THE UNIVERSITY
of EDINBURGH

The Entangled Predictive Brain:
Emotion, Prediction and Embodied Cognition

Mark Miller

PhD in Philosophy
The University of Edinburgh
2018

Declaration of Authorship

I, Mark Miller, declare that the thesis has been composed by myself and that the work has not been submitted for any other degree or professional qualification. I confirm that the work submitted is my own, except where work which has formed part of jointly-authored publications has been included. My contribution and those of the other authors to this work have been explicitly indicated below (described in "Author's Contributions", p. 21). I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others.

Mark Miller, 12th July 2018

Abstract/Lay Summary

How does the living body impact, and perhaps even help constitute, the thinking, reasoning, feeling agent? This is the guiding question that the following work seeks to answer. The subtitle of this project is emotion, prediction and embodied cognition for good reason: these are the three closely related themes that tie together the various chapters of the following thesis. The central claim is that a better understanding of the nature of emotion offers valuable insight for understanding the nature of the so called 'predictive mind', including a powerful new way to think about the mind as embodied

Recently a new perspective has arguably taken the pole position in both philosophy of mind and the cognitive sciences when it comes to discussing the nature of mind. This framework takes the brain to be a probabilistic prediction engine. Such engines, so the framework proposes, are dedicated to the task of minimizing the disparity between how they expect the world to be and how the world actually is. Part of the power of the framework is the elegant suggestion that much of what we take to be central to human intelligence - perception, action, emotion, learning and language - can be understood within the framework of prediction and error reduction. In what follows I will refer to this general approach to understanding the mind and brain as 'predictive processing'.

While the predictive processing framework is in many ways revolutionary, there is a tendency for researchers interested in this topic to assume a very traditional 'neurocentric' stance concerning the mind. I argue that this neurocentric stance is completely optional, and that a focus on emotional processing provides good reasons to think that the predictive mind is also a deeply embodied mind. The result is a way of understanding the predictive brain that allows the body and the surrounding environment to make a robust constitutive contribution to the predictive process. While it's true that predictive models can get us a long way in making sense of what drives the neural-economy, I will argue that a complete picture of human intelligence requires us to also explore the many ways that a predictive brain is embodied in a living body and embedded in the social-cultural world in which it was born and lives.

Contents

Preface

How does the living body impact, and perhaps even help constitute, the thinking, reasoning, feeling agent? This is the guiding question that the following work seeks to answer. The subtitle of this project is emotion, prediction and embodied cognition for a good reason: these are the three closely related themes that tie together the various chapters of the following thesis. The central claim is that a better understanding of the nature of emotion offers valuable insights for understanding the nature of the 'predictive mind', including a powerful new way to think about the mind as embodied.

Broadly, a cognitive process is embodied if it is deeply dependent upon aspects of the physical body outside of the brain, that is, parts of the body contribute to cognition in richly causal or even constitutive ways (Wilson & Foglia 2011). This project will extend this definition beyond the external physical body, to include a full living system that involves the hotter and bloodier dimension of the inner body (including the viscera, vascular, immune, and endocrine systems). Living embodiment of this kind is increasingly implicated in accounts of emotion and cognition (see Colombetti 2015; Stapleton 2013; Thompson & Cosmelli 2011; Damasio 2010).

General interest in researching emotion is at a high. Various fields of inquiry including philosophy, psychology, cognitive science and robotics have all begun to inquire into the role emotions might play in the functioning of the mind. This burst of popularity is due in part to advancements in our understanding of neural architecture. I will draw on recent neuroscience of emotion as a means of supporting some of the central claims in embodied ('full living body' style) cognitive science. Emotion is an ideal topic for those interested in embodying the mind. In an important sense emotions are rooted in both brain and bodily processes. One way to see emotion research and embodied cognitive science coming together is in recent network theories of the brain that are changing the way we think about the traditional ontological assumptions we have about emotion and cognition as easily separable processes (Pessoa 2017). In Chapter 1 I build on this research as a means of supporting some of the core ideas in embodied cognitive science.

In Chapters 2 through 6 I apply this line of thinking specifically to a new and exciting view of the nature of mind and cognition, which takes the brain to be a probabilistic prediction engine (Friston 2010; Howhy 2013; Clark 2013). Such engines, so the framework proposes, are dedicated to the task of minimizing the disparity between how they expect the world to be and how the world actually is. Part of the power of the framework is the elegant suggestion that much of what we take to be central to human intelligence - perception, action, emotion, learning and language - can be modelled (and so at least partially understood) within the framework of prediction and error reduction (Clark 2013: 181). In what follows, I will refer to this general approach to understanding the mind and brain as 'predictive processing'.

The predictive processing framework is in many ways revolutionary. However, there is a tendency for researchers interested in this topic to assume a very traditional 'neurocentric' stance concerning the mind. I argue that this neurocentric stance is completely optional, and that a focus on emotional processing provides good reasons to think that the predictive mind is also a deeply embodied (in the 'situated living organization' sense described above) mind. The result is a way of understanding the predictive brain that allows the body and the surrounding environment to make a robust constitutive contribution to the predictive process. While it's true that predictive models can get us a long way in making sense of what drives the neural-economy, I will argue that a complete picture of human intelligence requires us to also explore the many ways that a predictive brain is viscerally embodied in a living body and embedded in the social-cultural world in which it was born and lives.

Introduction

In order to offer important context to the arguments that follow, I will briefly introduce some of the relevant core concepts. First, I will briefly introduce the notion of embodied cognitive science. Second, I will outline how I see recent neuroscience of emotion offering support for the embodied cognitive science paradigm. Third, I will introduce the predictive processing framework - the primary model of cognition and mind I will be discussing throughout this thesis. Fourth, I will outline how different camps within the predictive processing framework are viewing its amenability to embodied cognitive science. Fifth and sixth, I will lay the groundwork for my specific line of argumentation that emotion research offers a more substantial bridge between predictive processing and embodied cognitive science. I conclude this introduction with a brief synopsis of each of the chapters.

Embodied cognitive science

Cognitive science and philosophy of mind have long made happy bedfellows. Empirically interested philosophers have learned to harness the cognitive sciences to help develop and test theories about the nature of the mind. In return, philosophers have developed broad theoretical frameworks for making sense of the empirical data. Of the many shared interests between philosophy and cognitive science perhaps the most exciting today is the notion that the mind may be embodied. That is, that cognitive function may depend, in part, on organismic resources found outside of the central nervous system.

Traditionally cognitive science defended a rather narrow view of cognition. The central nervous system was assumed to be the sole locus of mental activity. The body, when it was discussed at all, was imagined to be simply a vehicle for perceptual inputs and behavioural outputs to interface with the brain. As Shapiro writes, concerning traditional cognitive science, "[b]ecause

cognition begins and ends with inputs to and outputs from the nervous system, it has no need for interaction with the real world outside it" (Shapiro 2007: 339).[1]

Embodied cognitive science (ECS), by contrast, takes the position that cognition cannot be understood without also taking into consideration wider embodied and environmental dynamics. ECS is composed of a loose-knit collection of ideas about how the body and environment may make this contribution to cognition (see Brooks 1991; Beer 2003; Clark 1997; Kelso 1995; Noë 2004; Thelen & Smith 1994; Thelen et al. 2001; Thompson 2007; Thompson & Varela 2001; Maturana & Varela 1991).

One way researchers have argued for embodied cognition is by looking at how the body's structure and function has evolved in ways that help reduce the computational costs of cognizing. For example, an animal who has evolved two eyes set at a certain distance apart, and the ability to turn its head, has a computationally cheaper method for acquiring more accurate depth perception compared to, say, a single eyed fixed-headed version of itself (Churchland et al. 1994). In other words, the two eyed creature's body does some of its depth-perception for free.

Another popular approach to supporting ECS is a close analysis of the relationship between perception and action. A growing literature suggests that our perceptual abilities are built upon our behavioural abilities (Churchland et al. 1994; Clark 1997; Hurley 1998; O'Regan & Noë 2001, 2004). Hurley's (1998; 2008) work on "sensorimotor dynamics" is useful to consider here. Hurley argues that perception and action each emerge from a collection of subpersonal processes, many of which are shared between the two. Due to this integration activity in the motor system is able to directly impact perceptual processing and vice versa (see Hurley 1998: 363–364; see also O'Regan & Noë 2001). Hurley characterizes the relationship between perception and action as "constitutively interdependent": i.e., they interact with one another in such rich and reciprocal ways that an easy separation into distinct categories becomes impossible. What we perceive and how we behave are controlled by a single complex system: action constitutes perception, and perception is a form of action. This rich sensorimotor entanglement exposes a deep continuity between cognition and the body.

[1] A common image arising from early cognitive science was the 'mind as computer'. The central nervous system was likened to the central processing unit (CPU), with the brain and transducer systems translating sensory stimulation into chunky, manipulatable symbols that could be operated upon by the CPU. Some of those symbols could in turn be translated into behaviours. Cognition here is a highly insular ('neurocentric') affair, taking place entirely within the CPU-brain.

In both examples cognitive processing is described as being offloaded onto elements of the non-neural body. The mind from this vantage is not some immaterial Cartesian substance, nor is the brain considered the mind's minimally sufficient physical basis. Rather, cognition is reimagined as emerging from a much wider network of interactions that take place between the brain and body.

While considerations about gross morphology and sensorimotor dynamics go some way to refuting the neurocentric assumptions of early cognitive science, some researchers remain dissatisfied with these approaches.[2] These contributions may only offer a 'moderate embodiment': the body is described as helping to reduce cognitive costs, but in a way that maintains a strong separation between the brain as locus of cognition and body as support. A central aim of this thesis is the exploration of a new and potentially more complete approach to embodying the mind. I will argue that emotion research (specifically, neuroscience) can offer a more substantial view of the mind as embodied. While the study of emotion should be an obvious topic for ECS, as emotions are often described as composites of both brain and bodily elements, until very recently ECS has remained almost entirely silent about emotion. Emotion theorists in return have by and large failed to incorporate the lessons taught by ECS. I will argue that the body's most important contribution to cognition is not to be found in the external or superficial aspects of the body, but rather in the wetter, hotter and bloodier dimension of the inner body which includes the viscera, vascular, immune, and endocrine systems (see Craig 2002; Damasio 2010; Stapleton 2013; Colombetti 2014; Thompson & Cosmelli 2011). I will introduce my argument for embodying the mind through emotion in the following section.

Emotion-cognition entanglements

Philosophers and scientists have long been interested in understanding the relationship between emotion and cognition. Historically, emotion and cognition were viewed as separable, and often even opposing, processes (Aristotle 1991; Plato 1992). While researchers at the end of the 19th century, such as Darwin and James, had begun to recognize the important role that emotions play in the mind, their 20th century counterparts tended to view emotion and cognition as antithetical concepts. However, there has been a tremendous surge of interest in emotion within the cognitive sciences over the last two decades. Among the more provocative challenges to traditional doctrine is the growing body of evidence that the long assumed ontological boundaries separating emotion from cognition may be misguided (see Damasio 1994; LeDoux 1996; Damasio 1999; Rolls 2005; Phelps 2006).

[2] For good examples see Stapleton (2013) and Colombetti (2014).

Recent network models of the brain provide compelling evidence that these folk divisions do not map onto functionally segregated brain areas (Pessoa 2013, 2017; Lindquist & Barrett 2012; Parvizi 2009). Emerging from this work is a strong argument for seeing emotion and cognition not only as merely interacting, but as functionally interdependent and dynamically co-evolving (see in detail in Chapters 1, 2 and 3). In many ways, this is unsurprising. A tight interdependency between cognition and emotion is what allowed our animal ancestors to succeed in situated action control. For emotion and motivation to support fast adaptive behaviour these processes must be able to influence the rest the system in a timely manner. In our world, situations change in important ways from moment to moment: our level of energy, what opportunities the situation offers, and how we are faring in the task all matter to our survival. The necessity for animals to be extremely sensitivity to this internal bodily information should make us suspicious of a strong segregation between emotion, motivation and cognition (Cisek & Pastor-Bernier 2014; Lepora & Pezzulo 2015; Verschure et al. 2014).

Network models offer an account of the neural architecture that makes this smooth and ongoing influence possible. Where traditional models of the brain tended to assume that connections between discrete brain areas to be relatively simple and linear, network models describe the brain as composed of tightly organized neural-clusters, whereby every brain area is densely and reciprocally connected to every other area, and pathways leading from anywhere to anywhere are relatively short (Sporns 2006; Sporns et al. 2004; Sporns & Zwi 2004). For example, research now shows that every cortical area is either directly connected, or by way of only one or two intermediate steps, to every other cortical area (Hilgetag et al. 2000; Sporns et al. 2000). This rich interaction between brain regions has shifted the emphasis away from trying to understand brain function one area or region at a time. Instead, researchers are now investigating how coalitions or networks of different regions work together to support cognition and behavior.

I argue that this interconnection, interaction and mutual influence among components (or neural regions) motivates a view of cognitive and emotional processes as functionally inseparable within the brain (see in detail in Chapters 1, 2 and 3; see also work done in Colombetti 2014; Pessoa 2013; Lewis 2005). The tight coordination between cortex and subcortex (which is in turn tightly coordinated with information from within the organism's body) ensures that emotional and motivational value are continually folded into cognition and behavior (Pessoa 2009, 2015, 2017). For example, due to the rich reciprocal connections between the visual cortex and the subcortical pulvinar and amygdala (both widely linked to affective significance), vision is inevitably processed within an affective context. The result is that that can be no such thing as

pure vision (Churchland et al. 1994) - all vision will inevitably be *affective vision* (Pessoa 2015: 257). From this perspective, perception, action and cognition are more than merely biased by emotion and motivation: they are thoroughly, and inextricably, emotional and motivational in nature. I will argue that emotion and cognition should be thought of as "constitutively interdependent" in much the same way that Hurley proposes about action and perception (for other similar arguments see Colombetti 2014). In the process of generating intelligent behaviour, emotional and cognitive processes become so intermixed that a functional separation becomes impossible (Lewis 2005; Lewis & Todd 2007). This entangled architecture poses a significant challenge to the traditional ontological boundaries between thinking, perceiving, feeling and acting.

