates. Reconceptualising pain and touch as interoceptive allows us to see these as affective aspects of homeostatic behaviours and as supporting a basic kind of cognitive interaction with the environment. In the next section I will flesh out how we can think about the experience of affect as grounded in this interoceptive, homeostatic functioning.

[22] I will discuss this research in further detail in the final chapter.

Section 3: Grounding 'core affect' in interoception

Understanding how these basic sensory-motivational/action circuits are based in homeostatic regulation (and that the afferent limb of this regulation (i.e. interoception) can - in primates at least – give rise to activity which is associated with felt sensations) allows us to see that affect (understood as basic, primordial feeling, and as distinct from the emotions proper) is grounded in interoception. In addition, the direct links to areas which subserve action explains how basic affects – even in animals – have the capacity to intrinsically motivate the organism in question to keep its functioning within the bounds of viability. Interestingly, having such a powerful homeostatic behavioural system in place means that we can understand the behaviour of animals and infants without having to attribute emotions to them and importantly without being accused of behaviourism. While we may normally describe a mouse as being scared of the cat, it may be more accurate to say that the mouse is affected and - because affect is grounded in interoception, which is the afferent aspect of the homeostatic regulatory circuit, which also includes activity underpinning homeostatic behaviour - thus motivated to action, by the cat. However, even though affect under this view is a more primordial concept than emotion it is still a complex one. In this section I am going to explore the experience of affect in more detail and investigate how its aspects are grounded in the interoceptive and regulative activity discussed in the previous sections.

3.1 Dimensions of the experience of affect

The distinctions between (the experience of) affect and sensation, and affect and emotion were brought out in the late 19th and early 20th centuries as a result of debates centering around Wundt and Titchener (Barrett & Bliss-Moreau, 2009, pp. 168-173). Wundt argued that "simple feelings" are distinct from "simple sensations". For Wundt, sensations are accompanied by "sense-feelings" or an "affective tone" which are made up of more factors than just the 'pure' feeling of sensation. Wundt compares the relation of the affective tone and the simple feelings by comparing the affective tone to a chord made up of the tones c-e-g; the feeling of harmony (affective tone) connected with this chord cannot be separated from the simple feelings connected with each tone (Wundt, 1897, sec. 7.1). Like a harmony constituted by tones, the affective tone of a sensation, though it is constituted by simple feelings, is irreducible to those simple feelings: "The feeling that corresponds to a sensation, is as a rule, […] a product of the fusion of several simple feelings, though it is still as irreducible as a feeling of originally simple nature" (Wundt, 1897, sec. 7.4).

I understand Wundt to be arguing that a sense-feeling/affective tone is the psychological accompaniment to physical sensation. The pure sensory tone that one might normally attribute to the physical sensation is thus accompanied by the affective tone, and this affective tone is not a pure tone made up of just one dimension, but of several such that a shift in direction of one of the dimensions – towards one or other of the opposite ends - shifts the whole tone. Wundt distinguishes three "directions" designated by the two names that indicate their opposite extremes: pleasurable and unpleasurable feelings; arousing and subduing (exciting and depressing) feelings; and the feelings of strain and relaxation (Wundt, 1897, sec. 7.7). Barrett & Bliss-Moreau (2009) (henceforth BBM) translate Wundt's dimensions as valence, arousal, and intensity respectively. However, while Wundt talks in terms of intensity (and quality) earlier in the chapter, in regard to the affective system and the sensational system, this is not in such a way as to suggest that these are the two dimensions of affect (indeed if it were it would be unclear as to how "quality" is supposed to encompass both "pleasurable and unpleasurable feelings" and "arousing and subduing feelings" which it surely must if intensity is supposed to be "feelings of strain and relaxation"). Rather it is in analogy to the determinants of sensations:

> The varieties of simple sense-feelings are exceedingly numerous. The feelings corresponding to a particular sensational system also form a system, since, in general, a change in the quality or intensity of the affective tone runs parallel to every change in the quality or intensity of the sensations. At the same time these changes in the affective systems are essentially different from the corresponding changes in the sensational systems, so that it is impossible to regard the affective tone as a third determinant of sensations, analogous to quality and intensity. If the intensity of a sensation is varied, the affective tone may change not only in intensity, but also in quality; and if the quality of the sensation is varied, the affective tone usually changes in quality and intensity both. For example, increase the sensation sweet in intensity and it changes gradually from agreeable to disagreeable. Or, gradually substitute for a sweet sensation one of sour or bitter, keeping the intensity constant, it will be observed that, for equal intensities, sour and, more especially, bitter produce a much stronger feeling than sweet. In general, then, *every sensation is essentially accompanied by a twofold change in feeling.* The way in which changes in the quality and intensity of affective tones are related to each other follows the principle already stated that every series of affective changes in *one* dimension ranges between *opposites,* not, as is the case with the corresponding sensational changes, between greatest differences. (Wundt, 1897, sec. 3)

Even though Wundt talks about "[t]he way in which changes in the quality and intensity of affective tones are related to each other", it is clear from the examples given (changing from "agreeable to disagreeable" in the first, and producing a "much stronger feeling" in the second) and the emphasis on each dimension ranging between opposites, that "quality" and

"intensity" in regard to the affective tone, in this context, is referring to the valence, while the intensity corresponds to the extreme ends of the dimension. This means that a much stronger negative feeling equates to being much nearer the unpleasant end of the pleasantness dimension, rather than being a conglomorate of negative affect and high intensity. Compare this to the example of the sensation of sweetness above. The quality of the sensation is sweetness, and with an increase in intensity this does not mean an increase in sweetness as such; it is not increasing on some dimension of sweetness/unsweetness (or sourness), but rather it is increasing on the intensity of experience scale (that this has consequences for the pleasantness isn't in question here, I am just focussing on the actual sensations). I therefore suggest, contra BBM, that Wundt is here not suggesting that quality and intensity are dimensions of affect but rather using these terms as descriptive terms for the variations in affective tone to bring out the analogy and disanalogy with sensations whose dimensions are quality and intensity, and the relations between sensational intensity and affective tone.

Finally, Wundt's example of "strain and relaxation" suggests we should not understand this dimension as isomorphic with intensity:

> Feelings of strain, and relaxation are always connected with the temporal course of processes. Thus, in expecting a sense-impression, we note a feeling of strain, and on the arrival of the expected event, a feeling of relaxation. Both the expectation and satisfaction may be accompanied at the same time by a feeling of excitement or, under special conditions, by pleasurable or unpleasurable feelings. Still, these other feelings may be entirely absent, and then those of strain and relaxation are recognized as specific forms which can not be reduced to others, just as the two directions mentioned before. (Wundt, 1897, sec. 7.8)

And further:

> Again, the direction of pleasurable and unpleasurable feelings, is united with that of feelings of strain and relaxation, in the case of the affective tones of rhythms. The regular succession of strain and relaxation in these cases is attended by pleasure, the disturbance of this regularity by the opposite feeling, as when we are disappointed or surprised. Then, too, under certain circumstances the feeling may, in both cases, be of an exciting or a subduing character. (Wundt, 1897, sec. 7.8)

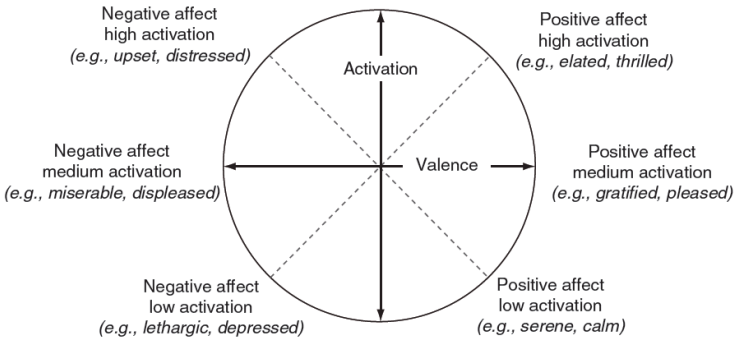
I suggest that strain and relaxation, rather than equating with feelings of intensity seem rather to correspond to feelings of anticipation and resolution, which should also be thought of as a dimension of core-affect[23].

[23] I suggest how this might be grounded in a predictive theory of attention in the next chapter.

3.2 Core affect and the 'circumplex'

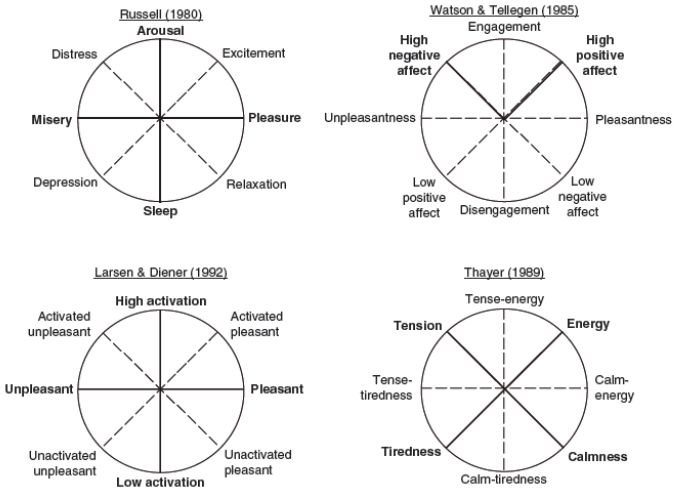
Following Wundt, BBM argue that affect is a psychological primitive, that is, "a fundamental, psychologically irreducible property of the human mind" (BBM, 2009, p. 167). They propose a concept of "core-affect" which, like Wundt's proposal, consists of the dimensions of valence and arousal. As mentioned above they interpret Wundt's third dimension as intensity and reject this as a separate dimension, as they propose that it can be accounted for in terms of the ends of the spectrums of valence and arousal.

BBM visualise the mental structure of core affect using what they call the 'circumplex', illustrated below. The circumplex represents the relations between the stimuli (round the circle) and the axes represent the psychological properties that quantify the similarity/difference between people's reactions to the stimuli (BBM, 2009, p. 179). Circumplexes are a result of heterogeneous descriptions – in this case the dimensions of valence and arousal - being projected onto geometric space. Something is heterogeneous when it cannot be compared using only one property as tokens simultaneously vary on another property as well.



"The affective circumplex. Hedonic valence is represented on the horizontal axis and arousal on the vertical axis" (Barrett & Bliss-Moreau, 2009, fig. 4.3).

The question as to what the dimensions of core-affect are is not uncontroversial. (BBM, 2009, pp. 182-183) note that other dimensions that have been proposed include: positive and negative activation; positive and negative affect; approach and withdrawal; and tense and energetic activation. All these dimensions can, like valence and arousal, be incorporated on the circular structure of the circumplex as they show on the diagram below:



"Multiple affective dimensions mapped in circumplex space. Primary (or main) dimensions are indicated in with black solid lines and labelled with capital letters. Secondary dimensions are indicated with gray dotted lines and are labeled in lower case letters." (Barrett & Bliss-Moreau, 2009, fig. 4.5).

One of the reasons for the dispute about what the dimensions of affect are, is predominantly grounded in the assumption that the descriptive aspect of affect should be isomorphic with its causal structure:

> Like Titchener, many scientists continue to believe that the descriptive structure of affect should be isomorphic with its causal structure, so that the best affective dimensions are those that are most causally plausible (i.e., the dimensions should reflect the processes that cause affective states). Accordingly, it has been claimed that certain dimensions (e.g., positive and negative affect) are more biologically basic, and therefore should be the preferred anchors of affective space. (Barrett & Bliss-Moreau, 2009, p. 185)

The dimensions of core-affect which BBM propose are based on its descriptive structure and not causal structure, and they see no reason to think that the dimensions ought to reflect the processes which cause affective states. They argue that many biological arguments of "basic" dimensions (such as positive and negative affect) which are proposed to anchor affective space are flawed, based as they are in claims that positive and negative affective states are realised in anatomically different parts of the brain; different hemispheres; or different neurotransmitters (2009, p. 186). BBM thus think that valence and arousal are best thought of as descriptive features of core-affect "that bear no resemblance to or inform us about how affect is caused. Simply put, content does not necessarily tell us anything about

process." (2009, p. 188)[24]. BBM still think, however, that descriptions can be scientifically useful even if they aren't casual. They argue that we should consider valence and arousal as the basic features of core-affect, rather than for example positive and negative activation/states, because they emerge as features across many domains of psychological response: judgments of emotion related language; judgments of facial behaviours; subjective ratings of emotional episodes, e.g. anger, sadness, and fear; subjective ratings of non-emotional affective states, e.g. fatigue, sleepiness, and placidity. Positive and negative activation, on the other hand, is only identified in self reports of experience and not in judgments of words or faces (2009, p. 188). They also cite ERP and neuroimaging evidence that valence is a basic aspect of face perception (2009, p. 189).

Given that valence and arousal are descriptive features of our psychological states, one might suppose that each dimension would be clearly apparent to us. Interestingly BBM show that while all humans can tell the difference between pleasant and unpleasant affective states, not all (although many) characterise their experience as high or low in activation, i.e. arousal (2009, p. 195). BBM distinguish two types of focus that one can have on one's affect; valence focus and arousal focus. Valence focus represents the amount of information about pleasure or displeasure (hedonic states) contained in the verbal reports of emotional experience (note that this does not have to do with the tendency to report such states – at least that is not what is being measured). Arousal focus, on the other hand, is the amount of information about felt activation or deactivation contained in the verbal reports (2009, p. 194). So valence focus (VF) and arousal focus (AF) reflect the extent to which valence or arousal are important descriptive properties of core-affective responding in that individual and these foci are related to what Barrett calls emotional granularity. High emotional granularity reflects equal focus on arousal and valence, whereas low emotional granularity reflects either more focus on valence or more focus on arousal. If the valence focus is stronger than the arousal focus, individuals have difficulty distinguishing between states which have the same valence but differ in arousal. For example negative states which differ in arousal are anger and sadness; someone with low emotional granularity may find it difficult to tell the difference between these two states and thus use the terms interchangeably to denote any state with negative valence. Those individuals who focus more on arousal than valence on the other hand have difficulty distinguishing between

[24] While I think that this is right to a degree, I think we can still ground these dimensions in the bodily and neural processes, as I will elaborate on later in the section.

arousal states that differ only in hedonic valence, such as nervousness and excitement, both high arousal states but which have very different hedonic valence (BBM, 2009, p. 195).

BBM show that individual differences in VF & AF relate to other psychological phenomena. For example, VF individuals are also more perceptually sensitive to hedonic information in the face of others; VF individuals describe themselves as being more sensitive to hedonic cues (as measured by the traditional personality measures of extraversion and neuroticism); VF individuals experience the world as "a rollercoaster full of drama" and this is linked to self esteem lability, in that they use the hedonic information that they perceive in their environments (which is far greater than others because of their sensitivity to this stimuli) to shape and change their sense of self (2009, p. 196). Arousal focus, in contrast, seems to be related to an enhanced sensitivity to ones own physical state (i.e., interoception). This was shown in the heart-beat study (Barrett, Quigley, Bliss-Moreau, & Aronson, 2004), reviewed in (BBM, 2009, p. 196), in which participants had to judge whether a series of tones were in sync with their heartbeat or not. The results were that those who scored higher on arousal focus showed enhanced sensitivity to their heartbeats:

> These findings indicated that people who have more awareness of the internal sensory cues coming from their body also experience more variation in the arousal-based property of core-affect. They clearly showed that people can, at times, detect specific information in their bodies, and this sensitivity is, in some way, related to the experience of emotion. (Barrett & Bliss-Moreau, 2009, p. 196).

BBM argue that the AF-interoception link helps to clarify the link between interoceptive sensitivity and emotional experience. Given that there have been inconsistent results from studies on the link between heartbeat detection and ratings of intensity of emotional experience they thus propose that it may be that interoceptive sensitivity is better conceptualized as relating to *arousal* rather than *intensity*. BBM argue that such correlations between AF, VF, and other psychological phenomena such as those above show that AF and VF are valid psychological constructs.

3.3 The dual face of affect: a mechanism for minimal appraisal

Barrett describes core-affect as "the constant stream of transient alterations in an organism's neurophysiological state that represents its immediate relationship to the flow of changing events" (Barrett, 2006, p. 39). *Core-affect* is the experiential dimension of *affect*, which is understood to be the changes in homeostatic regulation:

> Computations of value (whether an object is helpful or harmful) are represented as perturbations in a persons internal milieu – these changes are what we mean when we say that a person has an affective reaction to an object or stimulus. They are means by which information about the external world is translated into an internal code or representations (Barrett, 2006, p. 39).
>
> In a sense, core affect is a neurophysiologic barometer of the individuals relationship to an environment at a given point in time and self-reported feelings are the barometer readings. Feelings of core affect provide a common metric for comparing qualitatively different events (Barrett, 2006, p. 40).

From this description we can see that core-affect has two aspects; one which faces towards (or perhaps it is better to say is 'integrated into', or 'consists in') the neurophysiology, and one towards the experiential, descriptive level. So while it is right to think of core-affect being expressed through the dimensions of valence and arousal, both of which are situated at the psychological level, core-affect is fundamentally a biological phenomenon "grounded in the somatovisceral, kinesthetic, proprioceptive, and neurochemical fluctuations that take place in the core of the body" (Barrett & Bliss-Moreau, 2009, p. 171). Though it is grounded in these fluctuations, it is not identical to these; these may be thought of as (non-core) affect and the representations of these fluctuations as they occur is the neurophysiological aspect of core-affect.

Placing Barrett's work into dialogue with the work on interoception and the homeostatic emotions discussed earlier in the chapter we can work towards a more complete model for grounding core-affect. As the work on core-affect stands, we might postulate that arousal is grounded in the interoception of sympathetic nervous system activity, which though increased in highly arousing situations, is constantly at play in order to maintain homeostasis. However, valence, which is supposed to be the core of the affect system (Barrett, 2006, p. 39), is not so easy to ground in interoception. This becomes clearer when we consider other terms for that which 'valence' is used to refer: hedonic tone; utility; good-bad mood; pleasure-pain; approach-avoidance; rewarding-punishing; appetitive-aversive; and positive-negative (Barrett, 2006, p. 40). Valence seems to be entwined with action in many of these. Valence is rooted in the concept of value[25] and Barrett & Bliss-Moreau gesture towards this when – appealing to research by (Owren & Rendall, 1997; 2001) - they suggest that core affect represents a basic kind of psychological meaning:

[25] See Colombetti (2005) for a detailed discussion of value and valence.

> The basic acoustical properties of animal calls (and human voices) directly act on the nervous system of the perceiving animal to change its affective state and in so doing conveys the meaning of the sound. (Barrett & Bliss-Moreau, 2009, p. 172)

While I don't want to pursue the issue of whether *meaning* can be reduced to affective changes, we could instead think of meaning here as being the appraisal of value to the organism. As we saw in chapter one, 'appraisal' has many cognitive connotations. But I suggest that there is a minimal kind of appraisal which is not grounded in deliberation, and that is perhaps even more primitive than the kind of subpersonal appraising mechanisms outlined in chapter one. I suggest that appraising value to the organism can even be as basic as a response to a stimulus which is appropriate for maintaining homeostasis. In this minimal sense, homeostatic behaviours such as withdrawing from a painful stimulus or seeking water when thirsty are results of an appraisal that the current situation is incompatible with maintaining homeostatic viability. But as we saw in the discussion of homeostatic emotions based on (Craig, 2003a, 2003b), the action/behaviour is not so much a *result* of the interoceptive information, but rather is an integral part of the afferent homeostatic pathway in both primates and non-primates. Recall that in the basic pathway common to primates and non-primates (which rises through the brainstem to the limbic motor cortex) is directly involved in the loop receiving projections from the medial dorsal nucleus of the thalamus and sending projections on to the periaqueductal gray. And, in the primate specific pathway, the limbic motor cortex is also activated in virtue of direct projections from lamina I, and subsequently projects on to the right anterior insula, in addition to area 3a of the sensorimotor cortex (which projects directly to the primary motor cortex) receiving corollary projections from one of the afferent projections from the thalamus to the interoceptive cortex in the insula.

It can be seen from this that the 'motor areas' of the central nervous system are part of the very homeostatic loop itself, rather than functioning – at this basic level – as a result of deliberation conceived in either personal or subpersonal level terms. They are so entwined with the afferent homeostatic signals which ground interoception that it looks as though interoception is not merely the passive experience of the physiological changes in the body, but has the motor aspects already factored in. In other words, interoception includes motor information. So if, as I have been arguing, affect is grounded in interoception, as the sense of the changes of the physiological body, and these changes include information about and preparation for homeostatic behaviour, then interoception is by nature functioning as a basic appraisal machine adapting the system in response to perturbations from the environment.

Valence then, (behaviourally perhaps better understood in terms of approach/withdrawal or reward/punishment), arises as the felt aspect of this interoceptive/motor system. In other words valence is an *affective motivation*. This fits with Craig's articulation of pain as a homeostatic emotion:

> […] non-painful thermal stimuli inherently produce an affective motivation, a 'feeling' of pleasantness or unpleasantness that depends on physiological context, and they generate reflexive autonomic adjustments. These aspects directly signify the homeostatic role of temperature sensation (Craig, 2003a, p. 303).

I suggest that this inherent motivational aspect of valence (or rather the mechanisms of which valence is the experiential aspect) explains why, as discussed in the section on pain dissociations brought out in (Grahek, 2007), in all but lobotomised patients, valence seems tied to the emotional-cognitive and behavioural components of pain. The lack of the motivational aspect of affective motivation in lobotomised patients may therefore be a function of the disruption of the fundamental homeostatic afferent loop in the brain. We might expect therefore, that all interoceptive sensations have this dimension of affective motivation, as if the line of thought presented in this chapter is right, interoceptive sensations ground affect. As explained earlier, Craig argues that pain is a homeostatic emotion, and should be categorised as interoceptive. It is clear that typical pain has affective motivation as a component. Likewise temperature, and sensual touch both have affective motivation as a component; there is a direct pathway to both interoceptive and motor centre's in the brain, and phenomenologically we can distinguish the kind of compulsion to move that these sensations bring about (or rather is brought about in virtue of the processes which underpin the sensation – I don't want to imply that this is a linear causal process) from the motivation to action that arises from exteroceptive sensing such as seeing or smelling something unsavoury. I will discuss the relation between affect and exteroception further in the next chapter, but for now I just want to bring out their inherent affectivity.

So far, following Barrett, I have focussed my argument on the valence and arousal dimensions of affect. What then of the third dimension of affect that Wundt proposed, which I argued was not encapsulated by intensity, and which is expressed in Barrett's core-affect as the extreme ends of the dimensions of valence and arousal? It may be that strain-relaxation can be accounted for in terms of valence if valence is indeed grounded in the motor aspect of the homeostatic/interoceptive system. Certainly if we think in terms of approach/withdrawal there seem to be some similarities with strain/relaxation. I suspect however that this dimension is distinct from valence even as I have grounded it above. Rather, it seems to be

linked to a more fundamental aspect of the simultaneous generation of sensation and motivation that characterises the homeostatic emotions. I will discuss this in further detail in the next chapter.

3.4 Summary of chapter

Using an analysis of the dissociations in pain phenomena I have argued that affect can be distinguished from emotion. I then argued that affect is grounded in interoception, with the distinct dimensions of core-affect grounded in different aspects of this; in particular that arousal is grounded in sympathetic nervous system activity and valence in the motor aspects of the homeostatic loop (affective motivations). I suggested that this would explain why, in the analysis of the pain components early in the chapter, we saw that valence and the emotional-cognitive and behavioural components could, in all but lobotomised patients, not be dissociated. It would also explain why, in the case of the pain substitution devices, there was a distinct behavioural dimension to the valence (i.e. to disconnect the device) even though there was no emotional-cognitive component (i.e., threat) or corresponding (to the emotional-cognitive component) behavioural component (i.e., escape the threat) which in typical circumstances would take the threating stimulus as its object. These aspects of affect may be thought of as a minimal appraisal mechanism grounding the 'homeostatic emotions' and as foundational for flexible, adaptive behaviour.

Chapter IV

The role of affect in perception and cognition

We have seen that affective experience may be grounded in interoception (the afferent wing of homeostatic regulation). In this chapter I argue that affect infuses perception, perceptual phenomenology and mentality more generally. Moreover, affective information is integrated in 'cognitive' processing such that it is no longer clear that it is helpful to talk in terms of 'affective' or 'cognitive' processing. In chapter five, I will proceed to argue that this integration can plausibly be taken to be indicative of affect being partially constitutive of cognition.

Section 1: Interoception and perception

The recurrent nature of neural processing provides the opportunity for previously processed information to feed into areas which underlie even very early vision. As I discussed in chapter three, Damasio proposes that the superior colliculus provides the opportunity for exteroceptive and interoceptive information to integrate. The reason for this is that the superior colliculus receives projections from interoceptive pathways as well as exteroceptive pathways and that the maps (neural representations) of the information from these pathways are stacked in a spatial register so that "the information available in one map for, say, vision, corresponds to the information on another map that is related to hearing or body state" (Damasio, 2010, p. 84). The superior colliculus receives information from cortical as well as from subcortical areas and its ability to integrate the signals from the various sensory modalities (rather than just respond to them independently) is dependent upon inputs from the association cortex, though the exact mechanism for this integration is still not entirely understood (Cuppini, Ursino, Magosso, Rowland, & Stein, 2010). Recent work in neuroscience, however, shows that in addition to integrating exteroceptive and interoceptive information, the superior colliculus may feed into visual perception at an even earlier stage.

1.1 Top down information can influence early vision

There is evidence that the integration above feeds into early visual processing. Integrated information from the superior colliculus is known to play a role in the winner-takes-all

selection of single targets (in the presence of distractors) for eye-movements (Nummela & Krauzlis, 2011). Nummela & Krauzlis argue that this information also plays a role at an even earlier stage; that even before the selection of a target, superior colliculus activity is important for the integration of visual signals for orienting towards that target. Thus it is not just motor plans for saccades which are a function of the integrated multisensory information but this information is involved right from the initiation of smooth pursuit eye movements which in some cases precede saccadic eye movements.

Both 'bottom-up' and 'top-down' processing feeds early visual processing. From the bottom-up perspective, it has been hypothesised that the multimodal integration in the superior colliculus forms a 'priority' map for the selection of behaviourally relevant stimuli (Fecteau & Munoz, 2006). Computational and psychological models of visual attention and object selection have already been using the concept of a theoretical 'salience map' to explain how we come to select objects in a scene for attention when (as is typically the case) the complexity of that scene precludes the possibility of the complete scene being represented. Fecteau & Munoz argue that this map is located in the oculomotor network which is spread across the frontal eye fields, lateral intraparietal cortex, superior colliculus, and the brainstem reticular formation (Fecteau & Munoz, 2006, p. 384). The organisation of the superior colliculus is such that the superficial layers and the intermediate layers receive input from different stages in the sensory-to-motor processing path, with the superficial layers receiving input from the early stages (retina, primary visual cortex, and areas V2 and V3), and the intermediate layers receiving input from the later stages (lateral intraparietal area, frontal eye fields, dorsal lateral prefrontal cortex, and inferotemporal cortex) (2006, p. 385). Salience in visual attention, defined as the "physical, bottom-up distinctiveness of an object" in respect to other objects in the same scene (2006, p. 382), and inhibition of return (the property of not returning to a salient object after the initial capture of attention – so as to allow attention to subsequently be focussed on other targets) both originate late in sensory-processing in visual search, "which means that the salience map is not a summary of visual processing occurring at early stages of the visual hierarchy" (2006, p. 386).

Top-down processing is also involved in early vision. 'Relevance' is a top-down process which influences how an object is processed (2006, p. 387). The input from bottom-up and top-down sources converge in the superior colliculus to produce "an amalgamated representation of 'priority' " (*ibid*) which is correlated with both salience and relevance. Despite their sharing neural correlates, Fecteau & Munez argue that salience and relevance

have unique contributions as they yield distinct neural signals with salience being reflected in the initial registration of the target, whilst relevance is "reflected in the elevated activity following the predictive cue" (2006, p. 387). This account of relevance fits well with Feldman & Friston's (2010) account of attention in respect to the generalised predictive coding framework for neuroscience which I will discuss in more detail in the next subsection.

There are two main points that I want to draw from the discussion above. Firstly, salience and inhibition of return – basic early visual properties – are shown to originate late in sensory processing. This means that even such basic properties of visual search as these could *potentially* be being informed by information that is not directly 'visual'. Secondly, the hypothesis that a 'priority' map (rather than a mere 'salience' map) is instantiated in the superior colliculus, and is part of the mechanism for early vision, entails that top-down (relevance) inputs are involved in processing which has previously been considered to be purely bottom-up. Top-down input is by definition previously processed, and this provides yet another opportunity for information from outwith the strictly visual network to be integrated in early vision. Given this integration it is unclear how we might demarcate processes that are 'mere' causal influences on the processing from that which is an 'actual part' of the processing and I will address this issue in chapter five.

The mere opportunity for affective information to be involved in early visual processing of course does not show that it *is* involved. However, now that we are open to the integrated and recurrent nature of early processing we can consider a model of visual processing by Barrett & Bar (2009) which builds upon such properties of neural functioning and argues that affect is integrated into visual processing at several stages - including the formation of the initial gist of a situation.

1.2 Affective perception

In their (2009) paper on affective predictions in object perception, Barrett and Bar put together recent research on visual processing in light of the generalised predictive coding approach to neuroscience[26] to argue that both 'gist-like' and specific visual information is laden with affective value.

[26] The generalised predictive coding model generalises the mathematical predictive coding model to the levels at which we can understand brain processes in relation to human functioning. Wherever I

At the heart of the generalised predictive coding approach is the hypothesis that the brain is essentially a prediction engine, and the information that we garner from the world is encoded in the errors in these predictions. The brain will continue to recalibrate and generate new predictions until the incoming sensory states match those predictions (Bar, 2009; Friston, 2009; Friston & Kiebel, 2009). Generalized predictive coding is currently being used as a unifying framework guiding understanding at various levels in neuroscience, from the statistics of neural firing, to the level of us as agents interacting in the world. Barrett and Bar address prediction somewhere in between these levels. They argue that perception works in a similar way to the Dutch style of painting in the sixteenth and seventeenth centuries; first the gist of a situation is sketched, then, over time through the recursive application of ever-smaller dabs of paint, a detailed picture emerges. The recursive (and ever finer) dabs of paint in this example, correspond to the recursive predictions which are generated as a result of errors in the predictions of sensory states. Their thesis is that object perception arises partly as a result of predictions about the value of that object to the agent (either generally or at this particular moment in time).

Drawing on research from Aude Oliva's computational visual cognition lab at MIT (see for example Oliva & Torralba, 2001), Barrett and Bar propose that the brain quickly makes an initial prediction about an object using low spatial frequency visual information. The details are then filled in by memory; given the overall gist of a situation or object in context, the brain is left to predict what the details might be given previous knowledge. Direct projections between the visual cortex and areas of the prefrontal cortex provide a pathway for this recursive (re)-creation of the visual experience of the object. The previous knowledge which is used to fill in the gist of the prediction is encoded in sensory-motor patterns which are stored for future use. They argue that the sensory-motor patterns are sensory in its fullest sense; they not only involve external sensations and their relations to actions, but also internal sensations – from organs, muscles, and joints and how external sensations have influenced these internal sensations (Barrett & Bar, 2009, p. 1325).

Barrett & Bar show that the connections between various brain areas give us reason to believe that representations of internal bodily (autonomic and endocrine) changes are part of visual processing right from the stage at which the gist of a situation is being processed by

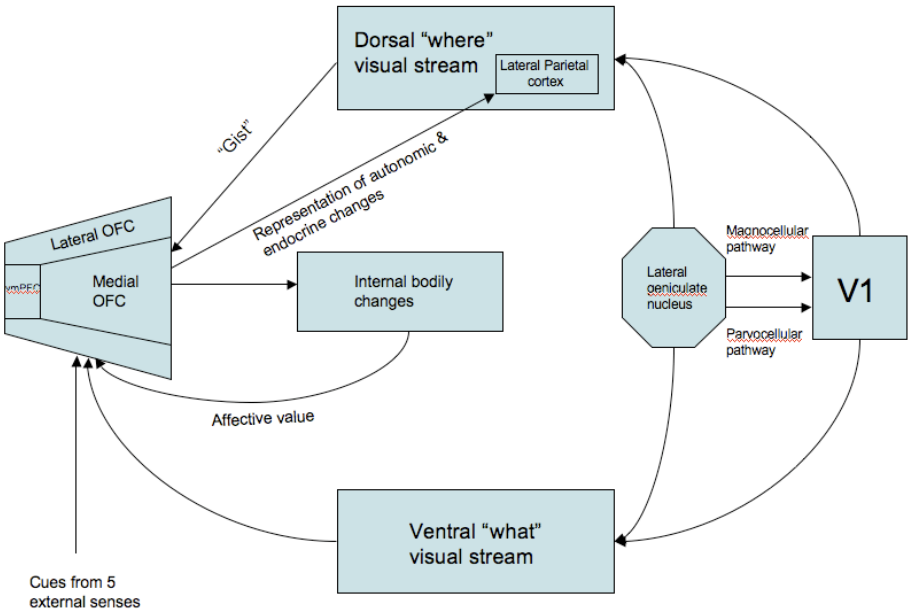
---

refer to "predictive coding" in this text I am referring to the generalised model expounded in (Bar, 2009; Friston 2009; Friston & Kiebel, 2009; Feldman & Friston, 2010).

the frontal systems, giving even perception at this paucity of specificity an affective flavour which helps code the relevance of the object of perception. The model of visual processing that they are proposing runs as follows. Visual information comes through the lateral geniculate nucleus (part of the thalamus) at which point a very unspecific "gist" of this information is sent through the fast magnocellular pathway through the dorsal visual stream, which includes the lateral parietal cortex, and also through fast magnocellular pathways to V1 and from there to the dorsal stream. The dorsal stream sends information on to the medial orbitofrontal cortex (mOFC) which then sends information to the autonomic and endocrine systems to effect bodily changes including preparation for action, and information about those changes that have been ordered to the lateral parietal cortex, feeding that information back into the dorsal stream. This shows that the processing of gist information is affective, as internal bodily changes are caused and (as discussed in chapter three) the representations of these changes are fed back through interoception. These feed into the lateral OFC helping to refine the gist each time with the information about affective value (in terms of information about how the body is faring) that these carry. The idea is that, on Barrett and Bar's model, each time round the processing loops, better and better predictions are being made and the perception of the object is getting less and less gist-like, and at the same time developing more and more meaning (in terms of biological relevance) for the agent in virtue of the affective aspect of the perception.

Highly specific visual information (as opposed to 'gist' information) gets sent on a different route towards the orbital frontal cortex. From the lateral geniculate nucleus it gets sent through slower parvocellular pathways to both the ventral visual stream, and also to V1 and from there to the ventral stream. Information from the ventral stream gets sent to the lateral OFC (rather than the medial OFC as was the case in the dorsal loop). Also feeding into the lateral OFC is information from the external senses and from the internal bodily changes that were effected as a result of processing in the medial OFC. The lateral OFC thus serves as an association area of all of this information from various senses including interoception. So even the more specific visual processing that builds upon the gist that is being created as a result of the dorsal loops is laden with affective value.

Laid out diagrammatically Barrett and Bar's model of visual processing looks something like this:



As my diagram above suggests, upon Barrett & Bar's proposal all the processing described is very recurrent and difficult to cleanly separate inputs and outputs to the process. Even in the simplified form that I have presented Barrett and Bar's proposal it is clear that affective value is feeding in at various levels (i.e. both dorsal and ventral) and becomes part of the very dorsal processing that it is feeding into. They suggest that it is likely that the real picture is even more complex, which would only further the argument that affective value is inherently a part of visual processing:

> Taken together, these findings indicate that it may be more appropriate to describe the affective predictions generated by the medial and lateral OFC as phases in a single affective prediction evolving over time, rather than as two separate 'types' of affective predictions (with one informing the other). This interpretation is supported by the observation that the medial and lateral OFC are strongly connected by intermediate areas; in addition, the lateral OFC receives some low spatial frequency visual information and the medial OFC some high spatial frequency information; and, magnocellular and parvocellular projections are not as strongly anatomically segregated as was first believed (for a review, see (Laycock, S. G. Crewther, & D. P. Crewther, 2007)). Furthermore, there are strong connections throughout the dorsal 'where' and ventral 'what' streams at all levels of processing (Chen et al., 2007; Merigan & Maunsell, 1993). Finally, the OFC has widespread connections to a variety of thalamic nuclei that receive highly processed visual input and therefore cannot be treated as solely bottom-up structures in visual processing. (Barrett & Bar, 2009, p. 1331)

While this model shows how affective information may be integrated in visual processing, Barrett and Bar leave attention out of their model. I argue that this can be remedied by

putting the third Wundtian dimension of core-affect into dialogue with a predictive coding account of attention. Feldman & Friston (2010) provide an account of attention as the process of optimising the post-synaptic responsiveness of the units reporting prediction errors. What this means in practice is that if the system regards some sensory information as high-grade (hence reliable), it can increase the gain on the error-units reporting prediction error, thus giving added weight to that sensory data. In such a case the units encoding that error become highly responsive (synaptic gain). This means they exert a greater force by which to resolve prediction errors. The searchlight of attention thus gets progressively finer until all prediction errors have been resolved, or a greater error arises elsewhere. Barrett & Bar's metaphor of the Dutch painter thus extended might look something like this. A gist of the situation is sketched, and those parts of it that correspond least to that which is being painted (i.e. where large prediction errors occur) are singled out as in need of more detail. The painter focuses in on one of these (increased gain on the units reporting the prediction errors), painting in slightly more detail here (increased gain results in added weight to sensory data coming in exerting greater force to resolve errors). Again, the parts of this that correspond least to what is being painted (large prediction errors) are singled out for more work, and so recursively the painting gets more detailed in those areas until those parts of the painting match up well with the object of the painting, or other inconsistent areas become glaringly obvious, persuading the painter to focus on them in turn.

The incorporation of attention into Barrett & Bar's model leaves it affective in terms of the afferent and efferent interoceptive signaling involved at various levels. Moreover, the addition of the attention story can ground Wundt's third dimension of core-affect, strain and relaxation, which I noted in the previous chapter is not accounted for in Barrett's model of core-affect. I suggest that the directions of strain and relaxation may correspond to the force and resolution of synaptic gain as evidenced in attention and the corresponding experience of being drawn towards that which captures our attention, or releases us.

There may seem to be a tension between Barrett and Bar's model of affective prediction and the Jamesian model of perception preceding bodily changes and the feeling of these changes. This can however be avoided if we are clear to distinguish perceptual processing (and the integration of affect in this) from the act of perception. As Barrett & Bar note (2009, p. 1328), affective responses have largely been ignored in cognitive science, and in emotion research it has been the predominant assumption that affect occurs after object perception. I suspect that part of the reason for this may be the tendency of emotion researchers to

conflate affect and emotion (as I rehearsed in chapter one). For example William James is explicit that emotions (identified as the feeling of bodily changes) occur after perception:

> Our natural way of thinking about these standard emotions is that the mental perception of some fact excites the mental affection called the emotion, and that this latter state of mind gives rise to the bodily expression. My thesis on the contrary is that *the bodily changes follow directly the* PERCEPTION *of the exciting fact, and that our feeling of the same changes as they occur* IS *the emotion.* (James, 1884, pp. 190-191)

So rather than the standard view: perception -> emotion -> bodily expression, James argues that the order is rather: perception -> bodily changes & feeling of bodily changes (simultaneously). In this passage, however I suggest that James is focussing on the emotions and not affect (that is on specific (overt) patterns of bodily changes rather than all internal bodily changes). Even though he does not explicitly make this distinction we can see it implicitly in the passage prior to the one above:

> I should say first of all that the only emotions I propose expressly to consider here are those that have a distinct bodily expression. That there are feelings of pleasure and displeasure, of interest and excitement, bound up with mental operations, but having no obvious bodily expression for their consequence, would, I suppose, be held true by most readers. Certain arrangements of sounds, of lines, of colours, are agreeable, and others the reverse, without the degree of the feeling being sufficient to quicken the pulse or breathing, or to prompt to movements of either the body or the face. Certain sequences of ideas charm us as much as others tire us. It is a real intellectual delight to get a problem solved, and a real intellectual torment to have to leave it unfinished. The first set of examples, the sounds, lines, and colours, are either bodily sensations, or the images of such. The second set seem to depend on processes in the ideational centres exclusively. Taken together, they appear to prove that there are pleasures and pains inherent in certain forms of nerve-action as such, wherever that action occur. (James, 1884, p. 190)

So James holds that there are feelings which do not have an *obvious* bodily expression, i.e. not in the way that (e.g.) fear is at least sometimes clearly correlated with accelerated heart-rate and other signs of high arousal. This suggests that there is a distinction lying here between emotional affect and pre-reflective affect that ought to be made where emotional affect pertains to patterns of bodily changes (in the way that Damasio envisages that I discussed in chapter two) which we focus on and bodily changes which are more subtle and shape our experience of the world (I will pursue this in the next section). However, given that emotions for James are just the feelings of the bodily changes then that gives us no basis to distinguish emotion from the experience of affect (or what Barrett calls 'core-affect'). If this is right then it might seem that even James, who might be considered the most embodied

of embodied emotion theorists, posits that perception of a stimulus occurs first and then affect/emotion. Matthew Ratcliffe argues however that, put in the context of James' other work, there may be a terminological ambiguity in the above passage between on the one hand, 'perception' as a physiological process, and on the other, as the phenomenological outcome of that process (Ratcliffe, 2005, p. 185). Likewise he also argues that "whether the 'object of perception' is the 'external cause of the perception' or the 'experienced outcome of the perception' is not always clear." (2005, p. 185).

Ratcliffe takes James to be claiming that "it is not an *object as perceived* but *sensory stimulation of our perceptual systems by an object* that automatically triggers bodily changes, just as one would expect from an automated reflex-like process" (Ratcliffe, 2005, p. 185). Cashed out like this, James' view is no longer at odds with Barrett and Bar's model of affective prediction. At the level of direct stimulation of the retina, there is no top-down influence from affective information. And, as is clear in Barrett and Bar's model there does not need to be a conscious experience of the perception for the bodily changes to be triggered. The stimulation of the retina part of the perceptual process may be the only point at which affect is not involved in perception however, as research by (Damaraju, Huang, Barrett, & Pessoa, 2009) suggests that even as early in visual processing as at V1 affective information modulates processing, enhancing activity and modulating functional connectivity.