Building on this view of cognitive and emotional processes as non-decomposable in the brain, I make the argument that the larger brain-body-environment system is also non-decomposable. Following arguments developed by Pessoa (2014), I suggest that structure-function mappings of the networks themselves are highly dynamic (see in detail in Chapter 1). In other words, the function that a given network performs will vary over time in a highly flexible and context-dependent manner. It is this latter finding which I take to support the non-decomposability of the brain-body-environment system. Network accounts propose that the function of any particular neural network can only be understood if the wider 'global variables' that constrain its functioning are taken into consideration, for example, neurotransmitters, bodily arousal and slow-wave potentials (Pessoa 2014: 408). The fact that neurotransmitter and arousal systems are capable of influencing the organization of brain networks invites the question as to why this boundary between the brain and body should be considered as a sort of "magical membrane" (Hurley, 2010)?. Why should we think that the factors capable of influencing the function of a network will reside only inside of the brain? I will propose instead that we think of bodily states of arousal and action-readiness (that manifest as changes in the body's vascular, visceral and motor systems) as global control parameters that influence the formation of large-scale networks in the brain. To determine the precise functional contribution a network is making to behavior requires zooming out, and having in view the whole organism in its interaction with the environment. Emotional-cognitive processes don't only take place inside of brains, but are processes that involve constant interaction between the brain and the whole living body of the organism in an ecological setting.

Predictive processing

An increasingly popular theory in philosophy and neuroscience is that the brain can be aptly thought of an engine of prediction (Bubic et al. 2010; Friston 2010; Howhy 2013; Clark 2013). This engine, so the story goes, makes sense of the world by actively predicting what will happen next. As it turns out, much of what we take to be central to human intelligence - perception, action, emotion, learning and language - can be understood as emerging from this prediction process (Clark 2013: 181).

This so called 'predictive processing' (PP) story begins with a brain that has evolved to help make understanding and responding to the world possible. How does that brain learn about the complex world of objects and events when all it seems to have available to it are the relatively abstract sensory signals? This challenge has been referred to as the 'problem of perception' (Hohwy 2013). To make the problem more pronounced, consider the fact that any single object can induce a variety of sensory patterns: the same object can be encountered from different angles or in changing atmospheric conditions. Moreover, many different objects produce similar sensory signals: a picture of an object and the object itself, or a partially obscured object and a fragmented one (e.g. a cat walking behind a picket fence). To make matters worse, we must also be able to explain how the brain is able to separate out the salient information from the unimportant sensory noise. The PP model 'earns its salt' by offering an elegant solution to this challenge.

In extremely brief terms, the predictive brain gets a grip on the world by continually learning about the vast network of temporal-spatial regularities that reflect, and in some ways constitute, its environment[3]. It uses this knowledge base to make increasingly refined predictions about what objects and events are most likely responsible for the signals it receives from the environment. When these top-down predictions match the bottom-up signals well enough, a perceptual experience occurs. When it doesn't, the discrepancy between the prediction and the actual signal (referred to as "prediction errors") either moves forward through the system helping to refine future predictions (e.g. the hose alongside the house isn't a snake), or it provokes actions aimed at making the sensory stream better fit the predictions (e.g. placing my coffee cup back in its regular place on my desk). Finally, this entire prediction error minimizing

[3] Our world is filled with learnable regularities - both natural and synthetic, ranging from the very fast (e.g. patterns in a swiftly moving river) to very slow (e.g. the river eroding the land around it). Through tracking fast regularities we grasp the fine-grained details about a scene. Tracking slower regularities provide the kind of abstracted and enduring information that can help contextualize more local predictions, and help with more long term predictions. Moreover there are obvious interrelatedness between these levels. For example, the small tidal movements and eddies of the river are key factors in how the shoreline will be sculpted over many years, in return the contours of the slowly shifting shoreline is what produces and contains the tidal movements. Tracking one would give clues about the other, and vice versa. See Hohwy (2013) for a more complete discussion about this point.

routine is further tuned by a set of second-order expectations that track the reliability of the predictive systems own estimates given the state of the organism and the current context. It uses this estimation of reliability (referred to as 'precision weighing') to flexibly adjust the gain (like turning the volume up) on particular error units which increases the impact they have on the unfolding process. This allows the system to both highlight specific bodies of information, and to modify the degree to which it relies on bottom-up or top-down information (see Friston 2009, 2010). For example, while listening to your favourite song in the shower it would be useful to turn down the influence on the error units produced by the flowing water, and rely more on our clear memories of the song (Clark 2016: 92).

The novel addition this theory makes to traditional, feedforward-dominated perception research is that perception is not explained by incoming signals alone, but crucially also includes our top-down predictions about the shape of those signals and what they could mean. Perception here turns out to be something like a complex controlled hallucination: we see what we believe we see. Of course what we believe is continually tuned by the actual sensory signals, which works (in normal functioning brains) to anchor predictions to reality.

This approach openly opposes classical feedforward-dominated perceptual models. In the not too distant past the brain was commonly characterized as a passive organ. Dormant neurons were thought to patiently await incoming signals to jolt them into action. When signals did arrive, they were thought to roll in from the sensorium and flow upward through the neural hierarchy increasing in complexity along the way (Marr, 1982). In direct contrast, PP describes the brain as fundamentally proactive - the brain actively generates perceptions by continually attempting to recreate from the top-down the world of sensory signals.

A novel feature of the PP model then is a tight link between knowledge, perception, and action. The brain here uses prior knowledge to make probabilistic predictions about the structure of the environment, which cashes out as both perception and world-altering action. In the process of continually attempting to predict the next sensory input, and using any error to update those predictions, the brain learns about the actionable world. This partnership between knowledge, perception, and action allows the human brain to leverage a lifetime's worth of learning in order to respond to the newsworthy aspects of an otherwise chaotic stream of data.

Conservative, Radical and Enactive-Ecological PP

While PP theories are in many ways revolutionary, there is something about the paradigm that encourages antiquated assumptions to quietly slip in the back door. It is easy to see why. The brain here is depicted as a multi-level probabilistic prediction machine. Mentality is defined as the generation of predictive models by way of error minimization taking place entirely in hierarchically structured neurons. This has led some researchers to the conclusion that the mind should be seen as skull-bound and disembodied (Hohwy 2010, 2016).

However, while we are getting a clearer picture of the structural and functional nature of these informational flows within the predictive brain, the story about these flows can be pitched in two quite different ways. I will adopt Clark's distinctions of these views as Conservative Predictive Processing (CPP) and Radical Predictive Processing (RPP)[4].

The more conservative approach of CPP takes the predictive process to be primarily concerned with sensory inputs being accommodated by the correct selection of a hypothesis that best explains the sensory information. These hypotheses are constructed with reference to past learning and resulting estimations about the reliability of the hypotheses and signals. In this reading, prediction error is conceived of as data yet to be accounted for by the evolving hypothesis. Sensory information from the world and body helps the brain generate an increasingly more accurate model, one that mirrors the causal structure and richness of those external realities. From this perspective, our cognitive contact with such realities is found exclusively in this rich cortically-encoded inner-model, which is in turn used to select and guide behaviours. As such, CPP offers something of a predictive version of what Anderson (2014) has referred to as a "re-constructivist' approach to perception: the predictive brain thus works to rely on its own recapitulation of the world for navigation - and progressively throws out the 'real' world as a result.

Jakob Hohwy's research remains the standard when it comes to characterizing this more conservative view of PP. According to Hohwy the predictive mind is inherently neurocentric, and should be thought of as strongly secluded from the world. The predictive brain, according to Hohwy, works by continually generating best guesses (from the top-down) about what is in the world. Those guesses are deemed correct insofar as they are able to 'explain away' the incoming signals. The fact that the incoming signals are as they are becomes evidence that the

[4] To be clear, my aim in contrasting these approaches is not to explicitly argue against the CPP vision here. Rather, my goal in this project is to build on recent embodied PP models by highlighting the intimate relationship between emotion and cognition in the predictive brain. In this way, my research will help to develop RPP into a better version of itself, and thus prepare it for future debate with those less-embodied theories.

brain's generative model is accurate. The brain's use of these circular patterns of evidence forms what Hohwy describes as an 'evidentiary boundary', which he uses as a point of separation between the hypothesis generating mechanisms and the evidence that is being explained. The boundary proposed here is the edge of the sensorium: on the inside the skull-bound brain, and on the outside the body and world. It is this move that forms the basis of Hohwy's neurocentric vision of the predictive brain. As he summarizes, "this tells us how neurocentric we should be: the mind begins where sensory input is delivered through exteroceptive, proprioceptive, and interoceptive receptors and it ends where proprioceptive predictions are delivered, mainly in the spinal cord" (ibid.: 18).

This approach highlights the internalist, neurocentric and representationally rich qualities of the framework.[5] While this rendition may not be technically 'wrong' (indeed, regardless of how we take the framework to be, the math underlying these processes remains the same) it can be importantly misleading. It can give the impression that the brain is simply in the business of 'searching' for the best multi-level hypothesis for explaining the incoming sensory data. The trouble with this interpretation is evident - especially for advocates of ECS. This neurocentric interpretation makes adaptive behaviours secondary to representational fidelity. A moment's reflection, however, ought to convince us that it is action – not perception – that real-world systems really need to get right. Perceiving a structured scene is adaptively useless unless it enables you to behave in ways that lead to continued thriving (or avoid life threatening events; see Clark 2015 for further discussion of this point).

RPP opposes this neurocentric slant, highlighting instead that knowing what's out there is deeply dependent on doing something about it. Uncovering the various entanglements that exists between perception and action is a potent way of, as Clark has put it, busting the predictive mind out of the confines of the skull (Clark 2017).

Until recently researchers interested in predictive dynamics have overwhelmingly favoured discussions about perception (a notable exception is Wolpert & Kawato 1998). If action was mentioned it was commonly characterized as important only insofar as it helped reduce sensory prediction errors. Recent research reveals a different picture of the predictive brain - one that characterizes the whole predictive process as aiming to translate sensory information into successful, adaptive behaviours (for a good review see Friston et al. 2017). This shouldn't come as a surprise. The brain evolved first and foremost to facilitate life-preserving interactions with

[5] There is an enormous literature from ECS that challenges these sorts of views. A few notable sources are Churchland et al. (1994), Clark (1997), Pfeifer & Bongard (2006).

the environment. Seeing a predator would be of little value if doing so didn't also lead to behaviours that improve one's chances of safety.

According to Clark, the action-oriented RPP approach aligns well with a number of the core ideas in ECS. To establish this relationship RPP highlights two aspects of the framework commonly underappreciated by CPP: the central role that action plays in the predictive process, and a novel capacity for precision weighting to weave together faster heuristic-style actions with slower deliberative strategies in the pursuit of cost effective behavioural solutions. For Clark this is the key to seeing the predictive brain as an essentially embodied brain: an organism that has an imperative to reduce error, and is able to do so through altering both their perception *and* action, will consistently make use of whatever cheap opportunities it has available – including offloading processing on the physical body and cognitive scaffolds that the environment offers. It is this dedication to leveraging frugal error revolving strategies that helps emphasize how the PP framework and ECS might come together.

The result is a view of the predictive brain as an action-control system always on the lookout for ways to get 'more for less'. The brain is not characterized as some stronghold of accurate knowledge acquisition and rational action, but rather as an organ dedicated to producing just good enough behaviours (which commonly includes economical solutions that rely on the body and world). As Clark concludes, "Embodied, situated agents, all this suggests, are masters of 'soft assembly', building, dissolving, and rebuilding temporary ensembles that exploit whatever is available, creating shifting problem-solving wholes that effortlessly span brain, body, and world" (2015: 250). We can think of this "non-reconstructivist" approach to PP as offering what amounts to a mechanistic rendition of earlier embodied sensorimotor theories of perception (O'Regan & Noe 2001).

Still, there is something unsatisfying about this picture from an embodied mind perspective. It remains a picture in which the brain is the controller of the body. Decision-making can sometimes be offloaded onto the body in the form of fast and cheap heuristics when bodily habits are given high precision-weighting. Furthermore, bodily processes of homeostasis generate highly precise prediction errors that organize lower levels of processing in ways just described. Otherwise, however, the body of the organism appears almost as an afterthought. Such a view of PP is consistent with a view of cognition as only *modestly* embodied (Clark 2008).

In the thesis to follow I will explore a way of understanding the predictive brain that allows for the body and the surrounding environment to make a robust constitutive contribution to PP. I

do this, in part, by looking more closely at the nature of emotion (and its relationship to cognition) within the predictive organism as a whole. The resulting picture of the predictive mind, so I will argue, is not only embodied, but also perhaps enactive and ecological.

Within the recent literature on predictive processing, we can distinguish another (even more radically embodied) strand that interprets the brain in ecological and enactive terms (Allen & Friston 2018; Gallagher & Allen 2018; Kirchhoff 2018). Enactivists subscribe to a mind-life continuity thesis that takes cognitive processes and living processes to work according to the same fundamental organizing principles. Enactive theorists take adaptive regulation of coupling with the environment to be the defining characteristic of cognition, which they characterize as sense-making (Varela et al 1991; Thompson 2007; Di Paolo 2005). The organism gives meaning to its encounters with the environment in actively realizing and sustaining a self-produced identity. It produces, sustains and conserves its identity over time in its interactions with the environment, and in doing so enacts or "brings forth" a meaningful world. Cognitive and affective processes work together in this framework to steer the organism through the world in pursuit of what is significant. Colombetti writes that "Cognition from an enactive perspective is, rather, the capacity to enact or bring forth a world of sense, namely, an Umwelt that has a special significance for the organism enacting it…cognition as sense making entails that cognition is simultaneously also affective" (Colombetti 2014:18). Enactive theories of cognition take meaning to be bestowed on the environment from the lived perspective of individual agents in their interaction with an environment that includes other agents. This giving of meaning to the environment by the agent is reflected in the central enactive notion of sense-making.