1.3 Affect structures our perceptual phenomenology

Barrett and Bar's model can, I argue, support the phenomenological claim from James (above) that there are feelings which do not have an obvious bodily expression, but which infuse our psychological lives with meaning. If affective value is integrated so early on in visual processing then this gives credence to the claim that Ratcliffe draws out of James that emotions "partially constitute the world of experience whose contents form the material for our deliberations" (Ratcliffe, 2005, p. 187). These deliberations are then not distinct from this affect but are partially constituted by affect as well. This is made clear in a passage from The *Varieties of Religious Experience* (James, 1902), cited by Ratcliffe (2005, p. 188):

> Conceive yourself, if possible, suddenly stripped of all the emotion with which your world now inspires you, and try to imagine it as it exists, purely by itself, without your favourable or unfavourable, hopeful or apprehensive comment. It will be almost impossible for you to realize such a condition of negativity and deadness. No one portion of the universe would then have importance beyond another; and the

> whole collection of things and series of its events would be without significance, character, expression, or perspective. Whatever of value, interest, or meaning our respective worlds may appear embued with are thus pure gifts of the spectator's mind. (1902: p. 150)

The emotion that the world is stripped of in the above example is not an emotion such as 'fear' or 'anger' or 'happiness'. It is the feelings referred to in one of James' passages cited earlier, feelings which James did not consider to be correlated with obvious bodily changes (and therefore not emotion). We might think of these feelings in terms of not being the object of attention but rather part of the attention itself, and because they are not the object of experience they are transparent; they allow us to experience an object through them. This kind of transparency does not equate to invisibility; coloured lenses are transparent and yet at the same time colour our experience of the world. The colour is transparent, however, in that it need not remain in the forefront of our experience. While we initially notice the pink tint of the perceived world when we don rose-coloured glasses, this gradually fades to the background, shaping what we see but not itself being the object of experience. Likewise, as (Legrand, 2007) discusses, we can have bodily experience of our hand in two ways: either as object, as when it is touched, or as subject, as when it is doing the touching. Just as we can direct our attention to the colour of the lenses of the tinted glasses we are wearing, or to the hand that is being touched rather than doing the touching, it is possible to take one's affective state as the object of that awareness, and this is when James would consider it an 'emotion'. But Barrett and Bar argue that being the object of our experience is not the usual role for affect. The integration of perception and affect means that when affect is in the background we perceive it as a part of the world rather than our reaction to the world:

> 'Unconscious affect' (as it is called) is why a drink tastes delicious or is unappetizing (e.g. Berridge & Winkielman 2003; Winkielman et al. 2005; Koenigs & Tranel 2008), why we experience some people as nice and others as mean (Li et al. 2007) and why some paintings are beautiful while others are ugly (for a discussion of affect and preferences, see Clore & Storbeck 2006). (Barrett & Bar, 2009, p. 1328).

Affect, like the rose-coloured lenses, is not merely an addition to the perception, but constrains the perception. Just as there are wavelengths that cannot breach the rosy barrier of the lenses, the body's affective state constrains what is available to be perceived by coding for relevance and value and thus shaping attention and how the gist of a situation is filled in.

Ratcliffe then makes the step from the claim that *emotions structure our perceptual phenomenology* to the claim that "they are not distinct from cognition but constituents of

cognition" (Ratcliffe, 2005, pp. 185-186). Here Ratcliffe is proposing that this is how we should read James, but given his own work on moods and existential feelings (Ratcliffe, 2009, 2010) we should take Ratcliffe to also support this claim. The argument runs as follows.

Affects play a role in structuring the experiential world that forms the context for our various deliberations (Ratcliffe, 2005, p. 186). This follows from James' practical conception of intentionality according to which "(a) emotions are a constituent of intentionality and (b) experience does not merely 'present' but also partially constitutes or 'makes' its object" (Ratcliffe, 2005, p. 190). This means that the claim is not merely that emotions – as a constituent of intentionality – are a background, or enabling condition for cognition, but that cognition (as the object of experience) is also partially constituted by emotion. This is because, on the Ratcliffe-Jamesian view, the object of experience is partially created through the intentionality. This "world-making" is the practical orientation (or "attunement") of an organism to the world in virtue of its organisation (and thus its needs and abilities)[27]. Ratcliffe talks of this in terms of intentionality because in the phenomenological tradition intentionality is not just the "aboutness" of a mental state – or in other words being directed toward the object of that mental state – but rather it is "conceptualized in practical terms, as an orientation that does not merely reveal but also differently configures the experienced world". To make this concept more intuitive we can ground it in the more biological terms of an animals "Umwelt" (Von Uexkull, 1934) which literally means 'environment'. An animal's umwelt is just that part of the world that it can act in, given its access to the world via its senses, or in other words, it is "the set of environmental features to which a given type of animal is sensitized" (Clark, 1997, p. 24).[28]

There may seem to be a tension here between, on the one hand, a part of an objective environment being revealed to one in virtue of one's senses and, on the other hand, one's world being a subjective construct. The tension, however, is illusory; there is an external world and we have access to only the parts of that world that are made available to us through our senses. What the phenomenological approach brings out – that the more biological approach may leave implicit – is that the senses do not make parts of that world

[27] It is, I think, gesturing towards the same underlying concept as Gibson's "affordances" and the enactivist conception of cognition as 'laying down a path in walking' (Thompson, 2007; Varela, Rosch, & Thompson, 1991).
[28] The paradigmatic (and often cited) example of this is in regard to the tick, a creature which has only 3 receptors and three effectors, and whose world (umwelt) is thus very small/simple. For further discussion see (Clark, 1997, pp. 24-25; Mandik & Clark, 2002, Von Uexkull, 1934, pp. 10-11).

available to us directly "as is" (i.e. in full) but rather the world is translated through our particular sensory mechanisms and possibilities for interaction such that that experience is structured by these in a way we cannot eliminate. Of course that they are transcendental preconditions does not mean that they are logically necessary. Different creatures have different sensory apparatus, and so we might assume also a different structure to their experience. But they are so inherent in our experience that I am not even sure that we can imagine swapping, let alone eliminating these structures. I won't discuss this in any more detail as there is much too much work on this – from Kant through the Phenomenological tradition of the 19th and 20th centuries – for me to discuss here; Mandik & Clark (2002), in particular, look at this question in detail, arguing that selective representing does not imply a threat to the realist conception of the world. Let it suffice for me to say that I do not see in this talk of intentionality and world-making any cause for concern in regard to lack of objectivity of the world; as I see it depending on our sensorimotor structures different parts of the world are made available to us and this shapes our particular experience of the world. And, given that affect is intimately bound with our sensory capacities, affect also shapes (1) how we experience the world – "our world", and (2) how we can act in that world.

It is the suggestion that if affect shapes how we experience the world then it also shapes how we can act in that world that is at the heart of Ratcliffe's Jamesian claim that affect is constitutive of cognition. In chapter five I will return to the issue of whether the work discussed in this chapter is sufficient to warrant the claim that affect is partially constitutive of cognition. I will now return to exploring the work on core-affect outlined in chapter three by Barrett and colleagues. Given the work expounded in this section so far, I argue that this gives us good reason to think that affect not only forms part of the structure of experience but also shapes cognition subsequent to – or indeed in the absence of – conscious perception, and thus influences cognition (in a manner which I will argue in chapter 5 amounts to being partially constitutive of cognition) and that this position meshes well with empirical findings.

1.4 Core affect as a basic psychological ingredient of mentality

As discussed in chapter three, Barrett & Bliss-Moreau (2009) (henceforth BBM) refer to the experience of affect as 'core-affect'. BBM explain how recent studies show that core-affect is not only a basic psychological ingredient in emotion, but in mentality more generally. That is to say there is evidence that core-affect is also a basic psychological ingredient in

psychological phenomena that fall outside the traditional boundaries of emotion. Their lab has been investigating the role of core-affect in learning and vision (BBM, 2009, p. 198).

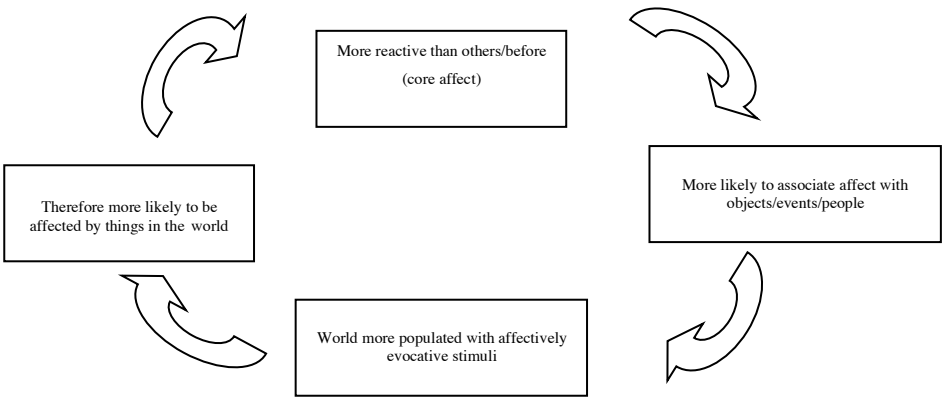
In regard to learning, core-affect supports the learning of what one should approach or avoid, desire, ignore and so on - the very basis of how we get by in the world. Given that there are only a few objects and situations which have the *intrinsic* power to perturb one's core affect (i.e., we are set up such that there are few perturbations of core affect that happen as a result of innate reactions to objects or stimuli), we must learn through associative affective learning[29]. Associative learning is where a neutral stimulus (a stimulus that does not yet have the power to perturb one's core affect) acquires affective value by being paired with affective stimuli (stimuli that changes one's core affective state). Such learning either takes the shape of Pavlovian/classical conditioning where the neutral stimulus is associated with stimuli that cause sympathetic nervous system responses (such stimuli include high pitched or loud noises and electric shocks), or evaluative conditioning. Evaluative conditioning is where a neutral stimulus is associated with stimuli which are explicitly evaluated by the individual to be liked/good or disliked/bad - that is, there is no direct core-affective response to the stimuli but the association is nevertheless made between the stimulus and those types of things which cause a core-affective response (BBM, 2009, p. 198). Through being associated with neutral stimuli in this way, affective responses pervade our experience.

Previous research (as briefly rehearsed in chapter three) has shown that people differ in their affective reactivity (with those who are more valence-focused reporting greater affective reactivity than those lower in valence focus or those who are arousal-focused). BBM show that this individual difference in affective reactivity supports individual variation in affective learning, predicting the magnitude of the learning. Those who are more perceptually sensitive to affective value (valence-focused individuals) demonstrate enhanced affective learning, and it can be shown that as perceptual sensitivity increases so does the magnitude of the EDA response (the electrodermal response on the fingertips which is used to measure the activation of the sympathetic nervous system) to the conditioned stimulus (CS+). These effects are enhanced for those individuals who score highly on neuroticism, a condition which correlates with those who are both highly valence-focused and tend to have high sensitivity to negative cues in particular.

[29] See Montague (2006) for one account of how dopamine and reward-prediction error signals might enable a system to "learn associations between experienced signals and temporally removed (but consistently associated) rewards" (Clark 2007, Review of Montague 2006).

In addition to being related to variation in negative affective learning, individual differences in core-affective reactivity also predict better rule-based affective learning, where the value of the object is communicated explicitly through symbolic means (for example being told by someone about danger, or pain etc.), and where this value was positive this learning was enhanced for those who score highly on extraversion. B&BM thus conclude that the findings "suggest that both associative and rule-based affective learning are enhanced for people whose core-affective states are often and easily perturbed" (p. 201).

The result of this is that "simple temperamental differences" in affective reactivity, affecting the degree of affective learning, can develop into "very different emotional lives" (p.202) in virtue of engaging a positive feedback loop which feeds more reactivity and more affectively evocative stimuli back on each other. What starts as an individual being more reactive to stimuli such that her core-affect is more easily and intensely perturbed, than that of others resulting in her being more likely to associate affect with objects, events and people. When she has associated these with affective states, they become affective stimuli themselves and so her world is more full of affective stimuli than someone who is less sensitive, and given that her world is more full of affective stimuli she is more likely to have affective responses to things in her world, which in turn means that her core-affect is being perturbed more often and more intensely than others and so the cycle starts again.



While it may seem unsurprising that affective association underpins affective learning what we should draw from this discussion is how affect structures what we then come to know or learn about the world, in just the sort of way that the Ratcliffe-Jamesian view discussed above, suggests, thus infusing subsequent experience with that affective value. This can be seen where BBM argue that core-affect is also a fundamental feature of conscious

experience. They say: "Neuroanatomical evidence strongly suggests that core affect provides a source of attention in the human brain (where attention is defined as anything that increases or decreases the firing of a neuron)" (BBM, 2009, p.202). Core-affect has an important role to play in perceptual functioning because if there is sufficient influence on the internal body state due to core-affect being perturbed by sensory information, then the processing of sensory information is prioritised such that the sensory stimuli which caused the perturbation is more easily seen and remembered. BBM suggest that "[p]ut another way, "feeling" and "seeing" (or "hearing" or "smelling" and so on) may not be all that independent of each other" (*ibid*).

B&BM present evidence that core-affect influences sensory processing throughout the brain through two routes; direct and indirect. The direct route is as below (I quote at length directly from B&BM as this is a clear and concise summary):

> "Parts of the neural reference space for core affect (such as the amygdala and lateral OFC) project directly to all sensory cortices and so can directly influence sensory processing. For example, the basal nucleus of the amygdala projects directly to all portions of the visual ventral stream, serving to modulate neural activity from the association cortex all the way back to the primary visual cortex (or V1) (for a review, see Duncan & Barrett, 2007). The sensory integration network in the central and lateral OFC projects to the visual association areas in the inferior temporal lobe (part of the ''what'' or ventral visual stream for object recognition) and the visceromotor network in the medial OFC projects to the visual association areas in the inferior parietal lobe (part of the ''where'' or dorsal visual stream for spatial localization and action preparation) (for a review, see Barrett & Bar, in press). The circuitry that realizes core affect also project indirectly to sensory neurons via three different routes. The amygdala, the visceromotor network of the OFC (including what is sometimes called the medial OFC or vmPFC), and the ventral striatum project to the ascending arousal systems the brainstem and basal forebrain (for a review, see Edelman & Tononi, 2000;Mesulam, 2000; Semba, 2000) that have diffuse, unidirectional afferent projections through- out the cortical mantle, acting as a ''leaky garden hose'' (Edelman, 2004, p. 25) to control the level of neuronal firing throughout the brain. (In fact, affective circuitry offers the only path by which sensory information from the outside world reaches the brainstem and basal forebrain; Mesulam, 2000). The amygdala and OFC (as well as the brainstem and forebrain nuclei) also project to certain thalamic nuclei that regulate the transmission of sensory information to the cortex and are partly responsible for forming and selecting the groups of neurons that fire in synchrony (called neuronal assemblies) to form conscious percepts (the things people are aware of seeing) (Zikopoulos &Barbas, 2007; for a review, see Duncan & Barrett, 2007)" (BBM p. 202-203).

The indirect route comes about as a result of the lateral pre-frontal cortex exercising top down/endogenous attention which (1) constrains the on-going processing throughout the cortex, and (2) helps to select the information that reaches conscious awareness by "directing

the formation and maintenance of the neuronal assemblies that underlie conscious experience" (p. 203). In particular viewing affectively potent stimuli results in enhanced neural activity in sensory areas including the visual cortex, and studies of negative core-affect have shown that this activity in the visual cortex is greater in states which are high in arousal, such as fear and anxiety (vs. those lower in arousal). This has the consequence that affect structures our perceptual phenomenology:

> "The pattern of projections from the neural reference space for core-affect to visual cortex suggests the intriguing hypothesis that what people literally see in the world around them may in part be determined by their core-affective state." (B&BM, 2009, p. 204)
>
> "… this research will inform an ongoing debate over the distinctiveness between affect and cognition suggesting that the distinction may not be an ontological distinction that is respected by the brain (cf. Duncan & Barrett, 2007). The most far reaching implication of this work is that "thinking" (e.g., sensing and categorizing or deliberating on an object) might not be a fundamentally different sort of psychological activity than "affecting" (i.e., constructing a state to represent how the object affects you" (BBM, 2009, p. 205).

Thus we see that by structuring our perception, affect also structures our capacity for cognition by determining what may be processed and by giving those processes relevance in terms of biological value. This amounts to an interesting causal claim in and of itself, in chapter five however, I will make the case that in this domain the distinction between casual and constitutive processes is so difficult to demarcate that affect is plausibly considered partially constitutive of cognition

To summarise this section, I have outlined two ways in which affect influences (and may be plausibly thought to be partially constitutive of) cognition:

(1) Affect structures our perceptual phenomenology & therefore is a precondition for other cognitive processes.

(2) Affect shapes our cognitive processes; these arise not just in a background of affect but are partially determined by affective state.

One might accept (1), that our senses (and therefore our conscious experience) are never wholly free from affect, but nevertheless argue that affect is a background or enabling condition for cognition, rather than being partially constitutive of cognition. On such a view affect would play a role in cognition similar to that of oxygen in the orthodox view of human

consciousness: oxygen is essential to keep us alive, which is necessary for human consciousness, but it serves only a supporting function – it is not part of the mechanism for consciousness. That is to say, if there were some other way of keeping the system alive that did not involve oxygen, the claim would be that whatever the mechanisms of consciousness are in humans, they would not be affected or changed in any way. Rather like how changing a diesel engine for a petrol engine would not affect how the gears or steering works in a car. In chapter 5 I will argue that affect should not be thought of as a background condition but as plausibly (partially) constitutive of cognition.

Claim (2) goes even further than claim (1), asserting that affect determines what – and how - cognitive processes occur. Taking the car analogy a bit further, this is analogous to a car having different possibilities for driving if it is running on petrol, as opposed to running on diesel; the fuel does not just enable driving but it also constrains what kind of driving can be done. Thus, taken together the picture I am supporting is that affect is not only playing a role in the structure of our perception and perceptual experience, but that it also constrains what is available to be cognized over and the way in which it is cognized over.

Is this sufficient for the claim that affect is constitutive of cognition? This depends on how you understand 'cognition' (and also on how you understand 'constitution' but more on this in chapter five). From the perspective of Phenomenology or Enactive Cognitive Science, I suspect that this is sufficient, because cognition for them comes down to our possibilities for acting in the world. If affect is constitutive of perception and a precondition for other cognitive processes then it feeds and constrains these possibilities for acting in the world, infusing them with affective value. Researchers working within the framework of a more traditional approach to the cognitive sciences or philosophy of mind, may however object that this does not show that affect is involved in the actual cognitive processes. This therefore would not necessarily change their models of cognition at the neural or computational levels. For this, we need evidence that cognitive processes such as attention or executive decision making, or other paradigmatic processes, are also infused or intertwined with affect such that it should be included in such models. In the next section I will argue that modern neuroscience shows that this is indeed the case.

Section 2: The role of affect in cognitive processing

As I discussed at the beginning of chapter three, 'cognition' is an ambiguous term in the cognitive sciences. So far I have predominantly been concerned with general cognition; whatever grounds our psychological capacities. In this section I will be a little more precise, and consider particular cognitive processes such as attention and executive function. A review of recent neuroscience shows that affect is involved in these cognitive processes. And, conversely that structures that have previously been labelled as 'affective' because of their apparent role in emotional processing (1) perform in ways very similar to 'cognitive' structures, and (2) are not only involved in 'emotional' processing. As we will see, this begins to undermine the traditional 'one area-one function' framework which up until very recently has been (and arguably remains) implicit in most cognitive science in favour of an explicitly network based framework. Such a framework will still be able to enable a distinction between emotion and cognition in terms of which networks underpin which behaviours (which we may want to categorise as cognitive or emotional) but does not acknowledge this distinction in virtue of anatomy or connectivity because the nature of the networks entails that 'affective' information is integrated into many (plausibly all) 'cognitive' processes.

2.1 'Affective' structures do 'cognitive' work

Pessoa (2008) argues that three of the motivations for distinguishing 'emotional' from 'cognitive' brain structures are that (1) emotional structures were thought not to be involved in cognitive processes such as attention; (2) emotional structures were thought to be independent of top-down factors; and (3) emotional processing was context independent. Pessoa reviews evidence that recent neuroscience shows that this is not the case. The amygdala is the paradigmatic 'emotional/affective[30]' structure. It is thought of as the "fear centre". It has been shown to activate in experiments on fear conditioning in rats; there to be a deficit in recognition of fearful expressions in patients with bilateral amygdala lesions; and to respond to fearful faces in neuroimaging studies (Pessoa, p. 149). However, Pessoa explains that processing in the amygdala has been shown to (1) be involved in a type of attention; (2) not be independent of top down factors; and (3) to not be context independent.

[30] Pessoa explicitly uses 'emotion' and 'affect' interchangeably.

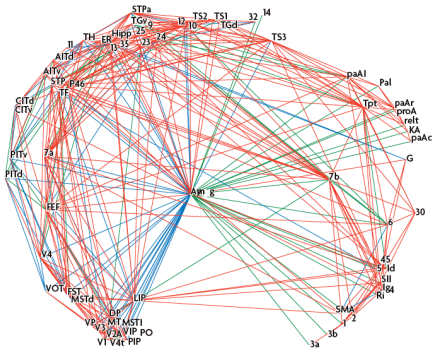
Attention is a paradigmatically 'cognitive' function. A central function of attention is to modulate sensory processing: attention to a stimulus increases neuronal firing rates and FMRI responses in the sensory cortex. But neuroimaging studies also show that amygdala activation is correlated with activation in the visual cortex (just as happens in traditional attention tasks). A number of other studies show that even when affectively significant stimuli mediated by the amygdala (such as faces or colours previously paired with a mild shock) are task irrelevant, activation in the amygdala increases detection performance and this is paralleled by increases in activation of the visual cortex. Pessoa argues that what this means is that something that seems to parallel traditional attention is going on in emotional processing and processing in the amygdala underlies this.

Pessoa argues that recent studies also show that processing in the amygdala is not independent of top-down factors. Although amygdala responses are observed in conditions of inattention and minimal sensory input they are also strongly dependent on attention (even for affectively significant stimuli). That is, attention modulates the amygdala so that increased attention results in an increased amygdala response. Amygdala responses have also been shown to be closely linked to perception so that when briefly presented with a fearful face the amygdala response differs depending on whether the subject reports perceiving the face.

Although I just want to focus here on the amygdala, for the sake of a full picture we should note that the converse also holds: a paradigmatically cognitive region – the Dorsolateral Prefrontal cortex (DLPFC) is also involved in integrating cognitive and emotional information. This was shown in a study in which response inhibition to negative words such as 'worthless' engaged the DLPFC even though non-negatively valenced inhibition tasks and tasks that involved negative-valence but no response inhibition did not (Pessoa, p. 150). Pessoa also cites a working memory study in which subjects were asked to keep in mind neutral and emotional pictures. The maintenance-related activity in DLPFC was modulated by the valence of the picture so that (relative to neutral pictures) pleasant pictures enhanced activity and unpleasant pictures decreased activity. Whereas when subjects were not asked to remember the pictures the emotional pictures did not affect the DLPFC response *(Ibid)*. Because the responses were relative to the task of remembering neutral pictures this result cannot be accounted for by appealing to the activation of working memory. We should conclude therefore that it is probably a mistake to distinguish anatomical components in virtue of their being 'emotional' or 'cognitive'.

2.2 Connectivity between 'cognitive' and 'affective' structures

Even though it doesn't look like we can anatomically distinguish emotional and cognitive areas, could there be an argument for distinguishing them in virtue of their connectivity? Pessoa argues that there is not. His argument runs as follows. The brain has a 'small world' topology, such that cortical areas are connected either directly or by just one or two intermediate areas (Sporns et al., 2004; Sporns & Zwi, 2004; Bullmore & Sporns, 2009). The prefrontal areas are among the most distant areas from the sensory periphery and so receive highly processed and integrated information. Highly processed information is supposed to bring greater flexibility and to support the abstract processing required for 'cognition'. However, the amygdala is connectively equally removed from the sensory periphery and so receives just as highly processed and integrated information. Moreover, the amygdala makes (and receives) widespread projections to the rest of the brain. It is one of the most highly connected regions of the brain – it is like the hub of a wheel.



"Quantitative analysis of the connectivity of the brain by Young et al." (reproduced in Pessoa 2008)

So if viewed from a perspective of connectivity, the amygdala is in the "geometric centre of the topological map" and even though it is a core affective region, it is "at least as well situated to integrate and distribute information as certain PFC territories." (Pessoa, p. 151). Other regions often thought to be purely affective such as the hypothalamus, OFC/VMPFC, ACC, also seem to be important connectivity hubs (p. 152).

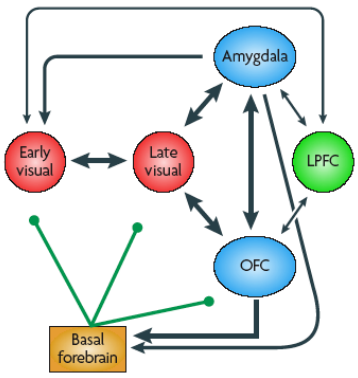
Why should a region's connectivity affect whether we think of it as emotional or cognitive? Because the more connected it is the more information it receives and integrates from areas traditionally conceived as 'affective' and 'cognitive'. And, the more it projects to both regions as well. If a region really were 'cognitive' or 'affective' we would expect it to

predominantly be connected to other regions of the same type and only minimally being connected to other structures because on a view where emotion merely influences cognition, emotional and cognitive processing would be highly modular (though interacting). However, the connectivity data shows that this is not the case. We should conclude therefore that the connectivity data suggests that we cannot distinguish 'emotional' and 'cognitive' regions in the brain according to the traditional criteria.

2.3 Integration of 'affective' processing at multiple stages

We have seen that our paradigmatic example of an emotion area; the amygdala, is involved in attention – a paradigmatically cognitive process – and that it is a connectivity hub, integrating information from - and projecting to - almost all the areas in the brain; both those traditionally conceived of as cognitive and as affective. Perhaps if regions in the brain cannot be distinguished as emotive or cognitive in virtue of their anatomy or connectivity there is at least a functional distinction. That is, perhaps circuits through regions traditionally conceived as 'cognitive' underlie a function such as sensory processing.

Pessoa reviews evidence from visual processing and executive control that suggests that this is not the case. In fact, 'cognitive' and 'emotional' contributions seem to be integrated at multiple processing stages such that they cannot be separated and the origin of any particular modulation is lost. In respect to visual processing, both the amygdala and PFC have reciprocal connections with visual sensory areas. And the DLPFC, which we saw integrates 'emotional' and 'cognitive' information, has reciprocal connections with early visual areas.

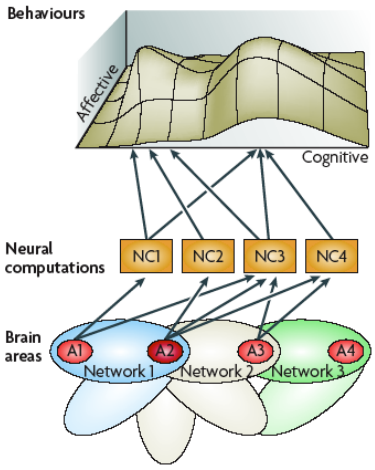


(Pessoa, 2008)

As can be seen from the diagram above from Pessoa 2008, there is top down processing from LPFC to the OFC, amygdala, and early visual cortex. The amygdala feeds into both

early and late visual areas and receives projections from late visual area. The amygdala and OFC both project to the basal forebrain which provides diffuse modulatory signals which enhance processing of contextually significant information in the early and late visual areas and OFC (p. 153). The moral of the story is thus that if there were a 'cognitive' or 'affective' origin of any modulation, it is lost in the processing due to the highly recurrent nature of neural processing.

Executive control is perhaps the most 'cognitive' of functions. Pessoa suggests that evidence shows that rather than executive control being underpinned by just the ACC and LPFC (regions traditionally conceived of as cognitive), 'affective' regions such as the amygdala, PFC and nucleus accumbens are involved because control involves taking into account costs, benefits, and goals. This broader neural circuit means that strategies for action incorporate value. This larger circuit also includes dopaminergic projections to the frontal cortex so reward prediction and expectation (including reward-prediction errors) features "in the temporal unfolding of control" (p. 153). This evidence leads Pessoa to conclude that the 'one area-one function' framework is misguided. Rather, functions are realized by neural computations receiving information from networks of brain areas. Instead he proposes a many-to-many framework in which behaviours are both cognitive and affective; the axes are not orthogonal. Any change along one axis means a change along the other (p. 154).



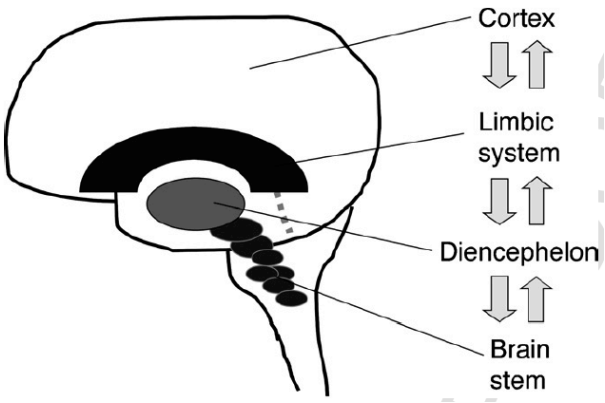
(from Pessoa 2008)

This work thus shows that while we may still be able to demarcate emotional and cognitive behaviours, it does not look as if the correct way to think about the work of the brain is in terms of 'affective' or 'cognitive' processing as the origin of any information becomes rapidly lost due to the highly recurrent nature of the brain. Moreover, given (1) that

structures in the brain that have traditionally been categorised as 'affective' do cognitive work and integrate as much information as those traditionally categorised as 'cognitive', and (2) that 'cognitive' structures are also involved in affective processing and integrate affective and cognitive information, we see that not only is affect integrated into 'cognitive' processing, but that this points towards giving up these labels at the level of neural processing.

2.4 Lewis and Todd's self-regulating brain

Lewis and Todd (2007) also think that a neural analysis overcomes the cognitive/emotional distinction. They argue that the folk-psychological distinction between emotional responses and cognitive interpretations can be mapped on to the vertical dimension of the 'neuroaxis'. The neuroaxis does not illustrate the anatomical positioning of brain structures but rather relates to the development of the neural tube. From the bottom (ventral) to the top (dorsal) the structures run as follows: brainstem; thalamus and hypothalamus; standard 'limbic' structures (amygdala; hippocampus) and basal ganglia; cerebral cortex, and culminates in the pre-frontal cortex. Each additional layer stretches out the time between stimulus and response.



The Neuroaxis, from Lewis and Todd (2007)

This division is different from the traditional one because (e.g.) the hippocampus would normally be considered to be cognitive because of its role in memory. Here there aren't emotional or cognitive regions – rather, ventral processing is more automatic (less processing between stimulus and response) and dorsal processing is more deliberate (more processing between stimulus and response). There exists upregulation (regulation starting from the ventral areas entraining the dorsal areas) and downregulation (regulation starting in

the dorsal areas and entraining ventral areas) but, critically, the ventral parts of the neuroaxis (the part that most corresponds to the folk-psychological category 'emotional') are always involved in processing; projecting to and receiving projections from structures higher up the neuroaxis. We all need a brain stem, whereas animals that have had their cortices removed can cope pretty well. The Anterior Cingulate Cortex (ACC) is the hub of the brain systems mediating effortful, deliberate and conscious self-regulation. There are different epicentres of self-regulation: epicentres can be lower in the neuroaxis and harness mainly subcortical areas (but which have connections with cortical areas) and thus support a predominantly unreflective drive-driven self.

Lewis and Todd focus mainly on two epicentres of self-regulation – one in the ACC reflecting deliberative, reflective, conscious behaviour, and the other in the limbic system, which consolidates habits of emotion appraisal and behaviour, reflecting more childish behaviour. The point is that whether the epicentre is limbic or cortical (ACC) processing is still emotional; still engages emotional structures (limbic structures) and neuromodulatory pathways. So although more reflective 'cognitive' processing activates areas higher up the neuroaxis this is coordinated with activity lower down the neuroaxis. Lewis & Todd (2005) summarise this nicely as follows:

> The brain stem and hypothalamus entrain limbic structures by means of neuromodulators and neuropeptides, locking in perceptual biases and associations, and they also recruit cortical activities in service of ancient mammalian and even reptilian agendas, which can be thought of as emotional *action tendencies*. Primitive agendas and requirements thus flow up the neuroaxis from its roots at the same time as executive attention, planning, and knowledge subordinate each lower level by the activities of the cortex. If not for the bottom-up flow, which can be seen as underlying motivated attention (*emotion as regulator*) the brain would have no energy and no direction for its activities. If not for the top-down flow, which underlies self-regulation processes (*emotion as regulated*), recently evolved mechanisms for perception, action and integration would have no control over bodily states and behaviour. It is the reciprocity of these upward and downward flows that links sophisticated cognitive processes with basic motivational mechanisms. (Lewis & Todd, 2005, p. 225)

Thinking in terms of 'emotion as regulator' thus gives us another way in which to understand how 'affective' information pervades processing, and can plausibly ground the phenomenological argument presented in section one - that affect pervades our perceptual phenomenology and mentality.

2.5 The amygdala and biological significance

Understanding the amygdala as a hub in the brain's networks, or as part of a ventral epicentre for behaviour, we can begin to see that it does indeed play an important role in emotional situations. But this role is not of the kind that we imagine when we are entrenched in the conception of the brain as a *one area - one function* machine. This framework resulted in interpreting data in such a way that the amygdala has generally become thought of as a "rapid-response fear module" (see Sander et al. 2003, for a detailed rejection of this view).

Now that we have a different framework in theoretical neuroscience in which to view data within, that of networks (Sporns, 2010; Sporns & Zwi, 2004) and prediction (Bar, 2009; Friston & Kiebel, 2009), we can see that there is evidence that the amygdala plays a far more important role then previously supposed. The role it plays is of coding for biological relevance (Sander et al. 2003). This can be understood in terms of the amygdala's function being to "direct the various sources of attention […] towards a source of sensory stimulation (such as an object) when the predictive value of that stimulation is unknown or uncertain" (Barrett et al., 2007). We can thus see that this model brings the amygdala into the kind of attention model proposed by Friston and Feldman, discussed in section 1.1. This attentional mechanism is affective in terms of predicting biological value and engaging physiological changes and physical actions which allow the system to better predict value (i.e. reduce prediction errors).

The literature which shows that the amygdala is involved in emotion processing does not necessarily imply either that the amygdala is an emotion structure or that it's activation pertains to emotional rather than non-emotional stimuli. Rather, conceived of as coding for significance, previous findings can be accounted for, while also accounting for the important role it plays in non-emotional processing (that is processing that would not traditionally be considered to be 'emotional'). Similarly (Pessoa & Adolphs, 2010) argue that the amygdala is not an emotion module but a core brain circuit with "broad connectivity with the cortex and other subcortical structures" which enables it to play a modulatory role in multiple networks:

> The precise functional importance of the amygdala in these networks remains to be investigated, but it is unlikely that it will map specifically onto emotion. Instead, we think that it corresponds to broader and more abstract dimensions of information processing, including processing of salience, significance, ambiguity, unpredictability and other aspects of 'biological value'. More broadly, we argue that the amygdala has a key role in solving the following problem: how can a limited

> capacity information processing system that receives a constant stream of diverse inputs selectively process those inputs that are the most relevant to the goals of the animal? (Pessoa & Adolphs, 2010, p. 780)

There is much more work that needs to be done here to show how this new conception of the activity of the amygdala fits in with the emerging theoretical neuroscience frameworks. But for our purposes the importance of this re-evaluation of the amygdala's role in processing should be clear. If activity in the amygdala is part of a network which codes for biological significance, and it is a hub projecting to, and receiving projections from, most areas in the brain, including those previously deemed to be 'cognitive', it is going to be very difficult to continue to hold the coarse distinction between either affect/emotion and cognition in terms of neural processing.

2.6 Summary

In this chapter I discussed evidence for the hypothesis that affect, in the form of afferent and efferent interoceptive information, is involved in perception. I discussed how the generalised predictive coding framework in theoretical neuroscience suggests that biological relevance is coded for in terms of prediction errors and I suggested that one aspect of Wundt's affect, strain-relaxation, could be grounded in the force with which errors are resolved. If affect is intertwined in perception then it seems reasonable to suppose that it structures our perceptual phenomenology. I reviewed work by James and Ratcliffe and argued that this is in fact the case. I argued that this in turn implies that core-affect is a basic ingredient of mentality and discussed recent work in neuroscience which shows that there is no grounds, in terms of structure, function or connectivity, for distinguishing emotion from cognition in the brain. Even if we are to distinguish epicenters of self-regulation as 'emotional' and 'cognitive', emotion still pervades cognitive processing and given the recent reconceptualization of the role of the amygdala in neural processing as coding for significance rather than 'fear' we can conclude that (1) the emotion-cognition distinction is no longer useful in regard to neural processing, and (2) affective information and biological significance are entwined in cognitive processing. In the next chapter I will discuss whether or not this is sufficient to ground the stronger claim that affect is partially constitutive of cognition.

Chapter V

Exploring the relation between affect and cognition

What kind of claim should I be making about the relation between affect and emotion given the story I have told in the previous two chapters? So far I have argued that: (1) it is useful to distinguish affect from emotion; (2) affect is grounded in interoception; (3) affect is involved in vision; (4) affect structures our perceptual phenomenology; (5) affect is a basic ingredient of mentality and shapes our cognitive processes; (6) affect is involved in cognitive processing (attention and executive control). A minimal claim is that affect is involved in cognition at various levels. This is interesting in and of itself, but is there a stronger claim that ought to be being made? In the previous chapter I talked in terms of affect influencing, integrating into, and being constitutive of cognition. While I was talking in loose terms there, here I want to tighten up my analysis and consider what such a claim amounts to, and how we should think about the relation between affect and cognition.

1. The causal-constitutive distinction

One claim that could be made is that together claims (3), (4), (5), and (6) above show that affect is partially constitutive of cognitive processes. Typically when we talk of constitution it is in terms of being opposed to 'merely causal' accounts of whatever is in question. There are two ways of thinking about constitution: In respect to constitutive conditions which are necessary for a thing to be the thing it is, (e.g.) it is a necessary condition for being a footprint that it was made by a foot. Alternatively constitutive factors and processes are thought to make up the core of the thing in question; and causal processes (often in the form of background and enabling conditions) provide the required context for the constitutive processes to occur. This latter conception of constitution seems to be the more relevant for cognitive science and so I will focus on this. On such an account of constitution one might claim that respiration is an enabling condition for neural functioning as neural functioning requires oxygen (as all eukaryotic cells do) but oxygen is not constitutive of neural functioning as: (1) so long as the oxygen could be replaced by something else that would keep the cells alive and well then neural functioning should not be affected in any way, and (2) the role that oxygen is playing is not at the right level to be of interest.

Really, (2) is just another way of putting the same assumption that is the foundation for (1): it is also just asserting once more that one has made a causal-constitutive distinction already and segregating phenomena in virtue of that. This might be thought of as a case of the ''causal-constitutive error' error' which Hurley (2010) discusses in regard to the internalism/externalism debate in the philosophy of cognitive science. The 'causal-constitutive error' is to take an explanation as constitutive when really it is only causal (Block 2005, cited by Hurley 2010, p. 106). For example, someone claiming that oxygen is partially constitutive of neural processes, would be derided by many philosophers of mind and science as making a causal-constitutive error. Hurley argues that – in regard to internalist and externalist explanations of cognition – those who assert that their opponents are falling foul of this error are in actual fact the ones committing the error. The error Hurley takes these deriders to be making is:

> … the error of objecting that externalist explanations give a constitutive role to external factors that are "merely causal" while assuming without independent argument or criteria that the causal-constitutive distinction coincides with some external-internal boundary. (Hurley, 2010, p. 106)

Essentially Hurley is accusing the internalists in question (principally Adams & Aizawa) of making a thinly veiled circular argument of the following form:

(1) Assumption: There is a casual-constitutive distinction
(2) If x is external then x is casual and not constitutive
(3) The processes the externalists appeal to are external
(4) Therefore: The processes the externalists appeal to are causal not constitutive

This is grounded in their assumptions that the capacities that she is giving an account of can be explained in terms of processes that are internal.

In regard to the internalism-externalism debate this means that the deriders have already decided that cognitive processes only take place in the brain and that anything outwith the nervous system is just playing a causal role. They thus argue, on the basis of this prejudice, that external factors are clearly causal and not constitutive. If this is how the position should be read then it is indeed a funny way for trained philosophers to argue, and I don't want to get into the ins and outs of their thinking here. Rather I want to use Hurley's rebuke as an example of how the very distinction of causal-constitutive has embedded within it the prejudice that things can be separated into 'that bit which is the real-deal' and 'stuff which interacts with it (or bits of it), and which might be necessary for it to exist in the first place but isn't really a part of *it*'.

Hurley focuses on the explanatory question rather than the metaphysical question. It is not always clear that explanation and metaphysics can be clearly delineated however. Is my question about whether affect is partially constitutive of cognition, metaphysical or explanatory? I tend to think more along explanatory lines, yet there is something about the way we talk about 'cognition' or 'affect' as if they are objects, or have some sort of essences, that encourages metaphysical leanings. It might be thought that the questions regarding the metaphysics or ontology of cognition and affect are perfectly reasonable but considering an argument from Ross & Ladyman (2010) gives us reason to think otherwise. They address the "alleged coupling-constitution fallacy" by arguing that metaphysicists of mind are playing with metaphors and thus cannot help us in discovering the boundaries of cognition. More specifically, they argue that "the metaphysical notion of constitution or composition is an abstraction that does not correspond to any general idea that figures non-metaphorically in science" and that "the notion of causation, insofar as it is relevant to science may not be applicable to fundamental physics, which casts strong doubt on its appropriateness as an explanatory element in any set of restrictions on unification of models" (Ross & Ladyman 2010, p. 157).