In the ecological-enactive rendition of PP (EEPP) that I will support (Chapters 4, 5 and 6), affect and cognition are both understood as aspects of the whole animal-environment system.[6] The living body is depicted as providing a point of view on the environment relative to which

[6] One of the key issues separating EEPP from RPP and CPP is the role of representations in PP. The generative model has sometimes been taken to call for explanation in representational terms (see. e.g. Clark 2015; Gladziejewski 2015; Williams 2017; Kiefer & Hohwy 2018, and for non-representational accounts of PP see, Orlandi 2016; Bruineberg, Kiverstein and Rietveld 2016; Hutto 2016; Kirchhoff & Robertson 2018; Gallagher 2017). EEPP starts from the idea of the organism as a whole as embodying a model of its environment (Friston 2013). Instead of characterizing the generative model as encoding knowledge, EEPP proposes to characterize the relation of the organism to its environment in terms of attunement developed over a past history of interactions. RPP argues by contrast that the hierarchical and temporally deep structure of the generative model has to be understood in representational terms. The role of the generative model in solving representation-hungry problems that call for reasoning about the absent, distal and counterfactual means that the generative model has to be understood in representational terms. I have where possible in my thesis tried to remain neutral on this vexed issue. Chapter's 4, 5 and 6 (co-authored with proponents of EEPP) tend however to favour formulations in EEPP terms. I don't however engage with the arguments from RPP for understanding the generative model in representational terms. My thanks to Julian Kiverstein for discussion of this issue.

the environment is perceived as having affective significance. The whole embodied organism in its environment is taken to be the fundamental unit of explanation - not only the processes within the organism's brain. Drawing on ideas from ecological psychology the environmental surroundings of the organism are conceived of as offering affordances or action possibilities to the organism. As I have noted, cognition is often distinguished from affect in the philosophical and psychological literature on emotion. Appraisal and evaluation are associated with cognition, while affect is identified with changes in the autonomic system in the body. Appraisal according to this ecological-enactive approach is instead understood as involving the whole living body of the organism (as it prepares to act on relevant affordances). The organism has a basic concern to improve its grip on its environment. The EEPP approach characterizes this concern as the 'tendency towards an optimal grip'. 'Grip' here refers to the organism's bodily stance in relation to its current situation. I (and my co-authors) use this term because grip is something the organism must actively maintain in relationship with the changing environment. I talk of grip as tending towards optimality (following Bruineberg & Rietveld 2014) because the organism is aiming at restoring relative equilibrium with its environment. In EEPP, prediction errors signal an increase in disequilibrium in the organism-environment system as a whole. Such disequilibria reflect a divergence in the sensory states (exteroceptive and interoceptive) the organism expects to occupy given the kind of organism it, its current state, and the niche it inhabits. The organism then acts to reduce this disequilibrium or to make it the case that it comes to occupy the sensory states that it expects to be in (given the life it leads). In Chapters 4, 5 and 6 I explore how precision and uncertainty should be understood based on this interpretation of the relationship between affect and cognition. In particular, I make the argument that precision is mediated by affective states (states of action readiness) taking place within the organism's body as a whole as it relates to the world (more on this below).

Interoception and the living body

Researchers have already begun thinking about how the PP framework may be extended to discussions about feeling, emotions and moods (Ainley et al. 2016; Allen et al. 2016; Apps & Tsakiris 2014; Barrett & Simmons 2015; Bruineberg & Rietveld 2014; Clark 2015; Gu et al. 2013; Seth 2013, 2014; Seth et al. 2012).[7]

[7] These approaches very much mirror PP models more generally, insofar as they range from staunchly disembodied and cognitivist (Seth et al. 2012) to deeply dynamical and embodied (Bruineberg and Rietveld 2014; Kirchhoff 2016).

Seth, for example, proposes that PP applies just as neatly to interoception as it does exteroception (2013). Interoception refers to our internal sense of physiological changes in the body - somatic, visceral, vascular, and motor (Craig 2002, 2003). According to Seth's Interoceptive Predictive Processing (IP) model, top-down predictions about the source of interoceptive signals counter-flow with bottom-up interoceptive prediction errors. Feelings arise from the ongoing integration of these various predictive representations. Error signals are hypothesized to be minimized in an analogous way to sensorimotor predictions: either the error modifies the model to fit the inner world, or autonomic reflexes are initiated which influence the body to fit the prediction. In this model autonomic reflexes are called on to fulfill interoceptive predictions.

While this work clearly includes the body into the discussion, Seth's model amounts to a relatively straightforward comparator-style theory of emotion. This means that while IP extends the PP framework to include signals form the body, it remains conceptually cognitivist - the body is only a source of information, as opposed to being constitutive of the process itself (see Allen & Friston 2018 for a similar conclusion). I will argue that taking this cognitive-centric perspective is optional, and give reasons to abandon this account in favour of a more dynamical and embodied view (Chapter 2).

Another popular predictive framework focusing on interoception has been proposed by Barrett and Simmons (e.g. the Embodied Predictive Interoceptive Coding model (EPIC); 2015). While Barrett and Simmons focus on the same brain areas as Seth (e.g. the insular cortex), by taking a network perspective of the brain they conclude that this process shouldn't be thought of as modular, but instead as dynamic. As they succinctly state,

> "It may be tempting to view the interoceptive system, as outlined in the EPIC model, as a modular system. However, the brain has a small-world architecture... augmented by 'rich-club' hubs (that is, highly connected nodes), which ... serve as the brain's 'backbone' for neural communication and synchrony. Several agranular visceromotor regions—including the anterior insula and cingulate cortices—are rich-club hubs, prompting the hypothesis that agranular visceromotor cortices send predictions to and receive prediction-error signals from cortices with greater laminar differentiation in an effort to create the kind of synchronized brain activity that is necessary for consciousness" (Barrett & Simmons 2015: 425)

Predictions about the internal body are described in this model as a sort of "pacemaker" signal, one that helps synchronize the various streams of information in order to bring forth a unified, embodied experience (Allen et al. 2016; Dehaene et al. 2014).

However, both the IP and EPIC models are only embodied in so far as bodily information is given a key role in influencing and contextualizing overall predictive dynamics. While cognition, perception and action are all continually adjusted by interoceptive information, that information plays its part entirely within the brain. In other words, the body itself is only important in so far as it adjusts the internal mental representations of the body. In both models, the body that matters is located squarely inside the head. Allen and Friston make the same point about these models, writing, "Whether such representations are connectionist [EPIC] or modular [IP] in nature is of little concern; both views paint a homunculus into the picture" (2018: 2472).

I will argue that a closer investigation of the relationship between circuits typically referred to as emotional or cognitive (as outlined in Chapter 1) allows for a much more embodied view of emotion (Chapter 2). Interoception, I will argue, should be seen as more than just another stream of information that the predictive brain makes use of. Rather, affective changes in the organism's body play a constitutive role in the predictive process itself. Central to my claim is recognizing an intimate, although commonly overlooked, relationship between precision weighing and affectivity (Chapters 3, 4, 5 & 6).

Re-positioning precision

Radical Predictive Processing (RPP) makes an argument for embodiment that is reliant on precision weighting (as noted above). Specifically, it relies on precision weighting's capacity to delicately adjust the various balances within the predictive systems to provide an economical combination of simple reflexes and complex goal-driven action sets. This fluidity may make precision weighting look a little like a 'magic modulator'. Without clearly understanding precisely how precision performs this task, some might worry that RPP may have inadvertently imported an un-explanatory free variable into our explanatory schema. I argue that the threat of magic modulation can be averted by appreciating the role of subcortical processing in the estimation, orchestration and distribution of precision weighting (Chapter 3). Network approaches to understanding the entangled nature of emotion and cognition (within the brain, and between the brain and body) offer a fuller description of how precision weighting performs this important function. However, including these elements into the story requires that we understand precision a little differently; crucially, it implies that the predictive mind is much more embodied than previously assumed.

Kanai and colleagues (2015) have proposed that precision weighting is orchestrated, in part, by subcortical regions such as the pulvinar complex (the largest nucleus of the thalamus). However, Kanai and colleagues limit their investigation to the ventral pulvinar (and its relationship with the visual cortex) which results in their account overlooking many of the important embodied elements that would otherwise be expected given the important role the pulvinar plays in emotion and motivation. I will expand their work by providing a view of the pulvinar that is consistent with dynamical network characterizations of the brain (specifically, Pessoa 2014; see Chapter 3). This model characterizes the pulvinar as a central hub where various streams of information (including massive inputs from within the body) are integrated. Through this integration, perceptual processing and decision-making are directed towards signals that have affective significance for the organism. My contribution thus opens up new ways for thinking about how interoception helps to tune the predictive system towards what matters to the organism.

I go on to develop this line of thinking about the relationship between precision weighting and affective significance by proposing an account of precision weighting set by *error dynamics* (Chapters 4, 5 & 6). 'Error dynamics' refers to the temporal comparison of error reduction rates (see Joffily & Coricelli 2013; Van de Cruys 2017). For a predictive organism to thrive it needs to be sensitive to error and error reduction, as well as the *rate* at which errors are being managed. If we think about error minimization happening at a certain speed, then the predictive organism must also be sensitive to changes in velocity - accelerations and decelerations in error minimization. This sensitivity indicates how well or poorly the organism is doing at reducing uncertainty. Error dynamics, as such, are second-order processes closely related to precision weighting. Where first-order prediction errors can be thought of as acting as information *in* the system (and are used to guide behaviour and update predictions), second-order prediction errors act as information *for* the system (representing how well or poorly the system is proceeding over all).

It has been proposed that error dynamics are made available for the system as emotional valence – feelings of pleasure and displeasure (Joffily & Coricelli 2013; Van de Cruys 2017)[8]. When things are going well for the predictive agent (and its behaviours are resulting in a more certain future) it feels good. When it's struggling to get a handle on the scene, or is unable to manage the complexity of some task, it feels bad. To be clear, this feeling should not be seen as

[8] This notion of valence as emerging as a form of 'prediction error dynamics' has already found a home in both artificial intelligence and robotics circles (Schmidhuber 2010; Kaplan & Oudeyer 2007), and research on intrinsic rewards and adaptive behaviours in humans and non-humans (Kaplan & Oudeyer 2007).

something over and above the tracking of error dynamics, but rather the feelings are a reflection of the quality of the organism's engagement with the environment (see also Polani 2009). As Van de Cruys writes, "Emotions [given this addition] appear as the continuous non-conceptual feedback on evolving —increasing or decreasing— uncertainties relative to our predictions. The upshot of this view is that the various emotions, from "basic" ones to the non-typical ones such as humour, curiosity and aesthetic affects, can be shown to follow a single underlying logic" (2017).

I will argue that error dynamics play a central role in setting precision weighting (Chapters 4-6). I use this discussion about error dynamics and affect to show why we shouldn't think of precision as merely a brain event, but rather as intimately related to affective changes in the body that track and leverage opportunities to optimize the organism's relationship with the environment. I draw a conceptual connection between error dynamics and a 'tendency towards an optimal grip'. I argue that organisms make use of these feeling (arising as part of their embodied tracking of their overall condition in the world) to fine-tune precision estimations. This points to a crucial role for bodily feelings in ensuring the organism keeps expected uncertainty to a minimum in its interactions with the environment. Equivalently, this means the organism succeeds in improving its overall condition in relation to its environment. This is to say it exhibits in its behaviour a tendency towards optimal grip on the things that matter in its environment (in Chapter 6 I present an account of addiction as a case in which the tendency towards an optimal grip breaks down). By presenting affective changes as part of the precision machinery, I offer an elegant way to fold (embodied) value back into the predictive system. This addition opens the way for a much more fully embodied vision of the predictive mind to emerge.

The central aim of this thesis is to offer a theoretical bridge between PP, affective neuroscience, and a rich 'living-body' style ECS. Brought closer together, these frameworks form a synergy: dynamic network views of emotion can help fill out some of the neurobiological underpinnings still needed by the PP framework, and in so doing open the way for a more deeply embodied vision of the predictive mind to emerge. In return, PP can offer a rich suite of analytical tools, and a sophisticated theoretical landscape, for making sense of why thinking, perceiving, feeling and acting collaborate in the brain as they do. The end result is an understanding of the predictive brain as one part of a much wider entangled system that includes the visceral, active body and the social and physical environment.

Summary of chapters

The chapters of this dissertation are written as independently readable papers and so can be read in any order. That said, they do form a coherent argument, which can be best understood through a brief synopsis of each chapter in order.

- **Chapter 1** I develop a framework for thinking about a radical embodied cognitive neuroscience. The first part of the chapter argues that emotion and cognition are inseparable processes at the level of the brain. The second part of the chapter argues that emotion should be understood as a dynamic process that depends upon the whole living body of the organism. In the third part of the chapter, I propose a tight fit between environmental affordances and the patterns of action-readiness that manifest in the organism's body in the form of states of arousal and valence. Given the inseparability of emotion and cognition in the brain, and the deep dependence of emotional processes on the whole body of the living organism in its engagement with the environment, I conclude that cognitive processes are likewise deeply dependent upon these wider organism-environmental dynamics. I take the description offered above of emotional and mental experience as the emergent result of dynamic parallel loops stretching between cortical and subcortical regions, and between brain and body, to provide a perfect example of an embodied cognitive system.

- **Chapter 2** applies the central ideas from **Chapter 1** to recent PP accounts of interoception. I explore how PP can inform, and be informed by, recent debates in the philosophy of emotion. In particular, whether we ought to think of emotions as taking place entirely within the head, or instead as embodied phenomena. The first part of the chapter outlines Anil Seth's recent PP account of interoception, highlighting in particular its proposed association with traditional cognitivist theories of emotion. In the second part of the chapter I offer a critique of these cognitivist theories based on more recent dynamic systems approaches to emotion – approaches which aim to dissolve the ontological boundary between emotion and cognition at both the psychological and neurobiological levels. In the third part of the chapter I make the case that a PP account of interoception would align much better with these embodied and dynamic systems models of emotion.