Why is it that, even when we know this, we are still drawn to the metaphysical notions of cause and constitution? Ross and Ladyman appeal to work by Lakoff and Johnson (1980) and Lakoff (1987) that shows that our language has inherent in it an implicit metaphysics - the doctrine of containment – on which:

> The world is a kind of container bearing objects that change location and properties over time. These objects cause things to happen by interacting directly with one another. Prototypically, they move each other about by banging into one another. At least as important to the general picture, they themselves are containers in turn, and their properties and causal dispositions are to be explained by the properties and dispositions of the objects they contain (and which are often taken to entirely comprise them). (Ross & Ladyman, 2010, p. 150)

The case that is relevant to us here – the relationship between affect and cognition – strikes at slightly different intuitions than the internalism-externalism debate in the philosophy of mind. But some are the same. While in that debate the intuitions are about the containers of the body or the brain, here also some might suppose that affect happens in the physiological body and is therefore not cognitive, while what happens in the brain – even if it is causally influenced by processes in the body proper – may be cognitive (this is a necessity claim not

a sufficiency claim)[31]. And so, the container metaphor is at work here – with the brain container containing that which is (or has the possibility of being) constitutive of cognition with everything outside it merely casual. Alternatively our intuitions may lead us to suppose that neural spiking is the container of the cognitive process while neurotransmitters, neuromodulators and glial cells just play a causal or modulatory role, rather than being constitutive of the cognitive process.

It seems natural to us to think of background or enabling conditions as separable from the explanatory target; as outwith the container of 'that which is really at issue'. Ross and Ladyman suggest that "the ubiquity of the containment metaphor derives from the fact that as tool builders we humans are naturally interested in isolating systems in such a way that we can transport them around without significantly changing the kinds of processes we can use them to effect (Cartwright 1989)" (Ross & Ladyman, p. 161-2). This way of putting it makes clear that transportability does not mean that the background conditions are not part of the process, but rather that they are easily available. Think of a fire: in oxygen rich environments we can carry around a fire with us wherever we go, so it seems that the oxygen is a background condition. But this is just a function of the happenstance that we live in an oxygen rich environment – it says nothing about whether oxygen is constitutive of combustion or merely a background condition. Compare to affect. There is a continuous stream of affective information at bodily, neural and phenomenological levels. Why should we think that this means that it is merely a background condition for cognition? Because we could plausibly keep everything else the same (or replace it) and separate out the processes that underpin, say, some types of memory in a sort of brain-in-vat in microcosm scenario? But, if everything is kept the same, that will not tell us any more about whether those processes constitute that memory process than moving the fire to another room will tell us about whether oxygen is partially constitutive of combustion.

The casual constitutive distinction does not seem to be a very helpful one in scientific explanation unless it is relative to an explanatory project. As Ross and Ladyman (2010, p. 163) point out, in the mature special sciences one can make models of systems, but these are not constitutive models in the terms of our naïve metaphysics; the relations are model-relative and variables are treated as endogenous or exogenous relative to a predictive or

[31] Note that the intuitions must not simply reduce to whether neurons are the substrate of cognition. Under this kind of objection neurons while necessary for supporting cognition are not sufficient, otherwise they would have no claim for keeping cognitive processing in the brain and spinal cord; neurons abound in the body proper not only in the peripheral nervous system but populations of neurons capable of information processing are present in the heart and enteric nervous system (gut).

explanatory purpose. Likewise, if it suits our purposes, we can background information about affect (and other cognitive processes) if we want to bring a certain type of memory to the fore, but in doing so one is not making any claims about the necessity or non-necessity (or constitutiveness or non-constitutiveness) of the processes that have been backgrounded for the foregrounded process. And the possibility of foregrounding certain processes does not imply that they are unpluggable from the background purposes.

2. Explanatory separability and difference/contribution phenomena

The confusion may arise because of a conflation of the method of controlled experimentation with claims of constitution. In controlled experimentation one tries to hold certain factors constant and manipulate others to see if a particular effect varies. But, as Hurley points out, this method seeks factors that are explanatorily separable and explanatory separability should not be conflated with a causal-constitutive claim. Hurley explains explanatory separability as follows:

> If the A and B factors are explanatorily separable, then either the contribution made by A factor to explaining X is independent of the level of or relations among B factors, or vice versa. But if the contribution of A factors to explaining X depends on the level of or relations among B factors, and the contribution of B factors to explaining X also depends on the level of or relations among A factors, then A and B factors are not explanatorily separable. In coupled dynamic systems, for example, the parameters of one system are the variables of the other system, and vice versa. Or, consider the nonseparability of bodily phenotype and extended phenotype in explaining the presence of a certain genotype (Dawkins 1982). If X depends not just on the factors that are varied but also on the levels of and relations among the factors that are "controlled" whichever way around we allocate variation and control to the A and B factors, then explanatory separability fails. Explanatory separability also fails if the A and B factors vary together in the relevant possible worlds, so that the factors in one set cannot hold constant while the others vary and their contributions to explaining X thus cannot be separated.
>
> Perhaps we can reindividuate sets of potential explanatory factors, so that they are explanatorily separable. But perhaps not. X may depend nonseparably on all the potentially explanatory factors and relations among them; they may be interdependent so as to form an explanatory unit. (Hurley, 2010, p. 108-109)

So is oxygen explanatorily separable from neural processing? No, but as Ross and Ladyman suggest we can background it by – for the sake of a particular explanatory model – treating it as an exogenous variable while treating variables such as spike frequency or neurotransmitter type as endogenous. Explanatory inseparability does not, however, mean that one cannot measure the difference various processes make to the explanatory target. Elliot Sober (1988) provides a way of thinking about causes that can help us to distinguish

types of causal contribution. He compares the casual contribution of gravitational force and electrical force on the acceleration of a particle to the way genes and environment contribute to a trait such as height. Sober argues that there are two questions that can be asked about casual contribution in the case of the particle: (1) what *contribution* did gravity (or electricity) make to the particle's acceleration? And, (2) what *difference* did gravity (or electricity) make in the particle's acceleration?

The contribution can be addressed by the following counterfactuals: (a) how much acceleration would there have been if the gravitational force had acted, but the electrical force had been absent? (b) How much acceleration would there have been if the electrical force had acted, but the gravitational force had been absent? By answering these questions one can work out whether gravity or electricity contributed at all to the acceleration of the particle, and if they did how much of the acceleration was due to their influence.

However, Sober argues, these questions become unintelligible if we transfer them to the interaction between genes and environment. Clearly when looking at the height of an individual we cannot ask ourselves how tall the person would have been if genes had not played any role, or conversely if the person had no environment. Height is a function of both genes and environment. And, thus to understand the contribution that each makes we must think in terms of holding one dimension steady and manipulating the other. In order to do this, biologists look at populations rather than individuals and use ANOVA (analysis of variance) statistical techniques. One can then have a population of genetically identical plants and vary the environment. Any differences in the height of the plant will then be due to environment rather than genetics. Or, one can have a population of genetically different plants in an identical environment and differences in height will then be a function of the genes rather than the environment. This method thus only gives you the *difference* of contribution rather than how much environment contributes and how much genes contribute to height. What this means is that the causal contribution of genes or environment can't be measured. While electrical and gravitational forces are commensurable (measured in the same terms) and local (can be manipulated in a particular case) genes and environment are not commensurable and cannot be manipulated to show the difference in an individual, only at the population level.

Sober gives a nice example (credited to Peter Woodruff) of a cannon firing:

> Consider the fact that the distance travelled by a projectile shot from a cannon is influenced by both the muzzle velocity and the angle at which the gun is set. Suppose we fire a cannon and the shot goes half a mile. There is no saying how much each factor contributed to this outcome, nor which factor contributed more. The reason is that muzzle velocity and angle setting do not make their contributions in a common currency. (Sober, 1988, p. 317)

Sober goes on to say that we can answer the question of what difference each factor makes, either locally or non-locally. If it is the case that the cannon is one which has, say, two possibilities of angle (let us think of these as x and y) and two possibilities of powder charge (let these be a and b) then one can work out the difference using local facts by working out the distance of a projectile fired when at angle (x) and powder charge (a); angle (x) and powder charge (b); angle (y) and powder charge (a); and angle (y) and powder charge (b). If the cannon has only one possible angle, however, then the difference must be worked out non-locally by comparing the distance of the projectile fired from that cannon (with various powder charges) with another similar canon with a different angle (but the same range of powder charges).

So, what we should take from this is that when trying to analyse the causal contribution of components of an effect (be it height, projectile distance, or, I will argue, behaviour and cognitive processing) we need to (1) work out what the components are, and, (2) analyse whether the phenomenon is an effect which is analysable in terms of the causal contribution of its components or whether the phenomenon under investigation is one which is only amenable to analysis in terms of the difference each component may cause. This can be done by identifying whether it is possible to eliminate completely one of the factors and still retain the effect; it is not possible to eliminate either genes or environment and still have an organism – therefore height is a phenomenon amenable only to difference contribution. Likewise, it is not possible to eliminate either the angle or powder charge of a canon and still have a projectile with velocity.

Once one has determined whether we should understand the contribution of a phenomenon's components to that phenomenon in terms of casual contribution or difference (for ease let us name these *casual contribution* and *difference contribution*, and likewise *causal phenomena* and *difference phenomena* for the phenomena whose contributions can be analysed in terms of causal contribution and difference contribution respectively) we then need to (3) work out whether the difference contribution can be worked out using local facts or whether non-local facts must be appealed to.

Target Phenomenon: Phenomenon components	Commensurable (measurable in same terms)	Local (can be manipulated in a particular case)	Type of contribution of the components
Particle acceleration: Electrical & gravitational forces	✓	✓	causal
Height: Genes & environment	x	x	difference
Cannon projectile distance: 2 possible angles & 2 possible charges	x	✓	difference
Cannon projectile distance: 1 possible angle & 2 possible charges	x	x	difference

Let us try to do this by analysing human behaviour (or *Cognition* with a capital 'C') into the components of affect and cognition. How can we determine whether flexible adaptive behaviour/Cognition is a causal phenomenon or a difference phenomenon? Is it possible to eliminate affect and still be Cognitive? Is it possible to eliminate cognitive processing and still be Cognitive? Or, by eliminating either would there be no longer a phenomenon under investigation (as in the case of height)? In the height case we can see easily that without genes or an environment there can be no organism to have the predicate of height.

Should we therefore argue that affect and cognition as the components of behaviour are analogous to the contribution of genes and environment in height, and *not* to the contribution of electrical and gravitational forces in particle acceleration? People often talk about emotion in terms of what contribution it plays assuming that it is dissociable from cognition. But if affect and cognition are comparable to the above example – the role of genes and environment in height – it is nonsensical to talk of how much emotion or cognition contributes to behaviour or how much affect contributes to cognitive processing; it only makes sense to talk of how much difference emotion or cognition makes to behaviour, or how much difference affect can make to cognitive processing. But as such these components are incommensurable and the difference can only be measured as a non-local phenomenon. Thus it would be nonsensical to talk of the causal contribution of emotion or cognition to behaviour in an individual or the causal contribution of affect/interoceptive processing to cognitive processing. If this is right then it may only appropriate to talk in terms of the contribution of affect or emotion to cognition in relation to population studies, and not individuals.

3. Explanatory separability and orthogonality

There is one other way in which the emotion/cognition and affect/cognitive processing distinctions might differ even from the case of the single cannon firing (the difference of the contributions for which had to be worked out non-locally). In the case of the cannon, manipulating the angle made a difference to the projectile velocity, but it made no difference to the gunpowder charge. That is, one can hold the gunpowder charge constant and manipulate only the angle of the gun; gunpowder charge and cannon angle are orthogonal even though they both contribute to the effect of the projectile velocity in terms of the difference they make to that phenomenon[32]. In such a case we might think of gunpowder charge, cannon angle, and projectile velocity as being explanatorily inseparable, even though gunpowder charge and cannon angle are orthogonal. It is thus possible that affect and cognitive processing could be explanatorily inseparable and yet orthogonal. This brings us closer to the spirit of explanatory inseparability in the sciences as explained in the quote from Hurley above, in that one can look for variables, which, though inseparable from the others in the grand explanatory scheme, can be isolated and varied to manipulate overall effects.

If affect can be manipulated without thereby altering cognitive processing (and vice versa) then affect and cognitive processing are orthogonal and despite their explanatory inseparability it may make sense to think of affect as not constitutive of cognition. But what would it mean for affect and cognitive processing to be non-orthogonal? Firstly, altering affect in any way one would thereby simultaneously alter cognitive processing. This is easier to visualise in terms of a state space whose parameters are affect within which the variables of cognitive processing reside. If the (affect) parameters are manipulated then the entire shape of the state space is likely to shift, giving a new topology of attractors and repellors which will change the shape of the cognitive processing taking place, or constraining which cognitive processes can take place. An example of such a case would be difficulty in recall of happy memories during a depressive episode (Teasdale et al. 1980; Fogarty & Hemsley 1983), which fits with how Ratcliffe (2008) talks about moods shaping possibility spaces. Likewise the converse should also hold. It ought to be the case that by changing the variables within the state space the entire state space is thus affected and changes shape. This would

[32] Likewise in Sober's explication of the relationship between genes and environment in regard to the height of corn plants (or people) genes and environment are orthogonal, in that one can hold the genes steady and manipulate the environment and vice versa. This is very likely not actually the case given what we know now about how the environment can actually cause certain genes to be expressed and how it can cause methylation of certain genes and so there is I think good reason to think that genes and environment are also non-orthogonal but I will not develop the argument for that here.

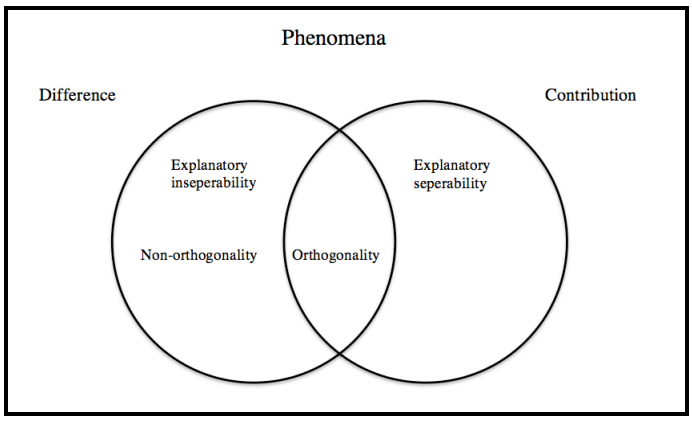
be the case in, for example, emotional regulation strategies which work by focussing attention on a task such as doing maths (engaging the dorsal regions of the neuroaxis as related in chapter four) and which subsequently inhibits activation in the ventral regions allowing relief of negative affect (see for example, Lewis & Todd, 2007; Holzheimer & Mayberg, 2011).

So far then I have outlined 3 ways we might understand affect as partially constitutive of cognition, without being constitutive in the metaphysician's sense (i.e. compositional and in opposition to casual factors). Let us call this kind of constitution which is relative to an explanatory project, and which does not imply the naïve metaphysical doctrine of containment, "constitution*":

(1) Affect and cognitive processing may be explanatorily inseparable
(2) Affect and cognition may be difference phenomena (and thus only manipulable non-locally)
(3) Affect and cognition may be non-orthogonal, such that neither can be manipulated without also thereby manipulating the other.

Non-orthogonality implies explanatory inseparability as, if by manipulating A, B is also manipulated, then one can clearly not hold A constant and manipulate B which is required for explanatory separability as outlined by Hurley above. Explanatory inseparability does not however imply non-orthogonality as the cannon example above showed; cannon angle, gunpowder charge and projectile velocity stand as an explanatory unit, but cannon angle and gunpowder charge are orthogonal. Orthogonality therefore can imply either explanatory separability or inseparability.

Being a contribution phenomenon implies explanatory separability as one addresses contribution by asking counterfactuals designed to elicit what the change would be to the explanatory target if one or other of the factors were absent in order to then work out their level of contribution. Explanatory separability thus implies that the target is a contribution phenomenon too. If the explanatory target is a difference phenomenon this implies that the factors involved are explanatorily inseparable as they are not amenable to the kind of local dissection described by the counterfactuals but rather difference must be assessed at non-local population levels.



Given the relations between processes that I have outlined so far in this chapter how can we understand the relation between affect and cognition? Are affective and cognitive processes (1) commensurable or incommensurable (2) local or non-local (3) do they have a causal or difference contribution[33] (4) are they explanatorily separable or inseparable (5) are they orthogonal or non-orthogonal?

4. Commensurability and orthogonality

Lets start with commensurability. What would it mean for affect and cognition to be commensurable? Affect is interoceptive information. One form of this information comes from the interoceptive receptors throughout the body. While this information has its own pathways in addition to feeding into other pathways it is essentially just neural information like any other. That is, it has the same currency as that information coming in from the exteroceptive senses and previously processed information that is now endogenous to the brain. Given the integration of information from affective sources with that from non-affective sources and the lack of a basis to discriminate between emotion and cognition on the basis of structure, function, and connectivity that I outlined in the previous chapter and can be seen in more detail in (Pessoa 2008, 2010), it would seem wise to dispense with the affect-cognition distinction at the neural level and just describe it all as cognitive processing. Whether that processing will lead to activity or thoughts that we tend to categorise as affective/emotional or cognitive will depend on which networks are involved and how they are integrated with other networks which subserve homeostatic, behavioural, sensory, or other internal processing. This is clear in the concrete abstraction of neural processing that connectionist networks provide. Information fed into the input units is parsed in the hidden

[33] I am assuming that their contribution is to the phenomenon of Cognition (with a capital 'C') i.e., flexible, adaptive behaviour.

units to reveal distinct categories of, for example grammatical structures (see Elman 1991), despite all the information being operated on being in the form of unit/synaptic weights. Affective information in the form of interoceptive information can thus be integrated with information from other sources and then proceed to play a role in processing which is particularly relevant to homeostatic operations or not. We might want to distinguish these two types of information as passive affect (interoceptive – afferent - information) and active affect (information pertaining to efferent homeostatic operations). It might be thought that active affect is incommensurable with non-affective cognitive processes. However, this is only because of its direction of information flow (physiological changes), and not a result of the currency of the information, as can be seen in Barrett and Bar's model of perception (Barrett & Bar, 2009) where efferent copies of these processes feed back in to the perception processing.

Affective information, however, does not only come in the form of afferent interoceptive neural information (or efferent information pertaining to physiological changes). The neuroendocrine system, chiefly in the form of the Hypothalamic-Pituitary-Adrenal (HPA) axis, allows the organs, immune system and nervous systems to be coupled through hormonal interactions. Though hormones act at a slower timescale than neural information transfer they are not so slow as to be inconsequential to cognitive processing – blood gets pumped around the body fairly rapidly. Likewise, neuromodulators and gasotransmitters are released by various brain areas and change the shape of the neural processing. It is typical to think of these interactions as modulating the processing rather than mediating it. However, we must be careful not to identify the mediation-modulation distinction with the causal-constitutive distinction. Clearly a modulator is changing a process, but at the same time it is also mediating the process that is the result of the change. If it is removed the process which it effected no longer comes about. The modulation intuition arises because one can think of an on-going process which, with the addition of x, changes the way it occurs, rather like turning the volume up or down on a process. But this is just an artefact of the choice of temporal slice we are looking at. If we just look at the period around the change the intuition is no longer so strong that x is performing a modulatory rather than a mediatory function. There does not seem to be an inherent reason to consider neuromodulators (and thus also hormones) to be modulatory (causal) rather than mediatory (constitutive). Indeed, the gasotransmitters work in a very similar way to the neuromodulators in that they are expressed as a cloud of molecules which pervade the surrounding cells. Unlike neuromodulators they do not require specific receptors in those cells to interact with them.

Rather the gaseous molecules are able to permeate cell membranes. Gasotransmitters are considered to be signalling mechanisms rather than modulatory molecules, and given the similarities between them and the neuromodulators I suggest that we should not read too much in to the 'modulator' part of the name 'neuromodulator', that is the term serves mainly to distinguish those molecules which act diffusely (and in the case of neurohormones also enter the blood stream) rather than just being sent from the pre-synaptic neuron to the post-synaptic neuron as signalling 'transmitters' are.

Re-evaluating neuromodulators in this light – as mediatory of processing like neurotransmitters and gasotransmitters, rather then 'merely modulatory' – is important, because I suspect that with this at the forefront of our minds the intuition that the molecular messengers are less 'part of' the neural processing than the spiking neurons is less likely to come about. In animal brains this 'gooey stuff' is explanatorily inseparable from the 'spiky stuff' (the electrical spiking function of neurons). Dopamine (a neuromodulator) is well known to be integral to memory, attention, and learning. Likewise the gasotransmitters nitric oxide (NO), carbon monoxide (CO), and hydrogen sulfide (H_2S) facilitate induction of long term potentiation (LTP) in the hippocampus (Wang 2002), a process which is thought to be one of the key neural mechanisms for Hebbian learning. Research on using models of gasotransmission in evolutionary robotics also gives us good reason for thinking that the gooey stuff and the spiky stuff are explanatorily inseparable.

5. GasNets and particular embodiment

In evolutionary robotics artificial neural networks (ANN) are evolved to control autonomous robot agents with the purpose of succeeding in a particular environmental task. The exact details of the nature of this evolution are not important to my argument, but the gist of the endeavour is that one starts with a group of neural networks hooked up to a simulation of the task environment[34] and a group of the more successful networks (according to a particular fitness function) as exhibited in an evaluation phase where the controllers are put to the task, along with some of those close by, are selected to 'grow' while the rest are culled. Recombinations of the 'genotype' and mutations among these are induced and the process begins again, resulting – over many generations – in networks that succeed in the task.

[34] In principle one could evolve the controllers while hooked up to the actual task environment but the amount of generations required to evolve successful controllers usually means that time-wise this is usually prohibitive. Instead controllers can be transferred to real-environment situations after they have been evolved in simulation, typically having been evolved to function despite a high level of noise to enable easier transference.

Standard artificial neural network models are designed so as to model electrical transmission between nodes. This can be adapted to also model gasotransmission, such that the activation of a node is not only a function of the inputs from the connected nodes but also of the concentration of gas at that node (for more detail see Husbands 1998; Philippides, Husbands & O'Shea 2000; Smith, Husbands, Philippides & O'Shea 2002, p. 162-164).

Smith et al. compare the evolvability and adaptivity of a GasNet (ANN with gas model) solution with a NoGas (typical ANN) solution to a task in which the robotic agents, starting from an arbitrary position in a blackwalled arena had to find and navigate towards a white triangle while ignoring a white square. In this GasNet two gases are modelled. Nodes in the network may emit no gas or one of two types of gas. Nodes that emit one of the gases will do so only if activity in the node increases beyond its electrical threshold or if the concentration of one of the gases increases beyond a threshold of 0.1 (Smith et al. p. 164). Drawing on previous research (Smith 2002), the authors show that basing evolution of solutions on the GasNet class (rather than the NoGas class) "consistently produces successful robot control solutions in significantly fewer evaluations" with these results holding "over a number of different evolutionary algorithms, with a number of different mutation and recombination rates used.." (Smith et al. p. 166). What this means is that GasNets seem to be more evolvable than NoGas networks. The authors argue that there are three reasons for the heightened evolvability of GasNets (1) that GasNets are more amenable to being tuned to the particular characteristics of the environment (2) the features of the gas diffusion mechanism enable switching between stable states to be based on input patterns received over time (3) being able to base output on input patterns received over time means that the GasNet can easily filter out noisy input by requiring input to be consistent over several time steps (Smith et al. p. 174-175). Note that the GasNet and NoGas solutions had evolved functionally equivalent timing mechanisms so on its own the presence of a timing mechanism should not have played a difference in the evolvability of the GasNet solutions.

The first point, that GasNets are more amenable to being tuned to particular characteristics of the environment, means that they should be more adaptive (as a species over evolutionary time) than NoGas models. The authors show that the GasNet controllers are more amenable to being tuned by re-evolving the GasNet and the NoGas controllers but this time only allowing the parts of the genotype that were involved in the timing mechanisms to be affected. The controllers were then re-evolved in two different environments; where the motor speed is set to double what the controllers were originally evolved for and where the

motor speed is set to a quarter of what they were evolved for. The difference in the number of generations which were required to re-evolve controllers which had 100% fitness was dramatic. In the double speed environment both controllers initially drop to only 20% fitness, but the GasNet only required 10 generations to reach 100% fitness compared to 409 generations for the NoGas. In the quarter speed environment the GasNet initially dropped significantly less in fitness than did the NoGas, however the authors claim that "the difference in the number of generations required to reach 100% fitness is much larger than might be predicted by this fitness difference (30 generations on average compared with 591 generations)" (Smith et al. p. 177-178). Similar, but less dramatic, differences in re-evolution rates in the double- and quarterspeed environments were also evident when the experiment was repeated but this time without restricting evolution to just the timing mechanism. So, while GasNets look to be extremely useful for evolving and tuning timing mechanisms this does not seem to be the sole reason for GasNets tunability.

This research is particularly enlightening for my argument for two reasons. Firstly, rather than relegating the gooey stuff to the affective heap and ignoring it as irrelevant to the proper cognitive processes, they have integrated it in to computational models of behaviour. And in doing so they have shown that it isn't merely a background or enabling condition, but that it plays a key role in evolvability leading to populations that can quickly adapt to a learning task and a particular environment. It should be noted that the solutions that evolve are very sensitive to environmental changes as they have evolved for a particular environment. However this can be countered by evolving the GasNets to a number of different (noisy) environments which will result in them evolving to be robust. Secondly, the fact that both GasNet and NoGas solutions evolved (even if the NoGas solutions took longer) does not mean that the gas in the GasNet controller was not constitutive* of the artificial neural processes that underpinned the robots behaviour in the GasNet solutions. That is to say, the role that the gas played in the GasNet solutions is not explanatorily separable from the role that the electrical nodes play. One cannot evolve a GasNet controller and then remove the gas and expect it to work. Nor can one explain how the solution works without appeal to the properties that the gases imbue the GasNet with as well as the electrical nodes.

Recall that while the GasNet and NoGas controllers had functionally equivalent timing mechanisms these resulted in quite different amenability to being tuned to a particular environment. The strategy employed by both was to time the duration of bright visual input

that is received in in the upper half of the visual field – this allows discrimination of triangles and squares as the former will result in a smaller duration of bright visual input that is received than the latter. A bright object finding mechanism is then put into play, but is inhibited if the duration of the visual input is sufficiently long. The GasNet timing mechanism was realised by a solution which utilised the emission of gas from a visual input node when activity in that node is high, with the time it takes for the concentration of the gas to reach the level at which it is sufficient to activate other nodes specified by the genotype. Likewise the time taken for the gas to decay once there is no longer any bright visual input is "a function of the genetically specified rate of gas concentration build-up" (Smith et al. p. 176). In the GasNet solution which they discuss, the timing mechanism is hooked up to motor output in the following way. When bright input to two visual nodes is sufficient, another node emits a gas. When the gas concentration is high enough, the right-forward motor node is inhibited. The effect of this is that while the robots are seeking the triangles (by rotating anti-clockwise) if a square is spotted, the bright input will be present over time, with the result that the right motor node is inhibited for a period of time which will mean that the agent will rotate clockwise and past the square. Whereas in the case that the triangle is spotted the concentration of the gas will only be enough such that the agent will curve back towards the triangle.

The NoGas timer solution on the other hand has the motor node fully connected to two other nodes, one of which responds to bright input. The system can settle into one of two equilibrium points; one if visual input is below threshold, and another when input is above threshold. Due to the nature of the recurrent feedback at work in the subnetwork when bright light ceases to be received the second state slowly decays back to the first, with the length of time the decay takes being a function of how near the system was to the second stable state. When there is sufficient bright light input, as is the case when the square is the object of bright light, the left motor node is inhibited resulting in the robot passing the object rather than approaching it.

However, despite the functional equivalence of these two subnetworks, the differences of their particular implementation makes itself apparent in the difference in tunability that I outlined above. Just the mere fact that both are successful solutions to the environment that they have evolved for, and therefore functionally equivalent in regard to that, does not mean that that level of explanation at which we see the functional equivalence is the correct one for us to understand what is really key to the ability of each controller to succeed. Rather, by

looking at the ease of evolvability and the mechanisms which underpin this amenability to being tuned to a particular environment we can see that the relevant level of explanation for the adaptive behaviour of the controllers is that which specifies the interaction of the gas and the nodes. The moral of this story is this: *In evolved systems this is not just the implementation, but is the relevant level for the algorithm of an adaptive system.*

6. Particular embodiment matters

To further pursue this argument, let us consider another example from evolutionary robotics. Typically in evolutionary robotics, algorithms are evolved in simulation and then transferred to hardware. Adrian Thompson has shown however that it is possible to evolve algorithms straight on to the hardware. In Thompson (1995, 1997) he discusses his work using a Field-Programmable Gate Array (FPGA) which is basically a very large silicon chip containing components, wires, and switches which determine how the components behave and how they connect to the wires (A.Thompson 1997, p. 103). An FPGA is essentially 'empty' in that using a computer one can configure the switches in the array how one likes to create a physically real electronic circuit. Thompson used the array to control a real-world robot. He used evolutionary algorithms to configure the switches and then was able to evaluate the circuit based on its performance on a task in the real-world (i.e. controlling the robot), modifying it based on evolutionary algorithms until performance was satisfactory. Thus, there was no simulation involved either of the control circuits or the robot and environment.

Evolving the array in the real world means that the simplifications that designers are forced to use to design circuits, out of practicality, like breaking down the system into modules or hierarchies, and abstracting away from the detailed behaviour of individual components and their interactions, can be abandoned. There is no need for top-down constraints:

> There is no analysis, simulation, or modelling, so no constraints need to be placed on the circuits to facilitate these. Evolution proceeds by taking account of the changes in the overall behaviour as variations (usually small) are made to the circuit's structure: this means that the collective behaviour of the components can be freely exploited without having to be able to predict it from a knowledge of their individual properties. Evolution can be set free to exploit the rich structures and dynamical behaviours that are natural to the silicon medium, exploring beyond the scope of conventional design. The detailed properties of the components and their interactions can be used in composing this system-level behaviour. It takes considerable imagination to envisage what these evolved circuits could be like: the kinds of systems we are familiar with (e.g. digital, discrete-time, computational, or

> hierarchically decomposed circuits) are but a subset of what is possible. (A.Thompson 1997, pp.107-9)

Thompson describes two experiments that are relevant to us here. One using a FPGA chip and the other using an evolvable hardware architecture he calls the 'Dynamic State Machine' (DSM), which – for our purposes – essentially does the same job as an FPGA chip (the DSM was used for one of the early experiments as suitable FPGAs weren't available). The DSM was set up to be connected to a real-world robot, whose task was to display wall avoiding and room-centering behaviour, being equipped only with two sonars for sensors and with motors that were not allowed to run in reverse. Thompson removed the 'clock' in the DSM which is a conventional constraint on a circuits dynamics that is typically put there to enable it to be modelled in Boolean logic. Instead he placed the clock under evolutionary control. The idea behind this was to convert the clock from being a constraint on the systems dynamics to being available as a resource "which can be used to further enrich the continuous-time dynamics of the circuit" (A.Thompson, 1997, p. 109). The control system was evolved in the array, i.e. as real hardware, to control the real motors in the robot for all the fitness evaluations. What is interesting is that the final evolved control system not only had a quite different solution to those which could be designed through conventional methods (and a solution which could not be modelled in Boolean logic because the continuous-time properties of the hardware are important to its operation) but also utilised only a tiny amount of electronics for its sensorimotor control structure.

The surprising properties of evolved hardware are even more evident in the experiment which Thompson describes, in which he uses the FPGA chip to discriminate between two inputs (1kHz and 10kHz) with the intended output to go to "a steady +5V as soon as one of the frequencies is present, and 0V for the other one" (p. 110). Thompson disabled the synchronising clock normally used with the FPGA in order to see if evolution could "exploit the rich natural unconstrained dynamics of silicon" to succeed in the task (p. 110). This is a significant ask of the FPGA:

> … all that is available is 100 FPGA cells, each intended to perform a single Boolean logic function, and each having a delay from input to output of just a few nanoseconds (billionths of a second). How could an arbitrary structure (potentially having many recurrent – feedback – connections) of these 100 simple high speed logic gates be evolved to discriminate perfectly between input periods *five orders of magnitude* longer than the delay through each component? Success would be significant: as well as vindicating the 'unconstrained' approach to hardware evolution, the resulting circuit (requiring no external components or clock) would be incredibly efficient in its use of silicon. (A.Thompson, 1997, p. 114)

Through the generations that were evolved solutions emerged that "would seem utterly absurd to a digital designer". Despite the FPGA being a digital chip, and the fact that the experiment was to evolve a recurrent network of logic gates, Thompson found that the gates themselves (in the chip) were not being used to 'do' logic. Instead evolution used "whatever behaviour these high-gain groups of transistors happen to exhibit when connected in arbitrary ways" (p. 118). Of particular interest is that in the final evolved circuit, about a quarter of the cells could be observed to be contributing to the behaviour. However, some of these cells which influenced the behaviour were not connected to the main part of the circuit, nor was there a route of connections between them and other connections by which they could influence the output pin! They were clearly contributing to the behaviour though, as attempts to clamp one of these 'unconnected' cells to a constant value resulted in the system malfunctioning. The only way that these cells could be contributing to the behaviour of the circuit is through exploiting the physical characteristics of the chip (other than the wire connections). Thompson suggests electromagnetic coupling or power-supply loading as possible means for this interaction. These properties are clearly reliant on the exploitation of physical space in the chip.

The exploitation of physical characteristics enables the systems to evolve solutions which have a greatly decreased computational complexity when compared against traditionally designed algorithms. However a result of this is that the line between algorithm and implementation has been blurred so that it is not going to be a trivial matter to implement an algorithm which has been evolved on a particular piece of hardware on a different piece of hardware. The peculiarities of a particular chip mean that if an algorithm has been evolved which exploits those, then that algorithm will not work on a chip that does not have those peculiarities. Just as this case shows that in evolved systems the algorithm may be partly implementation, we can also see that it causes problems for thinking in terms of constitution versus background conditions. What should we consider the role of the 'unconnected' cells to be? Intuitively one might suppose that the connected cells are constitutive of the behaviour and the 'unconnected' cells to be mere background conditions. Yet, there is no principled reason for supposing this. If either type of cell is clamped then the system malfunctions. It would be just a matter of electrical 'chauvinism' to suppose that the 'unconnected' chips are any less a part of the solution than the connected chips.

My argument is that likewise, many of the components of our physiology; affective information, gasotransmitters, neuromodulators and neurotransmitters, hormones etc. are not mere background conditions for cognitive processing but are as constitutive* as the neural electrical processes are. I think this argument can be taken even further however with an example of the exploitation of the properties of skin cells in the sensory system. Boulais and Misery (2008) review research that strongly suggests that cells in the epidermis contribute to the sensory transmission of touch, thermal sensation and pain.

Boulais & Misery argue that keratinocytes, melanocytes, Langerhans cells and merkel cells, along with neurons, are part of an integrated neuro-immuno-cutaneous system which consists of a "common language" shared by these cells "with the neuromediators as letters" (Boulais & Misery, p. 119). They argue that given that between the trunk and the distal parts of the limbs the numbers of nerve endings decreases without decreasing touch sensitivity, epidermal cells may relay signal transduction. In particular keratinocytes and merkel cells express some of the same proteins (which act as receptors) found in neurons (in particular C-fibres, Adelta fibres, and Abeta fibres). Keratinocytes have receptors which allow them to be sensitive to thermal and noxious stimuli, and the stimulation of these receptors results in the release of neuropeptides which can act as modulators for other epidermal cells, but most interestingly can also act as neurotransmitters onto target cells. While this is not thought to be the mechanism by which keratinocytes transduce signals to sensory neurons it is interesting to see that this kind of signal transduction between cells can occur without the spiking and synapses particular to neurons. Boulais and Misery suggest that a possible mechanism for signal transduction from keratinocytes to sensory neurons could be through purinergic receptors (such as receptors for ATP). They explain:

> It has been shown that ATP-activated cells can increase their intracellular calcium concentration, producing a calcium wave able to propagate to neighbouring cells. The ATP-dependent calcium waves so produced by keratinocytes can induce an increase in intracellular calcium concentration not only in adjacent keratinocytes, but also in sensory neurons. (Boulais & Misery, p. 123)

Keratinocytes are also receptive to neuropeptides and upon activation release mediating proteins. The authors suggest that this results in activation of neighbouring cells, and possibly through a cascade of paracrine signalling (signaling to nearby cells) lead to the depolarisation of nerve terminals.

> Thus, keratinocytes synthesise the key components which endow them to sense many physical variations and process the information perceived. The ion channels and neuropeptides originally found in the brain make the keratinocytes true partners for neurons. (p. 123)

The functioning of merkel cells provides an even stronger case for sensory transmission with neurons. Boulais and Misery explain that they (1) synthesise neuropeptides inside neurosecretory granules (2) these granules are mainly located near the sensory neurons which supply Merkel cells (3) they cluster around a slowly adapting mechanoreceptor (SAM) comprising a Merkel cell-neurite complex. This leads the authors to suppose that either merkel cells are mechanoreceptors themselves, which synaptically transduce the signal to sensory neurons, or they modulate the sensory function of neurons (p. 124). However the former seems particularly likely given that (1) "they express the most of the proteins involved in vesicle trafficking and recycling" (2) they have many of the components of glutaminergic transmission machinery, and (3) they have voltage-gated calcium channels which "are normally found in excitable cells and reveal synaptic capability, since quick calcium currents are believed to be involved in cell depolarisation and neurotransmitter release" (p. 124). The authors note that it may be the case that merkel cells are in fact the postsynaptic cells rather than pre-synaptic cells, but that it is likely however that they are still capable of activating the sensory neuron after their depolarisation and release of their neurosecretory granules (p. 125).

While Boulais and Misery make clear that we don't know for sure yet whether these epidermal cells should be considered as transmitting information to each other and sensory neurons or modulating each other and sensory neurons, what they do show is enough for my argument. The intuitiveness of the distinction between cognitive processes and merely affective ones, and between constitutive processes and background conditions dissolves in the light of how evolution exploits our particular embodiment. It is just not clear where those processes that are "cognitive" start and those that are "affective" or "modulatory" or "mere background conditions" stop. The result is that a specification of the algorithm for human cognition is going to be messily entwined with some parts of its particular implementation.

7. Summary

I have argued that the causal-constitutive distinction only makes sense in relation to a particular explanatory project. My explanatory project is Cognition, understood as flexible,

adaptive behaviour, and in relation to this I have argued that affect is constitutive of Cognition. I outlined two ways of understanding affect; either as interoceptive neural information as was the focus of the previous two chapters or as what I have termed the "gooey stuff" which is meant to cover the kinds of molecular signalling that we see both in the brain and as part of the neuro-endocrine system, and indeed between non neural (somatic and glial) cells. I argued that neural affective processing is commensurable with cognitive processing and, given the network nature of the brain outlined by Pessoa which I reviewed in the previous chapter, there does not seem to be any sense in differentiating these types of processing as 'affective' or 'cognitive'. It is all cognitive processing, thus in this way affect is partially constitutive of cognitive processing.

If we interpret affect as the "gooey stuff" we can see that this is incommensurable with the "spiky stuff". This could provide a foundation for thinking that it is a mere background condition or that it is "merely" modulatory. However the work on GasNets that I have reviewed in this chapter suggests that while the gooey and the spiky are incommensurable they are non-orthogonal. That is they are coupled such that manipulating one will effect changes in the other. This implies explanatory inseparability, and suggests that we should think of the contribution of affect to cognition in terms of difference rather than causal contribution. Further, it implies that questions about contribution of the gooey to cognition only make sense in relation to populations rather than in individual cases. Finally, given that evolved systems exploit their particular embodiment to find very good solutions for behaviour I suggest that it would be mere electrical chauvinism to disregard the gooey stuff as "merely" modulatory or as a background condition, rather than properly constitutive of cognitive processing.

Conclusion

The aim of this thesis was to show that the body matters for cognition in ways that go significantly beyond those seen in standard 'embodied cognitive science'. I have argued for this by showing how interoceptive information (in the form of afferent homeostatic signals) is interwoven in various forms of cognitive processing. We also saw how efferent homeostatic information may feed cognitive processing, and (in the last chapter) how ('gooey') molecular signalling throughout the nervous system and body proper (including but not necessarily limited to peptides, neurotransmitters, hormones and gasotransmitters) might need to be considered alongside the more familiar 'spiky stuff': the electrical 'spiking' function of neurons. The philosophical streams of both mainstream and 'embodied' cognitive science (see for example Bermudez, 2010; Clark 1997; 2001; Harre 2002) and even Chemero's "radical embodied cognitive science" Chemero, 2009) have in the main ignored both affect and 'gooey' signalling, and seem to have assumed that the relevant functional level for understanding cognition can safely abstract from (or factor out) such implementational details. We saw that in evolved systems processing routines can come to exploit all manner of messy 'implementational' properties to the point where it is no longer obvious how to distinguish the two.

1. The abundance of 'gooey' signalling

The neurobiology of the 'gooey' stuff is fascinating and complex and its contribution to cognitive processes is still in the process of being fully understood. But the prejudice against it playing an active (rather than a 'mere modulatory') role is beginning to dissolve, as more is understood about how the gooey stuff and the spiky stuff interact. For example, while not that long ago it was thought that hypothalamic hormones released as part of the neuro-endocrine system only acted downwards (to endocrine and exocrine glands in the body) it has now been demonstrated that hypothalamic peptides can also act upwards to parts of the brain (Kastin & Pan, 2010, section 1). Also, we now know that peptides from the periphery (body) can act on the brain; and that they can cross the blood brain barrier (Kastin & Pan sections 2 & 3). Not all peptides and hormones can cross the blood brain barrier but (1) there are areas in the brain in which the blood-brain barrier is not intact, providing possibilities for gooey access to neurons and glial cells, and (2) even where the blood-brain barrier is in place, there are other mechanisms that may allow the gooey stuff to contribute to neural

function, such as acting on cerebral endothelial cells to change their function (Kastin & Pan, 2010, section 3). Kastin & Pan also argue that the actions of peptides aren't necessarily limited to the period of their half-life in blood, rather they can initiate secondary effects and signal cascading. And, that peptides can have more than one action; they can have several cascades.

We know that cells other than neurons, such as glial cells, endothelial cells, and epithelial cells signal each other in many cases using the very same mechanisms as neuron-to-neuron signalling, such as calcium signalling (Berridge, 1988) metabolic and hormonal signalling (Levin et al., 2011), and (neuro)peptides (Kastin & Pan, 2010; Li & Kim, 2008). Just as gasotransmission in the central nervous system functions without requiring receptors in the receiving cells, peptides and calcium signalling also do not require specific receptors. This being the case, it is perhaps not surprising that the body proper and the central and peripheral nervous systems are intertwined in complex ways. But this also provides a platform for the objection that such signaling is 'merely modulatory' rather than mediatory, and thus is part of the background for, rather than constitutive of, cognitive processing.