- **Chapter 3** I explore what role emotions and subcortical contributions play in the PP framework. In the first part of the chapter I continue to develop the picture of online cognitive function emerging from the tight coordination of cortical and subcortical systems (**Chapter 1**). The second part of the chapter explores a role for subcortical circuits in setting "precision weighting". I build on a proposal from Kanai and Friston (2015) suggesting that the pulvinar (a portion of the subcortical thalamus) might play a role in setting precision weighting by placing their discussion within the context of dynamic network models of the brain (as seen in **Chapter 1**). The third part of the chapter argues that the close association between subcortical processing and precision engineering should lead us to think of precision engineering as more closely related to affective significance. Thus the contextualization that precision is proposed to offer turns out to be an affective one. By emphasizing subcortical contributions to setting precision weighting, and recognizing the rich reciprocal interaction between subcortical circuits and the internal feeling body, I open the door for thinking about a more fully embodied predictive brain: one that includes not just *gross morphology* and *sensorimotor integration,* but also the rich contribution that the inner dimension of affect (including changes in the endocrine system, immune system, vascular and viscera) makes to cognition. I continue to develop this line of thinking about the rich relationship between precision and affect throughout **Chapters 4, 5 and 6,** in which I present a series of test-cases for my framework**.**

- **Chapter 4** develops a novel account of why an agent that aims to minimize uncertainty would be motivated to seek out novelty. One might think that such an agent would inherently aim to avoid novel and surprising interactions with the world, perhaps even seeking out a dark (predictable) room and staying there. And yet various animals clearly find playfulness and curiosity to be highly rewarding. In this chapter I argue that the answer to this puzzle lies in understanding how bodily feelings help attune predictive organisms towards opportunities for *continually improving* their grip on what they care about. The first part of the chapter outlines the so called "dark room" challenge levelled against PP. I describe how action-oriented PP has so far attempted to overcome this challenge, and highlight a few reasons for being dissatisfied with the answers given so far. The most important criticism involves the widespread reliance on precision weighting while not yet clearly specifying the nature of how precision-weighting is set (as seen in **Chapter 3**). In the second part of the chapter I attempt to fill in this gap by suggesting that precision weighting is set in part based on *embodied error dynamics*. Error dynamics refers to the tracking of the *rate* at which error is reduced over time. Sensitivity to error

dynamics ensures that the predictive agent is continually steered towards opportunities for reducing uncertainty *at a particular rate*. Furthermore, I argue that this information about rate of change is given corporeally as states of affordance-related action readiness patterns that are simultaneously affective and behavioural. In the third part of the chapter I show how an agent that is sensitive to error dynamics will naturally be a curious agent, motivated to explore its environment. In **Chapter 6** I will go on to describe how a breakdown in this sensitivity to error dynamics is what accounts of the increasingly restrictive behaviours seen in substance addiction.

- **Chapter 5** extends this view of embodied error dynamics (constructed in **Chapter 4**) to explore how the PP framework might contribute to live debates in the philosophy of desire. The first part of the chapter reviews recent naturalistic (specifically, neuroscientific) approaches to thinking about desire as deeply related to reward learning. The second part of the chapter explores how desire and reward learning have been reinterpreted in PP terms. In the PP framework reward learning systems are thought to be involved in attentional processing, working to set precision weighting on prediction errors at multiple levels in the brain based on what the organism finds valuable or rewarding. In the third part of the chapter I argue for thinking about precision weighting as based on embodied error dynamics. Given my earlier arguments, I take this to mean that precision is based on affective tensions which arise as part of the agent's dynamic coupling with the environment. I therefore add an important enactive-ecological twist to the PP account of desire. The role of reward and value in modulating attention should be understood as taking place in the organism's body and not only in its brain, as the organism prepares to act on what matters to it in the environment. This approach further allows me to offer an account of the phenomenology of desire in PP terms, something which is currently missing in the literature. This vision of error dynamics as being closely associated with midbrain reward processing offers an bridge between this enactive-ecological PP framework and addiction research which I go on to explore in **Chapter 6**.

- **Chapter 6** draws on PP to explore whether or not addiction should be thought of as simply a disease of the brain. It is widely acknowledged that addiction is accompanied by structural and functional changes in the brains of addicts that directly correlate with pathological symptoms such as intense craving, compulsive drug-seeking behaviour and a breakdown in cognitive control. Based on this evidence the model of addiction as a brain disease is now widely accepted. However not everyone has embraced this model. Marc Lewis (forthcoming) has developed an alternative interpretation of the neurobiological

evidence. He argues that addiction is a form of maladaptive learning. In this chapter I agree with Lewis that addiction is not a brain disease. I argue instead that the pathological nature of addiction is not to be found in the brains of addicts but in the larger organism-environment system of which the brain is (of course) an important part. I construct this argument in large part on the basis of predictive processing models of addiction, however I once again suggest an important ecological-enactive twist on the current models. The first part of the chapter outlines the effects of addictive behaviours on the midbrain's reward circuitry. In the second part of the chapter I reinterpreted these effects in PP terms. In particular, outlining the view that reward learning is used to weight the precision of the prediction errors that lead to behaviours. I show how the "hijacking" of reward learning can be redescribed in PP as (roughly speaking) the substance determining how precision is weighted. Consistent with **Chapters 4 and 5,** I argue that it is in fact a sensitivity to embodied error dynamics that is thereby being hijacked in addiction. In the third part of the chapter I use these insights to develop an ecological-enactive predictive processing account of addiction. I argue that precision weighting relative to error dynamics is best understood within the larger context of an organism aiming to improve its 'grip' on the things that matter in its environment. I go on to develop a line of thinking about the destructive behaviours of addicts as the result of *disorganization* within the wider agent-environment system. From this perspective addiction is best characterized not as a change in any particular neural circuitry, but as a more general loss of attunement between the organism and its environment. I argue that PP, with the addition of embodied error dynamics, can offer a rich explanation of how neural computational processes in addiction contribute to the breakdown in this wider organism-environment system. This suggests a broader palette of possible interventions and helps bring culture, context, and neural processing into a single explanatory framework.

Author's contributions

The chapters in this dissertation have either been published in article form or are currently in review. Below I provide, if existing, the reference to the published version, and if co-authored I note the relative author contributions.

- **Chapter 1** has previously been published as:
  Kiverstein, J., & Miller, M. (2015). The Embodied Brain: Towards a Radical Embodied Cognitive Neuroscience. *Frontiers in Human Neuroscience*, *9*, 237.
  Author contribution: MM and JK conceived of the paper together and co-wrote the paper.

- **Chapter 2** is currently in preparation and has Mark Miller and Jelena Markovic as its authors.
  Author contribution: MM wrote the paper with extensive revision from JM.

- **Chapter 3** has previously been published as:
  Miller, M., & Clark, A. (2017). Happily Entangled: Prediction, Emotion, and the Embodied Mind. *Synthese*, 1-17.
  Author contribution: MM conceived of the paper, MM and AC constructed the core arguments of the paper, and co-wrote some parts of the paper.

- **Chapter 4** has previously been published as:
  Kiverstein, J., Miller, M. & Rietveld, E. (2017). The Feeling of Grip: Error Dynamics, Novelty and the Predictive Brain. *Synthese*, 1-23.
  Author contribution: MM conceived of the paper, MM and JK constructed the core arguments of the paper, and co-wrote some parts of the paper with revisions from ER.

- **Chapter 5** is currently in preparation and has Julian Kiverstein, Mark Miller and Erik Rietveld as its authors.
  Author contribution: MM conceived of the paper, and constructed the core arguments of the paper. MM and JK co-wrote some parts of the paper with revisions from ER.

- **Chapter 6** is currently in preparation and has Mark Miller, Julian Kiverstein and Erik Rietveld as its authors.
  Author contribution: MM conceived of the paper, and constructed the core arguments of the paper. MM and JK co-wrote some parts of the paper with revisions from ER.

Chapter 1

The embodied brain: towards a radical embodied cognitive neuroscience

Abstract

In this programmatic paper we explain why a radical embodied cognitive neuroscience is needed. We argue for such a claim based on problems that have arisen in cognitive neuroscience for the project of localizing function to specific brain structures. The problems come from research concerned with functional and structural connectivity that strongly suggests that the function a brain region serves is dynamic, and changes over time. We argue that in order to determine the function of a specific brain area, neuroscientists need to zoom out and look at the larger organism-environment system. We therefore argue that instead of looking to cognitive psychology for an analysis of psychological functions, cognitive neuroscience should look to an ecological dynamical psychology. A second aim of our paper is to develop an account of embodied cognition based on the inseparability of cognitive and emotional processing in the brain. We argue that emotions are best understood in terms of action readiness (Frijda, 1986, 2007) in the context of the organism's ongoing skillful engagement with the environment (Rietveld, 2008; Bruineberg and Rietveld, 2014; Kiverstein and Rietveld, 2015, forthcoming). States of action readiness involve the whole living body of the organism, and are elicited by possibilities for action in the environment that matter to the organism. Since emotion and cognition are inseparable processes in the brain it follows that what is true of emotion is also true of cognition. Cognitive processes are likewise processes taking place in the whole living body of an organism as it engages with relevant possibilities for action.

1.1 Introduction

Radical embodied cognitive science is a relatively new branch of cognitive science that looks to ecological psychology and dynamical systems theory to understand the contribution of bodily capacities to cognitive processes (Chemero, 2009; Barrett, 2011). It is "radical" in claiming that

cognitive scientists need new conceptual tools if they are to understand the ways in which cognition depends on the body in its interaction with the environment. The classical conception of the human mind as working according to the same principles as a digital computer encourages us to think of the body and the environment as providing at best inputs to, and receiving outputs from cognitive processes (Rupert, 2009). Embodied approaches to cognitive science by contrast stress the many and varied ways in which an animal's environmental niche offers resources for the animal to act on. The individual has bodily skills and abilities that are refined and perfected through practice for dealing adequately with the possibilities for action the environment offers. The explanatory tools cognitive science deploys must do justice to the essential contributions of bodily skills and environmental affordances to cognitive behavior. They must account for the ways in which the individual is able to expertly coordinate their behavior with a dynamically changing environment. Ecological psychology and dynamical systems theory provide the tools to meet this challenge. Moreover the conceptual tools these sciences offer are arguably better suited to explaining the complex, dynamical interactions between an animal and its environment than the standard computational tools of cognitive science. This gives us a pragmatic reason to add these explanatory frameworks to the explanatory toolkit of cognitive science.

Our aim in this paper will be to argue that cognitive neuroscience should look to ecological dynamical psychology for an understanding of the psychological functions the brain performs. The classical approach to cognitive science encourages the following view of the division of labor between psychology and neuroscience. Cognitive psychology provides analyses of the cognitive operations an individual must perform in order to carry out a cognitive task. Cognitive neuroscience then seeks to determine how these cognitive operations are carried out by brain regions and networks of brain regions. Consider vision as an example. The tasks the visual system performs are commonly decomposed and broken down into early visual processing in which an image is formed, and features of the surfaces in the surrounding environment are represented. At an intermediate stage in visual processing features of surfaces are used to construct object representations. Late processes then use object-based information to put objects into categories (Palmer and Kimchi, 1986). Cognitive neuroscientists design their experiments based on this type of functional decomposition of visual processing in the brain. The aim of their experiments is to use the knowledge we have accumulated from previous experiments to determine how these different stages in visual processing are carried out by the brain. There is no room in this view of the division of labor between psychology and neuroscience for the body and the environment to play anything other than a peripheral role in cognitive processes.

A radical embodied cognitive neuroscience will take its analyses of psychological function not from cognitive psychology, but instead from ecological psychology and dynamical systems theory. The advantage of doing so will be an analysis of psychological function that does justice to the contribution of the bodily capacities and the ecological setting to a given cognitive behavior.

The first part of our paper will point to findings that strongly suggest it is time for cognitive neuroscientists to look elsewhere for their analyses of psychological function than to classical cognitive psychology. We begin by reviewing recent research concerned with large-scale patterns of connectivity in the brain. We take this research to present a major challenge to any analysis of the brain into functionally specialized regions that carry out either emotional or cognitive psychological functions. Such a view of the brain as being made up of functionally specialized regions follows naturally from the view of the division of labor between psychology and neuroscience we sketched above. Cognitive psychology has the job of providing analyses of the functions that are computed in emotional and cognitive processing. Cognitive neuroscience looks to determine how these functions are performed by different regions and networks in the brain (see e.g., Posner et al., 1988). The work we will review on functional and structural connectivity in the brains of humans and other animals supports a different perspective on brain processes in which classical emotional and cognitive brain regions are in constant dynamic interaction. We will show how this work on functional connectivity makes trouble for attempts at localizing either emotional or cognitive functions to specific brain structures.

We argue next that psychological function is better understood at the level of the whole brain-body-environment system. The argument we offer for this second thesis is somewhat circuitous, but necessarily so because it enables us to engage with, and distinguish our proposal from other related theories in the literature on emotion and cognition. We begin by comparing our perspective with that of psychological constructionist (or dimensional) theories of emotion which interpret the integration of cognitive and emotional processes in terms of interactions between domain general neural networks (see e.g., Barrett and Satpute, 2013). We suggest (following arguments developed in Pessoa, 2014) that structure-function mappings are not fixed and static properties of networks. Instead structure-function relationships are dynamic, with the functions a given network performs varying over time in a highly flexible and context-dependent manner. We then argue that the functional contribution of a network is determined by the whole organism in its interaction with an environment that is rich with possibilities for actions. To

determine the precise functional contribution of a network to an animal's behavior we must look at how this network functions in the context of a wider organism-environment system.

The radical rejection of the classical computationalist explanatory framework by advocates of an embodied approach to cognitive science has been taken to call into question the place of the concept of representation in psychological explanation. Hutto and Myin have argued for instance that the dynamic engagement of an animal with the environment doesn't require the extraction, processing and manipulation of states with semantic or representational content (Hutto and Myin, 2013). We do not engage directly with this issue, but instead focus our arguments on defending the claim that embodied cognition is best studied at the level of the whole brain-body-environment system. We leave the implications of our arguments for the role of representation in psychological explanation as a matter to be discussed on another occasion.

The positive view we develop in this paper will be built up in two stages. First, we argue that cognition and emotion are inseparable processes in the brain. Second, we argue emotion is a dynamic process involving the whole body of the organism. The first two steps in our argument establish the inseparability of emotion and cognition in the brain and the deep dependence of emotional processes on the whole body of the living organism in its practical skilled engagement with the environment. We take these two steps to imply a third step: the conclusion that cognitive processes depend on the whole living body of a person or animal in its practical and skilled engagement with an environment of affordances. Since emotion deeply depends on the living body, so also does cognition.