In the previous chapter I outlined why I think that in this context the distinction, between background and constitutive processes is not helpful in our quest to understand cognition. While it may be helpful, in respect to a particular explanatory project, to background certain phenomena (say the gooey stuff) and focus on other phenomena (say the spiky stuff), this is useful only when the explanatory project is clearly delineated. In the case of cognition this is not the case. While we can carve up cognition into operations such as perception or memory, understanding cognitive processing as the substructure which allows us to be flexible and adaptive creatures provides no such clearly delineable operations (this is something I will come back to later in this conclusion).

2. Back to the GasNets

Philippides et al., explain that traditional artificial neural networks factor out chemical signaling as they have taken the mid-twentieth century model of neuron-neuron communication according to which:

> Neurons generate brief electrical signals (action potentials), which propagate along wirelike axons terminating at highly localized junctions (synapses) on other neurons, where the release of a chemical signaling molecule, or neurotransmitter, is triggered.

> The neurotransmitter is confined to the region of the synapse, and here the receiving neuron is equipped with receptors that directly translate the chemical signal into a brief electrical signal, either excitatory or inhibitory." (Philippides et al., 2005, p. 141)

One can simulate such a model entirely in terms of "electrical signals flowing between nodes in a network". The work I have outlined in this thesis shows that such a model was vastly incomplete. The reality of inter-cellular signalling is highly complex and intricate, and is still the subject of intense research to understand the precise mechanisms.

It might be argued that the main benefit of the diffuse gooey stuff is that it provides a richer internal dynamics for the system than can be provided by the kind of (principally) electrical model, allowing multiple time-scales to be in operation, and this in turn allows the system to find elegant, low-cost solutions. While Philippides et al. acknowledge that the rich internal dynamics may well play a role in the "much greater evolvability of the GasNets" (p. 153), they argue that the story is more complex than just reducing to dynamics. They evaluate the performance of three types of GasNet, the original and two modified GasNets; Plexus and Receptor. The two modified GasNets allowed for more flexibility and looseness of coupling, and were more successful (in terms of high evolvability to very good solutions) than the original GasNet. The most successful was the receptor model, and the authors argue that the reason for this is that this model of gaseous and electrical signaling had a bias towards loose coupling between these processes:

> The receptor model allows site-specific modulations, including no modulation (zero quantity of receptors) and multiple modulations at a single site. This provides a powerful context-switching mechanism that pulls the chemical and electrical processes further apart, allowing (but not forcing) looser coupling, while further increasing the potential for complex network dynamics. (p. 150)

The authors argue that the success of the Receptor model was due to loose coupling. If that were the case then "the distribution of fitnesses of mutations that affected either chemical or electrical signaling systems independently would be different", which they explain was the case; mutations that only changed the gaseous connections had superior average fitness and "a distribution biased towards high fitnesses" whereas the mutations which changed both gaseous and electrical connectivity were most detrimental "indicating destructive interference between the two mechanisms" (pp. 155-156). They conclude that:

> "systems involving distinct yet flexibly coupled processes are highly evolvable when there is a bias towards loose coupling between the processes; this allows the

> possibility of evolution tuning one against the other without destructive interference". (p. 157-8)

Loose coupling between two distinct but highly interacting systems does not imply that one of these systems is more constitutive of cognition than the other. Rather, it shows that the processes that produce the kind of flexible and adaptive processing that we see in cognitive creatures are likely to not be limited to a single type of substructural system. The heightened evolvability of loosely coupled systems seems to be at least partly a result of their producing more workable mutations, and thus a greater proportion of fitter mutants survive selection (p. 157). Many of these mutations in the genotype do not affect the phenotype (pp. 153-156) and thus a large variation in genotypes can arise providing a large base for possible changes that can, in subsequent generations, affect the phenotype in positive ways (of course in negative ways too, but as I just explained the greater proportion of fit mutants means that it is more likely that at least some of these will affect the phenotype in positive ways).

Philippides et al. note that although they have concentrated on changes to a plastic system over evolutionary time scales "very similar issues are likely to be important at the time scale of the plastic changes themselves" (p. 158). I think that this is right, and is why what we learn about the role of the gooey stuff in human cognition is relevant not only to understanding how and why our cognitive processes have evolved to be loosely coupled processes between the gooey and the spiky, but also what may be fundamental to flexible and adaptive systems in general. It may be the case that in order to be a flexible and adaptive system, the substructure of that system must have distinct processes that are able to be loosely coupled in the way that the gooey stuff and the spiky stuff is in animal cognition, in order to allow online "evolution" in the form of plasticity. To be clear, this is not a claim about the gooey stuff per se, but rather the functional properties of the gooey stuff. These could be implemented electrically. After all, remember that the GasNets simulate some of these processes with "virtual" gases. However the particular functions of the virtual gases (as opposed to the functions of the electrical nets) seems to be particularly good for loose coupling and enables the search space to be significantly constrained, while at the same time giving rise to a greater number of fit mutants to be selected on than is the case with purely electrical systems. It may be the case therefore that in practice the only kinds of stuff that possess the functional properties that enables the kinds of flexible adaptive processing we see in animal cognition is the gooey stuff that we have, in a 3D gooey and spiky environment. But I do not want to make such a strong claim here, it will come down to empirical testing to find out whether or not this is the case.

Philippides et al. note that the biology of gas diffusion in real brains, and the subsequent modelling of GasNets parallels the embodied cognition approach to cognitive science, but internally:

> In highlighting the functional importance of brain morphology, these phenomena take us increasingly further away from connectionist ideas and suggest that Pfeiffer's notion of ecological balance, which requires a harmonious relationship between an agent's morphology, materials and control, can perhaps be taken inside the head (p. 145)[35].

I think that this is just how we should understand the functional importance of all of the affective and gooey phenomena outlined in this thesis. We are not only cognitively embodied in virtue of our morphology, sensorimotor systems and possibilities for action in the world as the popular embodied cognitive science movement has been emphasising. In addition we are internally cognitively embodied. And beyond that which Philippides et al. suggest; it is not only inside the head that we should take this functional embodiment, but into the body proper too, including the affective systems, the neuro-endocrine systems and potentially non-neural bodily cells.

3. Standard embodied cognitive science

How does this proposal fit with the increasingly popular framework of 'embodied cognitive science'? As explained above, I think of my thesis in terms of an extension to embodied cognitive science. Clark (2008) distinguishes between two claims about the role of the body in embodied cognitive science which he argues reveals a tension. On his account, the claim which he calls the larger mechanism story (LMS) (and which he identifies his position with) is that:

> larger systemic wholes, incorporating brains, bodies, the motion of sense organs, and (under some conditions) the information-bearing states of non-biological props and aids, may sometimes constitute the *mechanistic supervenience base* for mental states and processes. (Clark, 2008, pp. 39-40)

The second claim, which Clark argues is potentially in tension with LMS is the special contribution story (SC). This is the claim that:

[35] They are referring here in particular to Pfeiffer & Scheier (1997), in my reference list under Pfeiffer & Scheier (2001).

> Specific features of the body (and perhaps the world) make a persistent, non-trivial, and in some sense special contribution to our mental states and processes. (Clark, 2008, p. 40)

In keeping with the focus on morphological and sensorimotor capacities in popular embodied cognitive science, Clark focuses his argument on accounts which propose that the specific morphological and sensorimotor features make a special and ineliminable contribution to cognition (such that we could not have the cognitive processes we do if these were absent or different). I am not interested in defending the targets of his argument here. Rather I suggest that the picture that I have presented shows that Clark's distinction between a 'larger mechanism' story and a 'special contribution' story is far blurrier than it might seem if we were considering embodiment as consisting solely of an agent's gross bodily structures.

My claims amount to arguing that specific internal features of the body (interoception, and all the affective processes that implicate the gooey stuff) make a persistent, non-trivial contribution to our mental states and processes. Should we think of this contribution as "in some sense special"? Not in the strong sense of being un-functionalizable. But to get the right functional equivalence we would need to reproduce parts of some very particular internal bodily mechanisms. How low we go to get this functional equivalence, that is, how tiny the functions that need to be replicated for a functionally equivalent flexible and adaptive system to ourselves, has yet to be determined.

The resultant story might therefore be considered a "smaller mechanism story" about the role of the body in cognition. This is clearly consistent with embodied cognitive science. In his own (1989) work on 'microfunctionalism' Clark argues that functionalism need not be identified with formal descriptions pitched at a gross level, but that what is essential to functionalism is merely that the "structure, not the stuff, counts" (Clark, 1989, p. 31). He then argues that 'microfunctionalism', by specifying internal state transitions "at a very fine-grained level" doesn't need to give a functional specification of each type of mental state but can give "an account of the kind of substructure needed to support general, flexible behaviour of a kind that makes appropriate the ascription to the agent of a whole *host* of folk-psychological states" (p. 31).

Here Clark is rejecting the kind of gross level functionalism about mental states that identifies the formal description of cognition with particular operations (in philosophy, these might be considered to be types of mental states, and in cognitive science, these operations would be cognitive activities such as perception or memory). Instead he supports looking for the relevant level of formal description at the processes that support "general, flexible behaviour". My position belongs in this broad camp, extending its scope from the connectionist neural networks in which Clark grounds his argument, to incorporate all manner of gooey interoceptive systems. My position is thus that microfunctionalism as presented in the 1980's was still too abstract and high-level. The 'smaller mechanism' story presented here might thus be thought of as a 'nanofunctionalist' story, where the functions that matter to cognition are pushed further down towards the implementational limit than was previously envisioned.

An embodied cognitive science grounded in nanofunctionalism in the way that I have set out, not only makes the distinction between mechanistic stories and special contribution stories of embodiment increasingly blurry, but also goes some way towards eroding the apparent divide between popular embodied cognitive science and the emerging enactive paradigm in cognitive science.

4. What *is* the relation between emotion and cognition?

I have argued that it is important to separate affect and emotion, and that affect is grounded in interoception, both that mediated by neural structures and gooey processes. My thesis is that these affective processes should be understood as interwoven with those processes traditionally considered to underpin cognition and thus that cognition is inherently affective, and thus properly embodied. It therefore follows that whatever story one wants to give about the mechanisms that underpin emotional behaviours or states, emotion is inherently bodily. In this way any nanofunctionalist theory of emotion will be bodily in a more fundamental way than the embodied theories discussed early in the thesis. It will not only rely on the perception of bodily states (cf. Prinz) or representations of value (cf. Damasio) but will posit a continuous stream of afferent and efferent information shaping and forming such effects. The traditional distinction between emotion and cognition, where they are placed in stark opposition to each other, is here abandoned. But this does not force us to adopt any particular more general theory of emotion. It simply requires that any such theory have affect at its core. The account on offer is thus compatible with the bulk of emotion research

in psychology and emotion science in general and helps provide a framework within which to interpret such findings.

5. Possible future directions of research

I have argued that the relevant functional level for understanding cognitive (where this now means 'cognitive-affective') processes may be very close to or even intertwined with the implementational level in natural cognitive systems. It remains unclear just how 'low' we must go if we are to replicate (or even properly model) human cognition. Must we replicate/model the entire body, or is there a level which really is "mere implementation". How would we know when we were at this truly implementational level? Might we have to replicate/model down to the atomic or even quantum level? This is not inconceivable given that our understanding of cell functioning relies upon explanations involving ions and ion channels.

I argued earlier that this extension of embodied cognition brings popular embodied cognitive science into dialogue with the enactive paradigm in cognitive science. It should be noted that there are fundamental tensions between these two approaches to embodiment which I have not covered in this thesis. Some of these tensions have been elaborated by Di Paolo in his work on "Deep Embodiment". One of these tensions involves rejecting the kind of functionalism said to inform popular embodied cognitive science. But the version of nanofunctionalism that I have set out here is entirely consistent with the key concepts in enactivism such as self-organisation, agency, autonomy, experience, the self, and sociality (for this listing, see Di Paolo Lecture on Deep Embodiment; Thompson 2007). While my thesis has not specifically spoken to any of these, the embedding of affect within cognition is fundamental to all of them. Connecting this with, for example, enactive notions of autonomy and sociality in a full way will require significantly more work. While this has been done to some extent in work on enactive emotions (for example, Colombetti & Thompson 2007; Colombetti 2009) combining this with detailed work on the physiological body is a future project.

6. Last words

This thesis began by arguing that even Prinz's 'embodied appraisal theory' ultimately fails to be any more embodied than mainstream appraisal theories in emotion psychology. I then

showed that Damasio's somatic marker hypothesis, though it was meant to reconcile emotion and cognition, ultimately perpetuates an artificial distinction between emotion processes and cognitive processes. I then discussed work in neuroscience which shows that processing even as supposedly 'cognitive' as that involved in attention and executive functioning cannot be separated from the processes that underpin emotion and affect, and that such processing can not be distinguished as emotional or cognitive in terms of structure, function or connectivity. I argued that it is useful to distinguish affect from emotion, and I showed how affect can be grounded in various facets of interoception. I then showed that affect should be considered *constitutive* of cognition in various ways including being involved in perception, structuring perceptual phenomenology, and being a core ingredient of many (I believe all, but cannot claim to have demonstrated this) forms of cognitive processing. I then discussed what it meant for affect to be partially constitutive of cognition and I argued that there are two main ways of understanding how affect could be constitutive of cognition (1) in terms of interoceptive processes, and (2) in terms of gooey signaling. I then presented a proof of principle for the possibility of some fundamental involvement of gooey signalling from work in hardware evolution and evolutionary robotics. The upshot is an extension of embodied cognitive science and perhaps a step towards bringing popular embodied cognitive science into a more productive dialogue with the enactive paradigm.

In summary, I have argued for a properly embodied model of cognition; a model which is embodied in terms of its internal, interoceptive structures and processes as well as its gross morphological and sensorimotor features. The result of this is that our particular embodiment matters, down to a level at which it is unclear what (if any) aspects of the bodily stuff could be factored out while keeping all the relevant functional properties intact. Affective processes, even those reliant upon gooey signalling, here emerge as integral to cognition. Cognition is thus not merely embodied in respect to morphology and sensorimotor processes; it is *properly embodied*, in all its gooey glory.

Bibliography

Abbott, N. J. (2002). Astrocyte-endothelial interactions and blood-brain barrier permeability. *Journal of Anatomy*, *200*(6), 629-638.

Arnold, M. B. (1960). *Emotion and Personality*. New York: Columbia University Press.

Auvray, M., Myin, E., & Spence, C. (2010). The sensory-discriminative and affective-motivational aspects of pain. *Neuroscience & Behavioral Reviews*, *34*(2), 214-223.

Bar, M. (2009). Predictions: a universal principle in the operation of the human brain. Introduction. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*(1521), 1181-1182.

Barad, M., Gean, P.-W., & Lutz, B. (2006). The role of the amygdala in the extinction of conditioned fear. *Biological Psychiatry*, *60*(4), 322-328.

Barrett, L. F. (2006). Valence is a basic building block of emotional life. *Journal of Research in Personality*, *40*(1), 35-55.

Barrett, L. F., & Bar, M. (2009). See it with feeling: affective predictions during object perception. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*(1521), 1325-1334.

Barrett, L. F., & Bliss-Moreau, E. (2009). Affect as a Psychological Primitive. *Advances in Experimental Social Psychology*, *41*, 167-218.

Barrett, L. F., Ochsner, K. N., & Gross, J. J. (2007). On the automaticity of emotion. In J. Bargh (Ed.), *Social psychology and the unconscious: The automaticity of higher mental processes*. New York: Psychology Press.

Barrett, L. F., Quigley, K. S., Bliss-Moreau, E., & Aronson, K. R. (2004). Interoceptive Sensitivity and Self-Reports of Emotional Experience. *Journal of Personality and Social Psychology*, *87*(5), 684-697.

Bechara, A., Damasio, A R, Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, *50*(1-3), 7-15.

Bechara, A., Damasio, H., Tranel, D, & Damasio, A R. (2005). The Iowa Gambling Task and the somatic marker hypothesis: some questions and answers. *Trends in Cognitive Sciences*, *9*(4), 159-162; discussion 162-164.

Bechara, A., Tranel, D, Damasio, H., & Damasio, A R. (1996). Failure to respond autonomically to anticipated future outcomes following damage to prefrontal cortex. *Cerebral Cortex (New York, N.Y.: 1991)*, *6*(2), 215-225.

Bermúdez, J. L. (2010). *Cognitive Science: An Introduction to the Science of the Mind*. Cambridge University Press.

Berridge, M. J. (1998). Neuronal calcium signaling. *Neuron*, *21*(1), 13-26.

Block, N. (2005). Review of Alva Noe, Action in Perception. *Journal of Philosophy*, *102*(5), 259-272.

Boulais, N., & Misery, L. (2008). The epidermis: a sensory tissue. *European Journal of Dermatology: EJD*, *18*(2), 119-127.

Brand, P. W., & Yancey, P. (1997). *The gift of pain: why we hurt & what we can do about it*. Zondervan.

Breiter, H. C., Etcoff, N. L., Whalen, P J, Kennedy, W. A., Rauch, S L, Buckner, R. L., Strauss, M. M., et al. (1996). Response and habituation of the human amygdala during visual processing of facial expression. *Neuron*, *17*(5), 875-887.

Broad, C. D. (1954). Emotion and sentiment. *The Journal of Aesthetics and Art Criticism*, *13*(2), 203-214.

Brooks, R. (1991). Intelligence without representation. *Artificial Intelligence*, *47*, 139-159.

Bullmore, E., & Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews. Neuroscience*, *10*(3), 186-198.

Burgess, N., Maguire, E. A., & O'Keefe, J. (2002). The human hippocampus and spatial and episodic memory. *Neuron*, *35*(4), 625-641.

Cannon, W. (1939). *Wisdom of the Body*. Norton.

Carnap, R. (1950). *Logical Foundations of Probability*. Chicago: University of Chicago Press.

Cartwright, N. (1989). *Nature's Capacities and Their Measurement*. Oxford: Oxford University Press.

Chen, C.-M., Lakatos, P., Shah, A. S., Mehta, A. D., Givre, S. J., Javitt, D. C., & Schroeder, C. E. (2007). Functional anatomy and interaction of fast and slow visual pathways in macaque monkeys. *Cerebral Cortex (New York, N.Y.: 1991)*, *17*(7), 1561-1569.