1.2 Localizing emotion and cognition in the brain

When we ordinarily think of emotion we often think of short-lived, transient episodes that wash over, and sometimes overwhelm us, gradually fading away after a relatively short period of time. Folk psychology makes distinctions between episodes of anger, disgust, fear, happiness, sadness and so on. When we think of our own emotions and those of others we do so using these folk psychological categories. How do these folk psychological emotion categories map onto processes in the brain?

Affective neuroscientists have posited a set of biologically basic emotions such as rage, fear, lust, care, and grief that can be localized to specific and dedicated networks in the human brain (Panksepp, 1998, 2012). Typically these basic emotions are associated with the brain stem, the

diencephalon (thalamus and hypothalamus), and limbic structures which are taken to be evolutionarily old, primitive parts of the brain, highly structured at birth and relatively isolated from learning. These parts of the brain are directly connected, and tightly coupled to autonomic, endocrine and immune systems in the body that work together to keep an organism's body in a state of homeostatic and metabolic equilibrium. They are also taken to be "automatic" in their processing (as contrasted with cognitively controlled processes), and are thought to be critically involved in impulsive behavioral responses such as fear and rage.

Cognitive processes (such as learning and memory, reasoning and planning) are often associated with the phylogenetically newer and more highly evolved cerebral cortex. The neocortex in primates has been described as the "crowning achievement of evolution and the biological substrate of the human mental prowess." (Rakic, 2009, quoted by Barton, 2012, p. 2098). The primate neocortex for instance shows a fivefold difference in volume when compared to that found in insectivores (Barton and Harvey, 2000). This growth is thought by many to be accompanied by an evolution of higher-cognitive functions. Systems in the prefrontal cortex (PFC) for instance make use of information that has been processed by other parts of the brain to ensure that we, in contrast to our animal ancestors, keep our emotional impulses in check. As the Victorian neurologist John Hughlings-Jackson put it: "the higher nervous arrangements evolved out of the lower keep down those lower, just as a government evolved out of a nation controls as well as directs that nation." (Jackson, 1884, p. 662, quoted by Parvizi, 2009, p. 354) The frontal lobes exert control over, and suppress our more animal desires, thereby ensuring that we act in ways that are contextually appropriate. These functional processes allow higher mammals to compare possible plans and strategies offline, and make a cost-benefit calculation as to which possible course of action is likely to be the most beneficial in the long run.

This understanding of cognition and emotion leads to a view of the mammalian brain as divided into cognitive "higher" regions (neocortex) and emotional "lower" subcortical regions. This division is perhaps best exemplified in Paul MacLean's discredited triune model of the mammalian brain (MacLean, 1952, 1990; for criticisms see Swanson, 1983; LeDoux, 2012). The lower, animal parts of the brain are understood (in line with the Hughlings-Jackson "Victorian" narrative) as standing in a linear, and hierarchical relationship to the higher neocortical regions. Why assume however that the only parts of the human brain to undergo change over the course of evolution were those located in the cortex? An alternative co-evolutionary hypothesis is that both cortex and sub-cortex underwent changes in a coordinated fashion. Those brain structures with major anatomical and functional links most likely evolved together (Barton and Harvey, 2000). Barton (2012) discusses for instance how the cerebellum (an area known to be involved in

the learning of motor skills) is larger in primates compared with other mammals. He adduces evidence for the threefold co- evolution of the diencephalon, cerebellum and the neocortex. There is also evidence that subregions of the amygdala are substantially more developed in monkeys as compared with rats (Chareyon et al., 2011). Pessoa has suggested, in line with the co-evolutionary hypothesis we just sketched, that this increase in size in the amygdala is likely to be linked to the degree of connectivity with other brain structures (Pessoa, 2014, pp. 413–414).

The old hierarchical model of the higher primate brain and the lower reptilian brain assumes a unidirectional flow of information from lower to higher brain systems. While the higher-brain systems depend on lower-brain systems for their functioning the reverse is not true. Parvizi (2009) offers an important critique of this "corticocentric myopia". In his words "higher functions of the brain are made possible by a reciprocal interconnection between cortical and subcortical structures rather than being localized only in the upper tip of the vertical neuroaxis" (p. 354). Parvizi describes the connectivity between the cortex and the rest of the brain in terms of "reciprocal interconnection". This means that there are complex relations of negative and positive feedback that characterize the communication between sub-cortex and cortex. Parvizi suggests that higher brain functions (such as executive control functions) happen in "the loops operating between the cortical areas and "lower" subcortical structures such as the basal ganglia; the basal forebrain; the thalamus; the cerebellum; and the brainstem dopaminergic and noradrenergic systems." (p. 358, supporting references omitted).

Consider as an example the hypothalamus, a region located just below the thalamus and above the brain stem. The hypothalamus is commonly associated with the coordination of homeostatic mechanisms such as hormonal and behavioral circadian rhythms and neuroendocrine processes. However, it is also bidirectionally connected to the cerebral cortex by at least four pathways that run via the basal forebrain and amygdala; the brainstem and the thalamus (Risold et al., 1997). Barbas and Rempel-Clower (1997) showed that in primates the hypothalamus projects to all regions (orbital, medial and lateral) of the PFC. The hypothalamus functions as a so- called "connectivity hub" that is optimally placed because of its extensive connections to have a near global effect on brain function (Pessoa, 2013, pp. 230–231). Importantly, areas of the PFC (orbital and medial PFC), insular cortex, hippocampus, and amygdala also link back to the hypothalamus. The connections between the hypothalamus and PFC are bidirectional and reciprocal, allowing for rapid coordination and synchronization of activity between "higher" and "lower" brain systems. This coordination allows for cognitive and affective processes to be mobilized together allowing the animal to behave flexibly, and in ways that are adapted to the particularities of a context of activity.

This picture of higher cognitive systems and lower emotional systems as being "vertically" integrated and tightly coordinated strongly argues against a corticocentric myopia. It suggests instead a view of cognitive and emotional processes as strongly interdependent (Lewis, 2005; Stapleton, 2013; Pessoa, 2013; Colombetti, 2014, ch.4). By "interdependent" we mean to refer to the degree to which cortical and sub-cortical systems influence each other. This degree of influence is measured by the information-processing operations the components that make up these systems each perform. The evidence (some of which we have reviewed above) points to a tight coupling or mutual influence between these systems. The operations carried out by components located in the cortex are constantly effecting, and being effected by, the operations that are taking place in components found in the sub-cortical systems. Functional connectivity isn't just about channeling information between functionally specialized brain regions. Instead it "generates complex system-wide dynamics that enable local regions to participate across a broad range of cognitive and behavioral tasks" (Byrge et al., 2014, p. 395).

Marc Lewis provides a clear example of the dynamical interaction between emotion and cognition in the brain (Lewis, 2005). He discusses the cognitive and neural processes that are engaged when a person experiences "road rage". We quote him at length because his example makes it clear why it is important to understand this interaction between emotional and cognitive processes as taking place at the level of the whole embodied person in the environment, a point we return to below.

> "Mr. Smart slams on the brakes when noticing the proximity of the car in front. Anger arises initially from frustration, as Mr. Smart wants to keep driving fast, but also from a sense of violated entitlement: he is in the left lane and should not have to slow down. Fear may also be triggered by the close call, eliciting further anger because of an intermediate evaluation of unmanly helplessness. These emotions arise rapidly, but they are paralleled by a co-emerging sense of the other driver as intentionally obstructive (and therefore blameworthy). Mr. Smart's highly focused visual attention, a derivative of anger, takes in the red color of the car ahead, as well as the expensive-looking design, and his anger is amplified by his sense of the unfairness of this show- off blocking his path (based on an implicit memory of some long- forgotten or fantasized rival). A stabilizing angry-anxious state, coupled with ruminative plans for vengeance (perhaps a blast of the horn), anchors attention to the head of the man in front. This lasts for a minute or two while Mr. Smart fashions and modifies plans to pass on the right. However, when the man peers over his shoulder, Mr. Smart evaluates this act as a taunt, generating shame and anger in an elaborated appraisal of humiliation, and calling for extreme action to save his self-image from further subjugation" (Lewis, 2005, p. 175).

For Lewis a change in emotional state is triggered whenever orderly behavior "is interrupted by a perturbation, resulting in a rapid loss of orderliness and an increase in sensitivity to the

environment" (Lewis, 2005, p. 174). In Smart's case the suddenly obstructed lane jolts his body from a state of low arousal (feeling of calmness and flow) to being highly aroused (feelings of fear and anger). Lewis writes: "living systems are like taut springs, ready to respond to small perturbations that are biologically meaningful" (Lewis, 2005, p. 176; see also Kauffman, 1993). Any cognitive or emotional event can be a trigger, so long as it sufficiently perturbs the system into a state of disorder. Once order has been disrupted, sub-cortical and cortical neural systems enter into a series of recursive feed- back loops, activity in the one system amplifying activity in the other (Lewis, 2005, p. 176). Consider for example the positive feed-back loop established between the psychological processes of bodily arousal on the one hand (realized in sub-cortical, limbic systems), and attention and recall (involving regions in the neocortex) on the other. Smart's feelings of violation at being halted in the fast lane triggers feelings of anger. The arousal generated by the anger motivates and focuses Smart's attentional resources onto the dangerous and transgressive elements in the environment: the driver of the red sports car. This increased attentional focus highlights the expensiveness of the automobile, which in turn triggers feelings of unfairness, which increase Smart's feelings of anger. Simultaneously Smart's angry arousal anchors his recall processes to similar past situations, and memories of past transgressors, which amplifies Smart's feeling of anger. Smart's bodily arousal motivates and directs his attentional and recall processes. These cognitive processes in turn intensify and prolong his state of anger arousal.

When the somatic and cognitive processes become appropriately coupled (anger in the form of bodily arousal directing attention and recall, attention and recall sustaining/amplifying bodily arousal) the brain systems supporting these psychological processes begin to settle into stable patterns of activity (Lewis, 2005, p. 177). Just as a group of birds quickly settles into an enduring flocking pattern, so also do Smart's emotional and cognitive processes temporarily stabilize and settle into a coherent and large scale anger-anxiety state. The "lower" neural processes that track bodily arousal, and the "higher" neural processes associated with attention and memory sustain each other, and generate an enduring emotional-cognitive state.

Lewis's example is framed in psychological terms, but the reciprocal and mutual influence he describes strongly speaks against any separation of emotion and cognition in the brain. Subcortical and cortical networks non-linearly interact in his example in such a way as to sustain temporary, large-scale patterns of organization over time. One might object that the activity in these networks take shape on the basis of interaction, or communication between functionally specialized brain regions. We think this is exactly the picture of the organization in the brain that is called into question once one rejects corticocentric myopia. In the Smart example, we see how

negative and positive feedback loops shift the brain from a state of relative disorder to a temporary, more or less short- lived pattern of global coherence. This shift from disorder to global coherence involves the formation of large scale networks in the brain. The relations of feedback that are critical for the formation of such networks mean that each element is directly or indirectly affecting every other element that makes up the network. This fundamentally challenges a picture in which each element performs a psychological operation (either emotional or cognitive) apart from its interactions with other elements. We suggest instead conceptualizing psychological function at the level of the processes taking place in the large-scale network as a whole. Large-scale networks implement psychological processes that are simultaneously both emotional and cognition in nature. They are in Lewis's words "amalgams" of emotion and cognition.

1.3 Are emotional and cognitive processes "psychological constructs"?

We have now given the argument for the first claim we wish to make in this paper that the brains of animals resist functional decomposition into separable emotional and cognitive components. In this section we begin making the case for the second step in our argument, which is a defense of the claim that the amalgam of emotional and cognitive processes we find in the brain deeply depend upon the whole living body of an organism. We begin by discussing psychological constructionism because research in this tradition would seem, at least at first glance, to lend further support for the view we have been developing. Psychological constructionism grew out of the dimensional theory of emotion. The dimensional theory claims that emotional episodes such as fear, anger, and sadness are combinations of more fundamental dimensions such as arousal (the strength and intensity of an emotion) and valence (the degree of pleasantness or unpleasantness) working in combination with cognitive processes of appraisal The dimensional theory can be traced back to Wundt and has received more recent defense in the work of James Russell and Lisa Feldman-Barrett (Wundt, 1897; Russell, 1980, 2003; Barrett, 2006). More recently, psychological constructionists have begun to argue that emotions do not map onto distinct regions and networks in the brain but are instead the result of dynamic interactions between large-scale networks that compute domain- general functions (Barrett and Satpute, 2013). This looks to be very much in keeping with what was argued in the previous section. The constructionist theory is often contrasted with categorical or discrete theories of emotion (briefly discussed at the beginning of section Introduction). Discrete theories of emotion posit basic emotions (e.g., anger, fear, sadness, happiness, disgust and surprise) that are species-universal, hardwired, and have unique physiological and neural signatures or profiles (see e.g.,

Panksepp, 1998; Ekman, 1999; Izard, 2007, 2011). Discrete theories of emotion have unfortunately often subscribed to the problematic division of the brain into emotional and cognitive systems criticized in the previous section (see e.g., Panksepp, 1998). Constructionists add something important to our earlier argument by calling into question the claim that emotions can be mapped onto specific functionally-specialized regions and networks in the brain. They argue instead that the brain is organized into domain-general, distributed functional networks, and emotions are the result of interactions between these networks.

A growing literature supports a view of function-structure mappings in the brain as many-to-many, thereby bolstering the case for a constructionist theory of emotion. There can be no one-to-one mapping of psychological function to anatomical regions or structures because brain regions and structures exhibit extensive pluripotency and degeneracy. Pluripotency refers to the well-established finding that one and the same region (e.g., Broca's area) can be involved in the performance of multiple functions e.g., language processing, movement preparation, imitation and imagery related tasks (see Anderson, 2010, 2014 for discussion of this and many other examples of pluripotency). Degeneracy refers to the finding that different neural structures can perform one and the same function (Edelman and Gally, 2001; Friston and Price, 2003; Figdor, 2010). Taken together these findings suggest a *many-to-many* mapping of structure to function at the level of brain regions.

This seems to present a challenge to the discrete theory of emotion. Pluripotency and degeneracy strongly suggest that each basic emotion is unlikely to have its own physiological and neural profile. Consider what we now know about the amygdala in this regard, an area that is often referred to as supporting a discrete theory of fear because it has repeatedly been shown to be involved in threat responses in rats and humans (see e.g., LeDoux, 1996; Öhman and Mineka, 2001; for critical discussion see Sander et al., 2003). The amygdala has however also been shown to be active when people are presented with novel, but emotionally neutral stimuli (Moriguchi et al., 2011; Balderston et al., 2011; Blackford et al., 2011). Herry and colleagues for example found increased activity in the amygdala when subjects were presented with unpredictable sequences of tones as compared with predictable sequences (Herry et al., 2007). The amygdala is involved in a wide variety of different functions, including "cognitive" functions such as value representation and decision-making (Sergerie et al., 2008; Pessoa, 2013, ch.2).

Recent meta-analyses have shown that other brain regions associated with emotion such as the anterior insula, pregenual and subgenual anterior cingulate and orbitofrontal cortex also show increases in activity for a variety of different emotion states. Lindquist and colleagues compared

the sets of brain regions that were consistently activated in studies of anger, disgust, fear, happiness and sadness. They found six distributed networks that consistently showed up in the studies they analyzed from 1990–2007. The networks and the regions of which they were composed were not associated with particular emotion categories, but were instead found to be active in all studies of emotion experience they analyzed. They found no brain regions that were functionally specialized—every region that was activated for one emotion category was also activated for at least one other emotion category (see Figure 5 in Lindquist et al., 2012 for a useful visual summary of the findings of their meta-analyses). The same brain regions can carry out a variety of distinct psychological operations and belong to different overlapping networks over time (also see Anderson, 2010, 2014; Colombo, 2013). What a brain region does at any given time will depend on the network with which it is affiliated.

Lindquist and colleagues argue their meta-analysis supports a constructionist or dimensional theory of emotions, and challenges discrete theories (which they label "natural kind" theories). Emotion in general has as one of its components bodily arousal which can (but need not) be combined with pleasurable or unpleasant feelings. We will henceforth refer to the latter dimension of emotion as "valence". Constructionists give this combination of arousal and valence the label "core affect". Core affect plays a role in homeostasis tracking endocrinal, visceral and muscular changes internal to the body that inform the organism that there is something in the environment of potential relevance or value. Barrett and Satpute (2013) take core affect to be neurally realized by a large-scale intrinsic network that has come to be called the "salience" or "ventral attention" network (Menon and Uddin, 2010). An intrinsic network is a network of widely distributed brain regions whose activations are tightly correlated across time when subjects are at rest, and their attention is not engaged by any external task or stimulus (Seeley et al., 2007; Bressler and Menon, 2010; Raichle, 2010; Yeo et al., 2011). Barrett and Satpute suggest that the salience network is made up of dorsal and ventral subnetworks. The dorsal subnetwork uses homeostatic and metabolic information from the body to guide attention and motor behavior. The ventral subnetwork realizes affective feelings that are experienced by a subject as pleasurable or unpleasant with different degrees of arousal. The salience network carries out "domain-general" functions, which is to say that this network of brain regions is active in a wide range of tasks (i.e., it is not domain or function-specific). What these tasks all share in common is they all require the orienting of attention to homeostatic or metabolically relevant information.

Core affect doesn't provide the basis for making folk psychological distinctions between emotions. It is a feature common to all of the emotions we ordinarily distinguish between in folk

psychology. Core affect takes on the character of different emotions only through the interaction of the salience network with other domain general networks in which such functions as categorization, language processing and executive control takes place. We focus on the role of categorization here since it will prove important for the argument we make for an embodied view of psychological function in the next section of the paper. It also provides an opportunity for us to briefly compare our theory with related work on the embodiment of emotion.

Constructionists argue that it is on the basis of the categorization of core affect that states of bodily arousal are made meaningful and related to a determinate object. Using the meta- analytic studies mentioned above as evidence they have recently argued that categorization takes place in the episodic memory, or default mode network (DMN) of brain regions that reconstructs past experiences for use in current processing (Bar, 2007; Wilson- Mendenhall et al., 2011; Lindquist and Barrett, 2012; Lindquist et al., 2012). Areas of the default-mode network (medial PFC, posterior cingulate cortex/precuneus, medial temporal lobe) were found to be consistently active during a range of emotional states (Kober et al., 2008; Lindquist et al., 2012). Constructionists have proposed the hypothesis that the DMN may function to model probabilistically the causes of current core affective changes (Lindquist et al., 2012, p. 125). The result of such models of the causes of core affect is the categorization of core affect as an instance of anger, fear, sadness or whatever (see also Barrett and Bar, 2009).

As already noted, this hypothesis certainly makes a good fit with the findings we reviewed in the previous section of extensive connectivity between emotional and cognitive brain regions. Emotion is the outcome of interaction between multiple psychological components, each associated with assemblies of neurons within distributed networks. Lindquist and colleagues write that "these networks combine and constrain one another like ingredients in a recipe, influencing and shaping one another in real time according to the principles of constraint satisfaction" (p.126). However, we shall argue that even this picture of function-to-structure mapping may need to be revised in the light of the arguments we made above in section one. Constructionists subscribe to a mechanistic view of psychological functions whereby emotion is decomposable into basic psychological operations which can then be mapped onto distinguishable networks or "flexible assemblies of neurons" that "fire together in a probabilistic way" (Lindquist and Barrett, 2012). We suggest however that functional connectivity may make trouble for such a functional decomposition, just as it did for attempts at localizing emotional or cognitive functions to specific components. We saw in the Smart example how large-scale networks form in the brain through positive and negative feedback. This means that either directly through local anatomical connections, or indirectly through long- range connections,

every brain region has the potential to influence every other brain region within a given network. The function and operations a particular region carries out will be determined by (but also determine) its interactions with the other elements to which it is connected in a network.

In this section, we've seen how the same region can play a role in carrying out very different functions over time. In order to determine what function a given region is performing we need to look at the network with which it is affiliated, but this is something that varies over time. Pessoa has suggested that "In the extreme, two networks may involve the exact same regions interacting with each other in distinct ways across time" (Pessoa, 2014, p. 408). The function of a region is thus not a fixed and static property, but is dynamic and context- dependent, varying with the network in which the region is functioning. A version of the problems we have raised with localizing function to structure may therefore also arise at the organizational level of networks. The finding that there is no one-to-one mapping of structure to function might also generalize to the domain-general networks appealed to by psychological constructionists to explain emotion in the brain. The function of a given intrinsic network may also be dynamic, with very different networks making the same functional contribution to behavior at different times (Pessoa, 2014; Fazelpour and Thompson, 2015). The same network of brain regions may contribute differently to behavior because of the way in which the elements of which it is composed are interacting.

The structure-function relation for networks, and not only for brain regions is thus also in dynamic flux. Pessoa has suggested that in order to determine the precise functional contribution of a given network we may need to look at "global variables" such as neurotransmitters, bodily arousal, and slow-wave potentials (2014, p. 408). The function a given network performs is dependent upon the wider context in which the network is active. How should we think about this context-dependence? We shall argue the context we need to take into account may include the rest of the body in its interactions with the environment, as is argued in radical embodied cognitive science. The mechanistic style of explanation that constructionists employ assumes however that networks have fixed domain-general functions. We saw this for instance in the constructionist proposal to divide the salience network into ventral and dorsal subnetworks with ventral regions directing the selection of visceromotor responses, and the dorsal parts being assigned the function of spatial orientation and motor selection. The research we have reviewed on functional connectivity challenges any such view of structure-functions mappings as fixed and permanent.

Is this simply a complication in the mechanistic theory of emotion that constructionists propose, or does it constitute a more serious challenge? In the next section we suggest it may be interpreted as supporting an embodied interpretation of the results reviewed above. Once we start to admit the role of global variables such as activity in neurotransmitter systems and valenced states of bodily arousal in influencing the functioning of a given network, why think of the boundary between the brain and the rest of the body as a sort of "magical membrane" (Hurley, 2010)? Why think that the factors that influence the function of a network reside only inside of the brain? In the next section we argue that bodily states of arousal and valence (which manifest as changes in the body's vascular, visceral and motor systems) shift the brain from state of relative disorder to temporary patterns of large scale coherence. The environment elicits patterns of action readiness that manifest in the body in the form of states of arousal and valence. We take these two points to establish the main claim of our paper that to understand the psychological function of a large scale network requires neuroscientists to pay attention to the whole organism-environment system.

1.4 The deep dependence of emotion and cognition on the living body

The constructionist, dimensional theory of emotion we discussed in the previous section conceives of emotional experience as the product of the interaction between different components. We've discussed the core affect and situated conceptualization networks above. Bodily arousal and valence (core affect) both occur as part of the life-regulation, homeostatic and metabolic processes of an organism, a being that strives to resist disorder and disintegration in its interactions with the environment (Spinoza, 1677/1894, *Ethics* III, 6 and 7; Thompson, 2007, part 2; Colombetti, 2014). We follow Giovanna Colombetti in arguing that these states of the living body provide an organism with a means of evaluating and appraising aspects of its surrounding environment in terms of their relevance or significance for the organism (Colombetti, 2014). "Relevance" is determined by the organism in relation to what the phenomenological philosopher Merleau-Ponty described as the "organism's proper manner of realizing equilibrium" with the environment (Merleau-Ponty, 1942/1963, p. 154, also see Bruineberg and Rietveld, 2014). When Merleau-Ponty writes of the organism "realizing" equilibrium with the environment, we take him to be referring to situations in which the organism is coping "smoothly" with the environment (to borrow a phrase from Dreyfus, 1991). Crucially, the organism never fully accomplishes equilibrium with the environment so long as it is alive. There is always room for further improvement. As long as the organism has needs and desires it will always be in a state of a state of relative disequilibrium with the environment, a

metastable state (Kelso, 2012; Bruineberg and Rietveld, 2014). Living systems act so as to reduce this disequilibrium thereby improving their situation, and taking them closer towards a state of equilibrium with the environment.

One of the core components of bodily affect is what the psychologist Nico Frijda called "action readiness" (Frijda, 1986, 2007). Affect makes the organism ready to act in ways that improve its grip on the situation in which it finds itself (Rietveld, 2008; Bruineberg and Rietveld, 2014; Kiverstein and Rietveld, 2015, forthcoming). The states of bodily arousal that are either negatively or positively valenced we take to be patterns of bodily action readiness. Constructionists also take core affect to be the "body's way of representing whether objects in the environment are valuable or not in a given context" (Lindquist et al., 2012, p. 124). The core affect network is described as orienting the "brain's processing capacity towards the most homeostatically- relevant and metabolically-relevant information—it performs a body-based source of attention within the human brain" (Barrett and Satpute, 2013, p. 366). This is all very much in keeping with the view we have just outlined, but with a number of crucial differences. We argue that this orientation to what is relevant in the environment should be conceived of as action readiness, where the latter needs to be understood in the context of a whole animal-environment system.

The living organism always finds itself in an environment offering many possibilities for actions. From these possibilities some are singled out as important to the organism because they are possibilities that elicit an action-readiness in the organism. The organism is drawn to act on those affordances that bring the organism closer to equilibrium with the environment, and move the organism further away from a state of disequilibrium. We are suggesting that interoceptive areas of the brain track changes in patterns of action readiness in the body of the organism as a whole. These bodily changes reflect the organism's state of relative equilibrium with the environment. When the body of the organism is aroused by some opportunity or challenge in the environment, the effect on the large-scale patterns of activity in the brain is that of destabilizing and disrupting the self-sustaining pattern of organization that has temporarily taken form. In the terminology of dynamical systems theory, the brain is caused to shift out of one attractor state. The brain then settles into new large scale patterns of activity (a new attractor state) that makes the organism ready to act in ways that reduce its disequilibrium with the environment.

In our view, bodily arousal in the form of action readiness already includes some appraisal or evaluation of the environment. The body of the organism is aroused in particular way by

opportunities or challenges the environment offers that matter to the organism. Due to the organism's skills and abilities that have been trained up in the past, the organism is already prepared to do what needs to be done to improve its situation in the world. Constructionists by contrast conceive of core affect not as states of action readiness, but more along the lines of raw sensations that are only given meaning through the cognitive process of situated conceptualization. Lindquist and colleagues describe categorization as functioning like a "chisel, leading people to attend to certain features in a sensory array and to ignore certain others." Categorization is said to take place in a network of brain regions (the DMN) that are engaged when remembering personal experiences (episodic memory) and when imagining future events (prospection) (Bar, 2007; Buckner et al., 2008). Categorization is hypothesized to take the form of representation of prior experiences (Barsalou, 2003). These prior experiences are used to infer what the most likely cause of the current affective changes in one's body might be, and it is this inference that allows for the integration of internal changes in bodily experience, and external sensory perception into a "meaningful psychological moment" (Lindquist et al., 2012, p. 124).

Constructionists might seem to be making common cause with recent embodied theorists of emotion such as Paula Niedenthal who take perceiving and thinking about emotion to involve the perceptual, somatovisceral and motoric reenactment or embodiment of the emotion in oneself. Like constructionists, Niedenthal also takes the concepts involved in thinking about emotion to be modal concepts involving a re-experiencing of past experiences. Consider for instance smiling. There is a clear difference between a felt smile and a false smile, a difference it turns out that can be traced to ways in which the muscles around the eye (the *orbicularis* oculi) contract (Ekman and Friesen, 1982). This so-called Duchenne marker is found in felt, but not in false smiles, and is apparently precisely localizable in the brain (Ekman et al., 1990, though one might question this in the light of the arguments given above). Niedenthal and colleagues have argued that the processes involved in recognizing a false from a felt smile in other people might likewise involve one unconsciously simulating offline the very same muscle movements one makes when genuinely or non-genuinely smiling (Niedenthal et al., 2010). Emotion is thus embodied because experiencing emotion, thinking about emotion, and recognizing emotion in others all require one to literally embody the emotion oneself in one's posture, expression, movements and gestures.