Clark, A. (1989). Microfunctionalism: Connectionism and the Scientific Explanation of Mental States. Research Paper, . Retrieved July 17, 2011, from http://www.era.lib.ed.ac.uk/handle/1842/1332
Clark, A. (1997). *Being There: Putting Brain, Body and World Together Again*. MIT Press.
Clark, A. (2001). *Mindware: An Introduction to the Philosophy of Cognitive Science*. OUP USA.
Clark, A. (2007). Choose your own reward. *Nature*, *445*(7129), 711-712.
Clark, A. (2008a). Pressing the flesh: A tension in the study of the embodied, embedded mind? *Philosophy and Phenomenological Research*, *LXXVI*(1), 37-59.
Clark, A. (2008b). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford Univ Pr.
Colombetti, G. (2005). Appraising Valence. *Journal of Consciousness Studies*, *12*(8-10), 103-126.
Colombetti, G. (2008). The somatic marker hypotheses, and what the Iowa Gambling Task does and does not show. *The British Journal for Philosophy of Science*, *59*(1), 51-57.
Colombetti, G. (2009). Enaction, sense-making and emotion. In J. Stewart, O. Gapenne, & E. Di Paolo (Eds.), *Enaction: towards a new paradigm for cognitive science*. Cambridge, Mass: MIT Press.
Colombetti, G., & Thompson, E. (2007). The feeling body: toward an enactive approach to emotion. In W. Overton, U. Mueller, & J. Newman (Eds.), *Body in mind, mind in body: developmental perspectives on embodiment and consciousness*. New Jersey: Lawrence Erlbaum Associates.
Craig, A. D. (2003a). A new view of pain as a homeostatic emotion. *Trends in Neurosciences*, *26*(6), 303-307.
Craig, A. D. (2003b). Interoception: the sense of the physiological condition of the body. *Current Opinion in Neurobiology*, *13*(4), 500-505.
Craig, A. D., Chen, K., Bandy, D., & Reiman, E. M. (2000). Thermosensory activation of insular cortex. *Nat Neurosci*, *3*(2), 184-190.
Cuppini, C., Ursino, M., Magosso, E., Rowland, B. A., & Stein, B. E. (2010). An emergent model of multisensory integration in superior colliculus neurons. *Frontiers in Integrative Neuroscience*, *4*.
Damaraju, E., Huang, Y.-M., Barrett, L. F., & Pessoa, L. (2009). Affective learning enhances activity and functional connectivity in early visual cortex. *Neuropsychologia*, *47*(12), 2480-2487.
Damasio, A R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *351*(1346), 1413-1420.
Damasio, A. (1999). *The feeling of what happens: body and emotion in the making of consciousness*. New York: Harcourt Brace.
Damasio, A. (2010). *Self Comes to Mind: Constructing the Conscious Brain*. Knopf Doubleday Publishing Group.
Damasio, Antonio R. (1994). *Descartes' error: emotion, reason, and the human brain*. New York: G.P. Putnam.
Davidson, R. J. (2003). Seven sins in the study of emotion: correctives from affective neuroscience. *Brain and Cognition*, *52*(1), 129-132.
Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876-879.
Denton, D. A., McKinley, M. J., Farrell, M., & Egan, G. F. (2009). The role of primordial emotions in the evolutionary origin of consciousness. *Consciousness and Cognition*, *18*(2), 500-514.
Dichter, G. S., Felder, J. N., & Bodfish, J. W. (2009). Autism is characterized by dorsal anterior cingulate hyperactivation during social target detection. *Social Cognitive and Affective Neuroscience*, *4*(3), 215-226.
Dretske, F. (1981). *Knowledge and the flow of information*. Cambridge, Mass.: MIT Press.
Dretske, F. (1988). *Explaining behavior*. Cambridge, Mass.: MIT Press.
Dubois, S., Rossion, B., Schiltz, C., Bodart, J. M., Michel, C., Bruyer, R., & Crommelinck, M. (1999). Effect of familiarity on the processing of human faces. *NeuroImage*, *9*(3), 278-289.
Duncan, S., & Barrett, L. F. (2007). Affect is a form of cognition: A neurobiological analysis. *Cognition & Emotion*, *21*(6), 1184-1211.
Dunn, B. D., Dalgleish, T., & Lawrence, A. D. (2006). The somatic marker hypothesis: a critical evaluation. *Neuroscience and Biobehavioral Reviews*, *30*(2), 239-271.
Edelman, G. M. (2004). *Wider than the sky: The phenomenal gift of consciousness*. London: Yale University Press.
Edelman, G. M., & Tononi, G. (2000). *Consciousness: how matter becomes imagination*. Allen Lane.
Egner, T. (2009). Prefrontal cortex and cognitive control: motivating functional hierarchies. *Nature Neuroscience*, *12*(7), 821-822.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*(2-3), 195-225.
Fecteau, J. H., & Munoz, D. P. (2006). Salience, relevance, and firing: a priority map for target selection. *Trends in Cognitive Sciences*, *10*(8), 382-390.
Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, *4*, 215.
Fellows, L. K., & Farah, M. J. (2003). Ventromedial frontal cortex mediates affective shifting in humans: evidence from a reversal learning paradigm. *Brain: A Journal of Neurology*, *126*(Pt 8), 1830-1837.
Fellows, L. K., & Farah, M. J. (2005). Different underlying impairments in decision-making following ventromedial and dorsolateral frontal lobe damage in humans. *Cerebral Cortex*, *15*(1), 58-63.
Fodor, J. A. (1983). *Modularity of mind: an essay on faculty psychology*. Cambridge, Mass.: MIT Press.
Fogarty, S J, & Hemsley, D. R. (1983). Depression and the accessibility of memories. A longitudinal study. *The British Journal of Psychiatry*, *142*(3), 232 -237.
Fox, E. (2008). *Emotion Science: an integration of cognitive and neuroscientific approaches*. New York: Palgrave Macmillan.
Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293-301.
Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*(1521), 1211-1221.
Gerrans, P. (2007). Mental time travel, somatic markers and "myopia for the future." *Synthese*, *159*(3), 459-474.
Goldin-Meadow, S. (2003). *Hearing Gesture: How our hands help us think*. Cambridge, Mass.: Harvard University Press.
Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*(1), 20-25.
Grahek, N. (2007). *Feeling pain and being in pain*. MIT Press.
Griffiths, P. (1997). *What Emotions Really are: The Problem of Psychological Categories*. University of Chicago Press.
Griffiths, P., & Scarantino, A. (2005). Emotions in the wild. In P. Robbins & M. Aydede (Eds.), *The Cambridge Handbook of Situated Cognition*. Cambridge University Press.
Gutbrod, K., Krouzel, C., Hofer, H., Müri, R., Perrig, W., & Ptak, R. (2006). Decision-making in amnesia: do advantageous decisions require conscious knowledge of previous behavioural choices? *Neuropsychologia*, *44*(8), 1315-1324.
Harre, P. R. (2002). *Cognitive Science: A Philosophical Introduction*. Sage Publications Ltd.
Heims, H. C., Critchley, H. D., Dolan, R., Mathias, C. J., & Cipolotti, L. (2004). Social and motivational functioning is not critically dependent on feedback of autonomic responses: neuropsychological evidence from patients with pure autonomic failure. *Neuropsychologia*, *42*(14), 1979-1988.
Herry, C., Bach, D. R., Esposito, F., Di Salle, F., Perrig, W. J., Scheffler, K., Lüthi, A., et al. (2007). Processing of temporal unpredictability in human and animal amygdala. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *27*(22), 5958-5966.
Holland, & Gallagher. (1999). Amygdala circuitry in attentional and representational processes. *Trends in Cognitive Sciences*, *3*(2), 65-73.
Holtzheimer, P. E., & Mayberg, H. S. (2011). Stuck in a rut: rethinking depression and its treatment. *Trends in Neurosciences*, *34*(1), 1-9.
Hsu, M., Bhatt, M., Adolphs, R., Tranel, Daniel, & Camerer, C. F. (2005). Neural systems responding to degrees of uncertainty in human decision-making. *Science (New York, N.Y.)*, *310*(5754), 1680-1683.
Hurley, S. (2010). The varieties of externalism. In R. Menary (Ed.), *The Extended Mind* (pp. 101-154). MIT Press.
Husbands, P. (1998). Evolving robot behaviours with diffusing gas networks. In Philip Husbands & J.-A. Meyer (Eds.), *Evolutionary Robotics* (Vol. 1468, pp. 71-86). Berlin, Heidelberg: Springer Berlin Heidelberg.
James, W. (1884). What is an emotion? *Mind*, *9*, 188-201.
James, W. (1890). *The principles of psychology*. New York, NY, US: Henry Holt and Co, Inc.
James, W. (1902). *The Varieties of Religious Experience: A study in human nature*. Bombay, New York: Longmans, Green & co.
Kandel, E. R., Schwartz, J. H., & Jessell, T. M. (2000). *Principles of neural science*. McGraw-Hill, Health Professions Division.

Kastin, A. J., & Pan, W. (2010). Concepts for biologically active peptides. *Current Pharmaceutical Design*, *16*(30), 3390-3400.
Kenny, A. (1963). *Action, emotion & Will*. London: Routledge & Kegan Paul.
Klein, S. B., & Loftus, J. (2002). Memory and temporal experience: The effects of episodic memory loss on an amnesic patient's ability to rememberrthe past and imagine the future. *Social Cognition*, *20*, 353-379.
Koechlin, E., & Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences*, *11*(6), 229-235.
Koechlin, E., Ody, C., & Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. *Science (New York, N.Y.)*, *302*(5648), 1181-1185.
Kouneiher, F., Charron, S., & Koechlin, E. (2009). Motivation and cognitive control in the human prefrontal cortex. *Nature Neuroscience*, *12*(7), 939-945.
Lakoff, G. (1987). *Women, Fire, and Dangerous Things*. Chicago: University of Chicago Press.
Lakoff, G., & Johnson, M. (1980). *Metaphors We Live by*. Chicago: University of Chicago Press.
Laycock, R., Crewther, S. G., & Crewther, D. P. (2007). A role for the "magnocellular advantage" in visual impairments in neurodevelopmental and psychiatric disorders. *Neuroscience and Biobehavioral Reviews*, *31*(3), 363-376.
Lazarus, R. (1991). *Emotion and adaptation*. New York: Oxford University Press.
LeDoux, J. (1996). *The emotional brain: the mysterious underpinnings of emotional life*. Simon & Schuster.
Legrand, D. (2007). Pre-reflective self-as-subject from experiential and empirical perspectives. *Consciousness and Cognition*, *16*(3), 583-599.
Levin, B. E., Magnan, C., Dunn-Meynell, A., & Le Foll, C. (2011). Metabolic sensing and the brain: who, what, where, and how? *Endocrinology*, *152*(7), 2552-2557.
Lewis, M. D., & Todd, R. M. (2005). Getting emotional: A neural perspective on emotion, intention, and consciousness. *Journal of Consciousness Studies*, *12*, 210-235.
Lewis, M. D., & Todd, R. M. (2007). The self-regulating brain: Cortical-subcortical feedback and the development of intelligent action. *Cognitive Development*, *22*, 406-430.
Li, C., & Kim, K. (2008). Neuropeptides. In The C. elegans Research Community, Wormbook (Ed.), *WormBook* (pp. 1-36). Retrieved from http://www.wormbook.org/chapters/www_neuropeptides/neuropeptides.html
Lind, S. E., & Bowler, D. M. (2010). Episodic memory and episodic future thinking in adults with autism. *Journal of Abnormal Psychology*, *119*(4), 896-905.
Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., & Barrett, L. F. (Forthcoming). The brain basis of emotion: A meta-analytic review. *Behavioral and Brain Sciences*.
MacLean, P. D., & Kral, V. A. (1973). *A triune concept of the brain and behaviour*. published for the Ontario Mental Health Foundation by University of Toronto Press.
Maia, T. V., & McClelland, J. L. (2004). A reexamination of the evidence for the somatic marker hypothesis: what participants really know in the Iowa gambling task. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(45), 16075-16080.
Maia, T. V., & McClelland, J. L. (2005). The somatic marker hypothesis: still many questions but no answers. *Trends in Cognitive Sciences*, *9*(4), 162-170.
Mandik. (2002). Selective Representing and World-Making. *MINDS AND MACHINES*, *12*(3), 383.
Mandik, P. (2005). Action-Oriented Representations. In A. Brook & K. Akins (Eds.), *Cognition and the brain: the philosophy and neuroscience movement*. Cambridge University Press.
Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H.Freeman & Co Ltd.
McAbee, G. N., Chan, A., & Erde, E. L. (2000). Prolonged survival with hydranencephaly: report of two patients and literature review. *Pediatric Neurology*, *23*(1), 80-84.
Merigan, W. H., & Maunsell, J. H. (1993). How parallel are the primate visual pathways? *Annual Review of Neuroscience*, *16*, 369-402.
Mesulam, M. (2000). Behavioral neuroanatomy: large-scale networks, association cortex, frontal syndromes, the limbic system, and hemispheric specializations. In M. Mesulam (Ed.), *Principles of behavioral and cognitive neurology* (2nd ed., pp. 1-120). New York, NY: Oxford University Press.
Milner, A. D., & Goodale, M. A. (1995). *The Visual Brain in Action* (illustrated edition.). Oxford University Press.
Minshew, N. J., & Keller, T. A. (2010). The nature of brain dysfunction in autism: functional brain imaging studies. *Current Opinion in Neurology*, *23*(2), 124-130.
Montague, R. (2006). *Why choose this book?: how we make decisions*. Dutton.

Nummela, S. U., & Krauzlis, R. J. (2011). Superior colliculus inactivation alters the weighted integration of visual stimuli. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *31*(22), 8059-8066.
Oliva, A., & Torralba, A. (2001). Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, *42*, 145–175.
Owren, M. J., & Rendall, D. (1997). An affect-conditioning modelof non-human primate vocal signaling. In D. H. Owings, M. D. Beecher, & N. S. Thompson (Eds.), *Communication*. Springer.
Owren, M. J., & Rendall, D. (2001). Sound on the rebound: bringing form and function back to the forefront in understanding nonhuman primate vocal signaling. *Evolutionary Anthropology*, *10*(2), 58-71.
Paemeleire, K. (2002). Calcium signaling in and between brain astrocytes and endothelial cells. *Acta Neurologica Belgica*, *102*(3), 137-140.
Panksepp, J. (2004). *Affective Neuroscience: The Foundations of Human and Animal Emotions*. OUP USA.
Panksepp, J., Fuchs, T., Garcia, V. A., & Lesiak, A. (2007). Does any aspect of mind survive brain damage that typically leads to a persistent vegetative state? Ethical considerations. *Philosophy, Ethics, and Humanities in Medicine: PEHM*, *2*, 32.
Di Paolo, A. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences*, *4*, 429--452.
Papez, J. W. (1937). A proposed mechansim of emotion. *Arch Neurol Psychiatry*, *38*(4), 725-743.
Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature Reviews. Neuroscience*, *9*(2), 148-158.
Pessoa, L. (2010). Emotion and cognition and the amygdala: from "what is it?" to "what's to be done?" *Neuropsychologia*, *48*(12), 3416-3429.
Pessoa, L., & Adolphs, R. (2010). Emotion processing and the amygdala: from a "low road" to "many roads" of evaluating biological significance. *Nature Reviews. Neuroscience*, *11*(11), 773-783.
Pfeifer, R., & Scheier, C. (2001). *Understanding Intelligence*. MIT Press.
Philippides, Andrew, Husbands, Phil, & O'Shea, M. (2000). Four-Dimensional Neuronal Signaling by Nitric Oxide: A Computational Analysis. *The Journal of Neuroscience*, *20*(3), 1199 -1207.
Philippides, Andy, Husbands, Phil, Smith, T., & O'Shea, M. (2005). Flexible couplings: diffusing neuromodulators and adaptive robotics. *Artificial Life*, *11*(1-2), 139-160.
Piaget, J. (1963). *The Origins of Intelligence in Children*. W.W. Norton.
Ploner, M., Freund, H. J., & Schnitzler, A. (1999). Pain affect without pain sensation in a patient with a postcentral lesion. *Pain*, *81*(1-2), 211-214.
Prather, M. D., Lavenex, P., Mauldin-Jourdain, M. L., Mason, W. A., Capitanio, J. P., Mendoza, S. P., & Amaral, D. G. (2001). Increased social fear and decreased fear of objects in monkeys with neonatal amygdala lesions. *Neuroscience*, *106*(4), 653-658.
Prinz, J. (2004). Emotions embodied. *Thinking about feeling*. New York: Oxford University Press.
Prinz, J. J. (2004). *Gut Reactions: A Perceptual Theory of Emotion* (illustrated edition.). Oxford University Press Inc.
Ratcliffe, M. (2005). William James on emotion and intentionality. *International Journal of Philosophical Studies*, *13*(2), 179.
Ratcliffe, M. (2008). *Feelings of Being: Phenomenology, psychiatry and the sense of reality* (1st ed.). OUP Oxford.
Ratcliffe, M. (2009). Existential Feeling and Psychopathology (and response to commentaries "Belonging to the World through the Feeling Body"'). *Philosophy, Psychiatry & Psychology*, *16*(2), 179-211.
Ratcliffe, M. (2010). The Phenomenology of Mood and the Meaning of Life. In P. Goldie (Ed.), *Handbook of Philosophy of Emotion* (pp. 349-371). Oxford: Oxford University Press.
Rolls, E. T. (1999). *The Brain and Emotion*. New York: Oxford University Press.
Rolls, E. T., Hornak, J., Wade, D., & McGrath, J. (1994). Emotion-related learning in patients with social and emotional changes associated with frontal lobe damage. *Journal of Neurology, Neurosurgery, and Psychiatry*, *57*(12), 1518-1524.
Ross, D., & Ladyman, J. (2010). The alleged coupling-constitution fallacy and the mature sciences. In R. Menary (Ed.), *The Extended Mind* (pp. 155-166). MIT Press.
Sander, D., Grafman, J., & Zalla, T. (2003). The human amygdala: an evolved system for relevance detection. *Reviews in the Neurosciences*, *14*(4), 303-316.
Scarantino, A. (2005, October 10). *Explicating Emotions* (PhD). University of Pittsburgh.
Schachter, S., & Singer, J. E. (n.d.). Cognitive, social, and physiological determinants of emotional states. *Psychological Review*, *69*, 379-399.

Schenk, T. (2010). Visuomotor robustness is based on integration not segregation. *Vision Research*, *50*(24), 2627-2632.
Scherer, K. R. (1999). Appraisal Theory. In T. Dalgleish & M. Power (Eds.), *Handbook of emotion and cognition* (pp. 637-663). Wiley.
Schoenbaum, G., Nugent, S. L., Saddoris, M. P., & Setlow, B. (2002). Orbitofrontal lesions in rats impair reversal but not acquisition of go, no-go odor discriminations. *Neuroreport*, *13*(6), 885-890.
Schwartz, C. E., Wright, C. I., Shin, L. M., Kagan, J., Whalen, Paul J, McMullin, K. G., & Rauch, Scott L. (2003). Differential amygdalar response to novel versus newly familiar neutral faces: a functional MRI probe developed for studying inhibited temperament. *Biological Psychiatry*, *53*(10), 854-862.
Semba, K. (2000). Multiple output pathways of the basal forebrain: organization, chemical heterogeneity, and roles in vigilance. *Behavioural Brain Research*, *115*(2), 117-141.
Shannon, C. E., & Weaver, W. (1963). *The mathematical theory of communication*. Illinois: University of Illinois Press.
Sherrington, C. (1906). *The Integrative Action of the Nervous System*. New Haven, CT: Yale University Press.
Shewmon, D., Holmes, G., & Byrne, P. (1999). Consciousness in congenitally decorticate children: developmental vegetative state as self-fulfilling prophecy. *Dev Med Child Neurol*., *41*(6), 364-74.
Smith, L. B., & Thelen, E. (2003). Development as a dynamic system. *Trends in Cognitive Sciences*, *7*(8), 343-348.
Smith, T. (2002). *The evolvability of artificial neural networks for robot control*. University of Sussex, School of Biological Sciences.
Smith, T., Husbands, Phil, Philippides, Andy, & O'Shea, M. (2002). Neuronal Plasticity and Temporal Adaptivity: GasNet Robot Control Networks. *Adaptive Behavior*, *10*(3-4), 161 -183.
Sober, E. (1988). Apportioning causal responsibility. *Journal of Philosophy*, *85*(6), 303-318.
Solomon, R. C. (1976). *The passions: emotions and the meaning of life*. Hackett Publishing.
Soussignan, R. (2002). Duchenne smile, emotional experience, and autonomic reactivity: A test of the facial feedback hypothesis. *Emotion*, *2*(1), 52-74.
Sporns, O. (2010). *Networks of the Brain* (1st ed.). MIT Press.
Sporns, O., & Zwi, J. D. (2004). The small world of the cerebral cortex. *Neuroinformatics*, *2*(2), 145-162.
Sporns, O., Chialvo, D. R., Kaiser, M., & Hilgetag, C. C. (2004). Organization, development and function of complex brain networks. *Trends in Cognitive Sciences*, *8*(9), 418-425.
Sterelny, K. (1990). *The Representational Theory of Mind: An Introduction*. Wiley-Blackwell.
Stewart, J., Gapenne, O., & Paolo, E. A. D. (2011). *Enaction: Toward a New Paradigm for Cognitive Science*. MIT Press.
Teasdale, J. D., Taylor, R., & Fogarty, Sarah J. (1980). Effects of induced elation-depression on the accessibility of memories of happy and unhappy experiences. *Behaviour Research and Therapy*, *18*(4), 339-346.
Thakkar, K. N., Polli, F. E., Joseph, R. M., Tuch, D. S., Hadjikhani, N., Barton, J. J. S., & Manoach, D. S. (2008). Response monitoring, repetitive behaviour and anterior cingulate abnormalities in autism spectrum disorders (ASD). *Brain: A Journal of Neurology*, *131*(Pt 9), 2464-2478.
Thompson, A. (1995). Evolving electronic robot controllers that exploit hardware resources. *Advances in Artifical Life: Lecture Notes in Computer Science*, *929*, 640-656.
Thompson, A. (1997). Artificial Evolution in the Physical World. In T. Gomi (Ed.), *Evolutionary Robotics: From Intelligent Robots to Artificial Life (ER'97)* (p. 101--125). AAI Books.
Thompson, E. (2007a). *Mind in life: biology, phenomenology, and the sciences of mind*. Harvard University Press.
Thompson, E. (2007b). *Mind in life: biology, phenomenology, and the sciences of mind*. Harvard University Press.
Von Uexkull, J. (1934). A stroll through the worlds of animals and men. In K. Lashley (Ed.), *Instinctive Behavior*. International Universities Press.
Varela, F. J., Rosch, E., & Thompson, E. (1991). *The embodied mind: cognitive science and human experience / Francisco J. Varela, Evan Thompson, Eleanor Rosch*. Cambridge, Mass: MIT Press.
Wallis, J. D., & Miller, E. K. (2003). Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task. *The European Journal of Neuroscience*, *18*(7), 2069-2081.
Wang, R. (2002). Two's company, three's a crowd: can H2S be the third endogenous gaseous transmitter? *The FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology*, *16*(13), 1792-1798.

Wedig, M. M., Rauch, Scott L, Albert, M. S., & Wright, C. I. (2005). Differential amygdala habituation to neutral faces in young and elderly adults. *Neuroscience Letters*, *385*(2), 114-119.
Whalen, P J. (1998). Fear, vigilance, and ambiguity: Initial neuroimaging studies of the human amygdala. *Current Directions in Psychological Science*, *7*, 177-188.
Wheeler, M. A., Stuss, D. T., & Tulving, E. (1997). Toward a theory of episodic memory: the frontal lobes and autonoetic consciousness. *Psychological Bulletin*, *121*(3), 331-354.
Wilson, F. A., & Rolls, E. T. (1990). Neuronal responses related to reinforcement in the primate basal forebrain. *Brain Research*, *509*(2), 213-231.
Wundt, W. M. (1897). *Outlines of Psychology*.
Zikopoulos, B., & Barbas, H. (2007). Circuits for multisensory integration and attentional modulation through the prefrontal cortex and the thalamic reticular nucleus in primates. *Reviews in the Neurosciences*, *18*(6), 417-438.