This is an intriguing idea, and the evidence for it is strong. However, there are a number of important differences between Niedenthal's embodied theory of emotion, constructionism, and our embodied theory. For Niedenthal, folk psychological distinctions between emotions can be

precisely mapped onto physiological states which can in turn be correlated with populations of neurons found in sensory, motor and affective regions of the brain (Niedenthal, 2007). She is careful to add that the re-experiencing of physiological states involved in perceiving and thinking about emotion (be it one's own or other peoples) need not be real physiological states of the body. Simulations of these physiological changes will do. The changes taking place in the living body of the organism turn out to be merely peripheral, and all the real action required for generating an embodied emotion takes place in the brain. We argue by contrast that states of the whole living body in the form of action-readiness drive the meta-stable, large-scale pattern of activations that take shape in the brain. Emotion is embodied because it is realized in states of action readiness that mobilize the organism, orienting the body to relevant possibilities and challenges.

Constructionists would, we suspect, also disagree with Niedenthal's embodied theory. She seems to be committed to a discrete theory of emotion according to which distinct physiological and neural states are associated with different basic emotions. Russell (2003) and Barrett (2006) have both argued against a mapping of emotions one-to-one onto behavioral expressions of the kind appealed to in Niedenthal's research. They argue for instance that behavioral expressions of emotion are enormously context-sensitive and exhibit massive situational variance. For example, in some situations I may express my sadness by crying, while in many other situations I may find this sort of open expression of feeling to be inappropriate. We agree with Colombetti (2014), ch. 2 however that a dynamical account of emotions of the type we have been proposing, can account for this context-sensitivity without completely rejecting a discrete theory of emotion. Colombetti argues (and we agree) that emotional episodes may be mapped onto "relatively stable patterns of brain and bodily (including behavioral and expressive) processes." (Colombetti, 2014, p. 48).

The role of past experience in orienting an organism to what is relevant looks very different when viewed from the perspective of the whole living animal in interaction with its environment. Constructionists describe the psychological process of situated conceptualization as involving the reenactment of past experiences, which leads to an understanding of the cause of one's current bodily state of core affect, and creates "a meaningful mental moment in the present" (Barrett and Satpute, 2013, p. 367). Emotion is responsible for giving an organism a meaningful experience of the environment just as we argue, but it does so only with the mediation of processes of situated conceptualization. We've suggested by contrast that affect in the form of action readiness orients the organism to the possibilities for action that matter most

to the organism at the time. The organism finds itself ready to deal adequately with the affordances of the environment, but it does so in large part because of its past experience.

Particularly important are the skills and abilities built up over long period of repeatedly encountering and dealing with the same or similar situations. In sports and music for instance "training, repetition and drill is the concrete foundations on which the structure of play gets erected (Noë, 2009, p. 118). However it is not only in these socio-cultural domains that practice matters; an animal's adequate dealings with the affordances of the environment is always a matter of skill. An animal can respond to an affordance well or badly, and it can get better over time at doing so (Rietveld, 2008). This past history of recurrent interactions with the environment is necessary we suggest for correctly anticipating the outcomes of one's current interactions with the environment. Past experience thus explains how the animal is currently ready to respond adequately to relevant possibilities for action. It is on the basis of this action readiness (which we have just argued enfolds the organism's past experiences) that the organism gives meaning to the environment, and certain possibilities for action stand out as immediately relevant to the organism now. We therefore agree with constructionists that past experience plays an important role in creating a meaningful moment of emotional experience. We disagree however about the form this meaningful moment of experience takes. We argue it takes the form of the whole organism being ready to deal adequately with the relevant affordances of its environment.

This in turn suggests a different interpretation of the constructionist finding that the grouping of areas that makes up the DMN are consistently active for a range of different emotional experiences. We speculate that the spontaneous activity found in this population of neurons gives the organism the ability to accurately and precisely *anticipate* the outcome of its interactions with the environment. (For more details on how we would understand these anticipatory processes we direct you to Bruineberg and Rietveld (2014)).

1.5 Conclusion

Our goals in this paper have been twofold. First we wished to show that cognitive neuroscience may need a different account of cognitive function to that which cognitive psychology supplies. Second we wanted to show that ecological psychology and dynamical systems theory under the heading of radical embodied cognitive science may be able to provide such an account of cognitive function. However if we do look to embodied cognitive science to play this role, this

means giving up on a brain-centred view of cognitive function. We will no longer be able to claim that the brain is the organ of the mind. Instead we will need to think about mind and the cognitive processes that make up the mind at the level of the whole brain-body-environment system. Let us briefly recap our argument.

We began by reviewing some of the problems cognitive neuroscientists have run into in mapping emotional and cognitive functions onto discrete and separate structures in the brain. Instead of discrete brain regions and networks performing specialized emotional or cognitive computational operations, we have discussed evidence that points to extensive mutual influence between classical emotional and cognitive areas of the brain. We then argued that this makes trouble for any attempt to localize function in specific brain areas. We turned our attention next to dimensional or constructivist theories of emotion that share our view that the different emotions as they are understood in common-sense psychology are unlikely to map onto distinct neural circuits. Constructionists argue instead for an account of the different emotions as constructed out of the activity and interaction among domain-general neural networks. However we argued that constructionists may face a similar problem to the discrete emotion theorists they oppose. They may find that the domain-general networks to which they appeal likewise do not have fixed and permanent functions, but may shift their functions in ways that depend on context.

The moral we think of this considerations about brain regions and networks not having fixed and permanent functions is that we need to think of cognitive function in the brain as context-sensitive. We then turned our attention to offering an account of this context-sensitivity. We argued for a view of affect as states of action readiness involving the whole body of the organism. States of action readiness manifest in the body as forms of arousal that are either positively or negatively valued. These bodily states prepare the organism to respond to relevant opportunities and challenges in the environment. These states of action readiness are tracked by interoceptive processes in the brain. The salience network very likely plays a central role in this process (Menon and Uddin, 2010). We then argued that patterns of large scale activity take shape in the brain in ways that are driven by the states of action readiness in the body as a whole. These states of action readiness are elicited by relevant affordances in the environment, and make the organism ready to respond to relevant affordances.

So far our arguments have focused entirely on emotion and how best to understand emotion in the brain. We take our argument however to point to the more general conclusion that cognitive function is best investigated at the level of the whole brain-body-environment system. We take such a conclusion to follow from what we've already argued about emotion and cognition

interactions in the brain. We've seen above that there any separation of emotional and cognitive processes in the brain doesn't hold up in reality. The brain areas that neuroimaging studies identify as being active when people perform tasks that engage emotional and cognitive processes turn out to be in constant and continuous interaction. We've also argued that emotional processes take place in the living body of the organism in its interactions with an environment rich with affordances. Given that there is no separating emotion and cognition it follows that cognitive functions likewise deeply depend on the whole living body of the organism in its engagement with an environment rich with affordances.

Chapter 2

Interoceptive inference: emotion-cognition interactions in the predictive brain

Abstract

In recent years, predictive processing and hierarchical inference have proven to be invaluable tools in describing the complex interaction that takes place between top-down cognitive processing and bottom-up sensory information (Friston et al. 2017; Friston 2010; Corbetta et al. 2008). While predictive processing frameworks have primarily been applied to exteroceptive signals and the ways that we model the outside world, there is growing interest in how the same functional models may be used to describe the processing of interoceptive signals.

Seth (2013) has recently proposed a predictive processing theory of emotional awareness. According to the model, cascading top-down predictions about the source of interoceptive signals counterflow with bottom-up interoceptive prediction errors. The integration of the various predictive representations results in the felt aspect of an emotion. The model is intended to extend traditional cognitive appraisal theories of emotion by filling out the neurocomputational mechanisms underlying the interaction between the affect (e.g. neural and physiological arousal) and appraisal (e.g. memories, evaluations, predictions) elements commonly considered to make up an emotional experience (Scherer 1984, 2001, 2009; Brosch & Sander 2013; Schachter & Singer 1962).

In this paper, we will argue that an interoceptive predictive processing account of emotion needs to be dynamic and non-linear in ways that refute the cognitivist assumptions of existing appraisal theories of emotion. The argument will be in part based on recent neuroscientific descriptions of the relationship between emotion and cognition. As we will see, predictive processing accounts of emotion fit well within dynamic network models of the brain that aim to dissolve the boundaries between emotion and cognition (Pessoa 2014; Lewis 2005). This resemblance sets interoceptive predictive processing firmly against many of the characteristics of appraisal theories of emotion. The aim of this paper will be to highlight some of the ways that predictive

processing can contribute to live debates in emotion theory, as well as suggest how affective neuroscience might in turn facilitate a better understand of our predictive processing theories of mind.

2.1 Interoceptive predictive processing

In recent years, the Predictive Processing (PP) framework has proven a powerful resource when it comes to describing the complex interactions that take place between high-level cognitive and low-level sensory information (Friston 2002, 2010; Corbetta et al. 2008; Howhy 2013; Clark 2013). Very briefly, according to the PP model the brain uses what it has learned about the underlying regularities in the world to make increasingly accurate top-down predictions about the nature of the bottom-up sensory signals it receives moment to moment. Perception in this model is not about sensory signals alone, therefore, but also the brain's best guesses about what those signals mean for the embodied agent. Discrepancies between these predictions and the sensory signals produce 'errors'. These errors help improve the predictive processes in a couple of ways: first, errors can go forward in the system helping to refine the predictions so as to form a better fit with the scene; or second, they can produce the behaviours needed to bring the external scene in line with the predictions. Ultimately, this whole prediction error minimizing regime is 'tuned' by second-order predictions about how reliable the different streams of information are given the predictive agent's current state and the current context. This estimation of reliability, referred to as 'precision weighting', allows the predictive system to flexibly adjust the relative impact (or gain) that certain prediction errors have on the overall process (Friston 2010).

While PP has primarily been used to investigate exteroception, there is a growing interest in how the this same framework may also apply to interoception (Ainley et al. 2016; Allen et al. 2016; Apps & Tsakiris 2014; Barrett & Simmons 2015; Bruineberg & Rietveld 2014; Clark 2015; Gu et al. 2013; Seth 2013, 2014; Seth et al. 2012). Seth (2013) has provided PP researchers an excellent foray into the realm of feelings with his proposed interoceptive predictive processing account (IP). According to Seth, the same predictive story can also be told about interoception. Here, interoception - the internal sense one has of physiological changes in the body, including visceral, vascular, motor and somatic information - is the collaborative result of top-down predictions about the source of the interoceptive signals converging with bottom-up interoceptive signals from the body. Error signals in this model are minimized in an analogous way to the sensorimotor predictions above: either an error modifies the model to fit the affective

state of the organism[9], or autonomic changes are provoked so that the body comes to fit the predictions. The integration of these various predictive representations at multiple levels results in a felt experience.

Seth proposes that IP may help extend appraisal theories of emotion (ATE) by filling out some of the neurocomputational mechanisms underlying the interaction between the affective (e.g. neural and physiological arousal, action tendencies) and appraisal (e.g. memories, evaluations, predictions) elements commonly considered by ATE to make up an emotional experience (Brosch & Sander 2013; Scherer 2001, 2009; Schachter & Singer 1962). While it is true that both IP and ATE aim at explaining how emotions emerge from more domain-general processes, they do so in importantly different ways. In particular, we will show, IP should be seen as dynamic and embodied in ways that refute the overly simplistic and cognitivist assumptions of ATE. Our argument will be based on a dynamic network models that aim specifically at dissolving the boundaries between emotion and cognition (Pessoa 2014). As we will see, IP fits well with such network models. This resemblance, we believe, makes IP irreconcilable with many of the characteristics of both traditional and more contemporary ATE accounts.

The remainder of the paper will proceed as follows: first, we will review ATE as a model, highlighting in particular the tendency to assume an overly simple relationship between affect and appraisal. Second, we will look at both recent network models of the brain (Pessoa 2014) and dynamic systems approaches to thinking about emotion (Lewis 2005) that describe the relationship between emotion and cognition as dynamic and non-linear; and third, we will argue that IP fits better with these dynamic models than with both traditional and contemporary ATE.

2.2 Appraisal theories of emotion

Traditionally, cognition and emotion were described as separable at both the neurobiological level (e.g. cortical vs. subcortical) and the psychological level (e.g. perception vs. arousal). Given that separation, the primary task for emotion researchers was to reconcile these disparate elements, helping to explain how they came together in the generation of an emotional experience. A major contributor to this project are cognitivist, or appraisal type, theories of emotion (ATE).

---

[9] We will follow Barrett and Bliss-Moreau (2009) in their more general use of the term 'affect'. They write, "In the science of emotion, "affect" is a general term that has come to mean anything emotional. A cautious term, it allows reference to something's effect or someone's internal state without specifying exactly what kind of an effect or state it is. It allows researchers to talk about emotion in a theory-neutral way" (ibid., 168).

ATE (e.g. Arnold 1960; Scherer 1999) are motivated by the fact that our cognitive evaluation (or appraisal) of a scene forms a central part of any emotional experience. Perception of an event on its own doesn't seem to be sufficient for producing an emotional response. Rather, emotions arise when an organism evaluates the significance of an event for its well-being (Brosch 2013). If the event is relevant to the organism, cognition induces the physiological changes that produce an emotional experience. For example, a rude gesture only gives rise to anger insofar as it is interpreted as offensive. One advantage of ATE, then, is that it is a handy explanation for the intentionality and rationality of emotions. In virtue of their cognitive component, emotions are about 'things in the world' and can be judged as rational or irrational, appropriate or inappropriate. The cognitive or evaluative aspect of an emotion allows ATE to do justice to the intentionality of emotions, while retaining the importance of bodily events.

'Appraisal' in these theories generally refers to processes that evaluate the significance of a stimulus for an organism's well-being (Brosch 2013: 370; Scherer 2009: 1309). ATE aims specifically at explaining how emotions emerge from the interaction between affective reactions and cognitive appraisals. Consider a famous example from Schachter and Singer's (1962) experiment. Participants in the experiment received an injection which they believed was for vision but was in fact adrenaline. They were then asked to wait in a room with a confederate who acted in a silly or an irritating manner. When later interviewed, the participants (but not the control subjects) identified their adrenaline induced arousal as either joy or irritation depending on the context that included the confederate. Schachter and Singer concluded that emotions are the result of largely undifferentiated patterns of neural and visceral arousal *and* the individual's best guess at what that reaction could mean. The famous catch line: an affective reaction is necessary, but not sufficient, for a particular emotional experience to occur. While the findings and methods have been open to criticism (for a review see Prinz 2003), the central conclusion that emotions take more than raw bodily activations is widely accepted today (see also Critchley 2005). They must also in some way include an interpretation of what is happening in the environment and its relationship to the organism. Notice that in Schachter and Singer's two-factor model the obvious separation between bodily affectivity (described as "a general pattern of excitation of the sympathetic nervous system"; 1962: 379) and the cognitive appraisals that make sense of those activations. This fits a general pattern in ATE models of characterizing cognitive appraisal and emotional affect as discrete entities.

How are these entities related? In traditional ATE theories, appraisal and affect are related to one another in simple and causally linear ways. Newer ATE models, however, take into account

possible dynamic and recursive relationships between cognitive and emotional processing. For instance, Scherer (2004, 2009) presents an appraisal model of emotion called 'The Components Process' model (CPM). This model focuses on explaining the dynamic unfolding of an emotion over time caused by an individual's evaluation of an event as significant to her goals or needs. According to appraisal theories, an individual is constantly evaluating whether objects or events around her are relevant to her goals, needs and values. Appraisal mechanisms check for criteria such as novelty, valence (whether the event is pleasant or unpleasant), relevance (the significance of the event for the agent or her social group), the implications of the event, coping potential (how well the individual can adjust to the consequences of the event) and normative significance (the compatibility of the event with one's goals and values) (Scherer 2009: 1309).

The CPM emphasizes that appraisal along these criteria occurs at multiple levels of processing. Appraisal of novelty, for instance, can occur quickly and automatically via subpersonal processes at the neural level. It can also occur as the result of a deliberate comparison at the level of conscious thought. Appraisal processes also require the coordinated activity of many neural structures, and the CPM includes bidirectional effects between appraisal processes and cognitive and physiological functions. For instance, a minimal amount of attention is needed for the appraisal process to begin but, if an event is perceived as relevant, more attentional resources will be recruited for the stimulus (Scherer 2009: 1314).

Taking an event to be relevant sets in motion the motivational systems needed to respond to that event. The results of the appraisal and motivational changes induce changes in the autonomic system and somatic system (such as increased heart rate and a changed facial expression). These changes are then given a unified representation by a multimodal integration area, and this representation is continuously updated according to changes in events and appraisals. The unified representation can then become conscious and available to be labelled as an emotion. The CPM sees the role of emotions as preparing the organism to respond to significant events and, in some cases, preparing the organism to take one of any number of possible courses of action in response to an event (Scherer 2009: 1308). Nonetheless, it is important to note that emotions are not a sufficient cause of action. Behaviours are seen as complex, with emotions being among the factors that prepare an organism for action.

Essentially, the CPM claims that emotion processing arises from interacting componential subsystems, including appraisal processes, physiological processes, motor expressions and action tendencies. The subsystems are interdependent such that changes in one subsystem result in changes in the others. Additionally, the appraisal process and subsequent physiological, somatic

and motor unfolding is recursive, with multiple feedforward and feedback mechanisms between subsystems.

Scherer's CPM specifically aims to do justice to the neurophysiological evidence on emotion processing, citing the following as "central design features" of emotion: "(i) the dynamic, recursive nature of emotional processing; [and] (ii) the central, and causal, role of multi-level cognitive processing of both antecedent events and response options" (Scherer 2009: 1307).[10] Nonetheless, it is by no means the only current appraisal theory to do so. Brosch and Sander (2013) present an "appraising brain" model of emotion that rejects the 'feedforward' view of information processing in the brain. Their model involves recursive processing cycles through which the brain develops an increasingly finer-grained evaluation of a stimulus. Because cognitive and physiological information is integrated into stimulus processing at multiple latencies, the model does not draw a hard and fast distinction between cognitive and physiological components of evaluation. In their account:

> "A linear view of information processing, where information is first analyzed in the sensory cortex then moves "upward" to regions underlying more complex processing to finally arrive in the prefrontal cortex, has been replaced with models emphasizing that information flow in the brain occurs in multiple sweeps of activation, with numerous feedforward and feedback loops that refine neural processing patterns and the underlying computations with each iteration […] This allows for the dynamic, increasingly more complex evaluation of a stimulus, highly compatible with the predictions of appraisal theory" (ibid.,166).

Although new ATE accounts take pains to accommodate neuroscientific evidence on recursive and dynamic processing they still fall prey to certain cognitivist assumptions, dividing emotional and cognitive processing into decomposable subsystems and privileging the latter.

In the following sections, we will argue that the ATE model is based on an inaccurate interpretation of the neural interactions underpinning emotional experience. As we will see, recent dynamic network models of the brain highlight neural connectivity that dissolves any clear separation between cognitive and emotional processing. In the next section, we draw on Giovanna Colombetti's work, which critiques ATE on the basis of this interconnectivity between emotion and cognition at the level of the brain.

---

[10] Although Scherer's model incorporates dynamic and recursive neural processing, appraisal criteria are processed sequentially and in a strict temporal order. This is a distinctive feature of Scherer's ATE model, so we will not focus our argument on it.

2.3 The cognition-emotion divide in ATE

Scherer claims to offer a dynamic and recursive account of emotion because it incorporates continuously changing cognitive and motivational processes. However, Colombetti (2007) argues that this is insufficient to make his account genuinely dynamic. She claims that the CPM is still modular because it "posits cognition (appraisal) as a separate component inside the emotion system, which merely interacts with other components" (2007: 97). Bodily events are not part of the self-organizing network of processes that give rise to cognition. The body influences cognition, mostly indirectly, but it does not form a part of the cognitive process itself. Cognition still occurs only in the head.

According to Colombetti, Scherer's model reproduces the flaws of cognitivist models insofar as psychological functions are assigned to discrete subcomponents of the system as opposed to being emergent features of the system. For instance, the affective and cognitive components of Scherer's model do not emerge from the interactions of several subcomponents (which are themselves neither cognitive or emotional). Rather, they are controlled by dedicated systems. Moreover, appraisal belongs exclusively to the cognitive system. Arousal and bodily events modulate cognitive processing, but do not play a part in appraising themselves.

ATE imagine cognitive appraisals as doing all of the heavy lifting when it comes to determining the character of the felt emotion: they appraise the context, categorize the stimulus as either pleasant or unpleasant, initiate physiological changes and action tendencies, and reappraise the object and situation over time so that the agent can adjust their response. Affect is often conceptualized as the body's reaction to cognitive appraisals. As Colombetti so elegantly puts it, "The body in the two-factor theory is by itself 'naked', so to speak, and it needs to be cognitively "dressed" to give rise to emotion" (Colombetti 2013: 88). Colombetti, using Susan Hurley's terminology (1998), calls this approach "vertically modular": though arousal and cognition interact, they are controlled by distinct and easily separable subsystems. This means that, though it claims to focus on the dynamic nature of emotional processing, the CPM is not genuinely dynamic in the way espoused by proponents of dynamical systems theory in cognitive science (e.g. Lewis 2005). According to a dynamic systems model, appraisal, arousal and other psychological functions emerge from the interaction of several microcomponents. Importantly, these microcomponents do not themselves have any specific function independently of the large-scale patterns in which they participate.

In the following section we will briefly review Pessoa's collection of findings on cognitive-emotional processing in the brain and examine why Colombetti and Lewis take such findings to pose a problem for models such as CPM.

2.4 Dissolving the boundaries between appraisal and affect

A once-popular neuroscientific paradigm imagined the human brain as evolving in layers with phylogenetically newer and more advanced circuitry overlaying and directing older more primitive-instinctual circuits (Herrick 1933; Papez 1937). From that perspective, cognition and emotion were naturally assumed to be separable at the neural level - being neatly mapped onto the higher-cortical and lower-subcortical areas of the brain respectively. However, as our neuroscientific techniques improve we are increasingly aware that such a dichotomous view of neural processing is too simple to capture the complex, reciprocal and self-organizing nature of human brain function. This is especially true when discussing the the division between cognition and emotion.

Drawing on a wealth of neuroscientific research Pessoa argues for the dissolution of the boundary between emotional and cognitive processes at the level of the brain (2013). Pessoa shows that brain regions typically characterized as emotional or cognitive perform functions that are characterized as both emotional and cognitive in nature. For example, the prefrontal cortex (PFC), which is often characterized as contributing to cognitive functions due to its central role in the maintaining and manipulating information in the brain, also plays an important role in emotional processing such as regulating approach and avoidance behaviour (Pessoa 2008: 150). Or consider that the amygdala - a paradigmatic emotional area traditionally believed to play a role in fear processing - is widely understood today to be key in directing attention and supporting associative learning (Hollard & Gallagher 1999). Pessoa concludes that "parceling the brain into cognitive and affective regions is inherently problematic, and ultimately untenable" (2008: 148).

Pessoa's argument rests on a dynamic network perspective of neural structure and processing (van den Heuvel & Sporns 2011, 2013a, 2013b). In contrast to more traditional models (that assumed connections between brain regions to be relatively simple), neural networking models describe the brain as being composed of 'functional clusters': regions within which every area is reciprocally connected to every other area (Pessoa 2014: 10). In turn, these clusters communicate and collaborate with one another via far reaching "connector hubs": centrally positioned brain regions that enjoy massive connectivity with the rest of the brain (Pessoa 2008: 154). A prime

example of such a hub is the amygdala, a subcortical area at the topographical centre of the brain that is densely and reciprocally connected with all but 8 cortical regions (Young et al. 1994).

Intelligent behavior, according to this model, emerges from the transient, large-scale, functional networks where circuits are continually being connected and disconnected in various patterns over time. In this model, behaviours are not implemented by individual circuits or areas, but emerge instead from the interactions of multiple areas (Pessoa 2008; Anderson 2014). One important conclusion that emerges from this perspective is that the functional profile of any given area will inevitably change relative to the transient large scale pattern the area is affiliated with at the moment (Pessoa 2008: 154; Mesulam 1990).

While there may be neural clusters at the far peripheries of the brain, due to their lack of connections with neighbouring areas, that may be characterized as either cognitive or affective; even these circuits contribute to overall function through their participation in much larger patterns of brain activity. The large-scale patterns that emerge are inevitably composed of what might be characterized as both emotional and cognitive circuits, significantly eroding our ability to cleanly divide processing into simply emotional or cognitive categories. Complex behaviours are seen as emerging from these large-scale emotional-cognitive patterns, which continually shift according to the context and the goals of the animal. Pessoa concludes, "there are simply no truly separate systems for emotion and cognition because complex cognitive-emotional behavior emerges from the dynamic interactions between brain networks" (Pessoa 2008: 148).[11]

Colombetti argues that dynamic network models such as these pose a problem for appraisal theories: if there is no division between cognitive and emotional processing at the level of the brain, then psychological accounts that characterize cognition as a distinct module are unjustified. To help illustrate this point, Colombetti references Marc Lewis' dynamic systems theory of emotion (Lewis 2005, 2015; Lewis & Liu 2011; Lewis & Todd 2005, 2007). Lewis draws on neuroscientific evidence to show that the broad psychological categories of 'emotion' and 'appraisal' cannot be cleanly mapped onto the brain. He argues that an inability to differentiate cognitive from emotional functions at the neural level implies that it is untenable to attempt to do so at the psychological level (since these distinctions lack a neural basis). Maintaining the division between cognitive and emotional processing thus creates a jarring difference between the way functions are individuated at the neural vs. the psychological levels.The distributed nature of neural processing means that the psychological functions -

[11] There is a growing literature in cognitive neuroscience that questions whether any neural resources are simple components in the way previously imagined (see Anderson's work on neural reuse; 2010).

typically taken as cognitive and emotional - interact in such rapid and mutually dependant ways that we cannot separate them from another. This is further illustrated in the following section where we examine Lewis' dynamical theory of emotions.

2.5 Lewis' dynamical systems theory of emotion

Marc Lewis outlines an approach to emotion research that emphasizes the inseparability of affective and cognitive elements. He uses dynamical systems modelling to help bridge emotion theory with the dynamic network models of the brain. In what follows, we will summarize the dynamic systems approach and Lewis's application of the theory to emotion.

Dynamic systems theory (DST) explores the mathematical means to model changing systems in temporally sensitive ways. As real world systems rarely exist in a vacuum, dynamic modelling often includes how closely related systems influence one another (Van Gelder 1995: 358). When neighbouring systems are so close that state changes in one system necessarily result in state changes in the other, those systems are described as 'dynamically coupled'. According to DST, coupled systems are not independent systems, but rather are better thought of as comprising a single changing system (Van Gelder 1995: 357; see also Kelso 1995). An important characterization of any such dynamic system is a tendency to 'self-organize' (Lewis, 2005). This means that no subsystem acts as the 'controller' of other systems, but rather, "coherent wholes emerge and consolidate from interacting constituents" (Lewis 2005: 173).

The language of self-organizing dynamic systems is an ideal tool for discussing mental phenomena. As Lewis writes, "dynamic systems operate through reciprocal, recursive, and multiple causal processes, offering a language of causality consistent with the flow of activity among neural components" (Lewis 2005: 169). Lewis picks out various psychological components commonly described as appraisal processes (e.g. perception, evaluation, attention, memory, and higher-order executive functions), and as affect processes (e.g. action tendencies, attentional orientation, and affective feelings). According to Lewis, when we shift from one emotional state to the next, elements of appraisal and affect collaborate in highly interdependent ways.

Consider the complex interactions that occur between affective and appraisal components as part of sexual attraction. Erotic cues impact various brain systems (perhaps via the amygdala) in fast and pre-attentive ways (Diano et al. 2017). This initial activation triggers innate sexual

reflexes (patterns of arousal and action readiness) and learned sexual scripts (memory and planning). These arousals and expectations further entrain attentional processing on the erotic cues, which in turn produces a feedback loop between the autonomic arousal and the narrowing of attentional and decision making processes - which ultimately, if unimpeded, leads to approach behaviours. Once the change in emotional state is initiated, these affective and appraisal elements quickly become coupled. As Lewis writes, "according to this model, appraisal activities and emotional response activities cause one another, each activating, propelling, and guiding the other, reciprocally and recursively... hence what evolves is not just an appraisal but an emotion-appraisal amalgam or 'emotional interpretation'" (Lewis 2005: 176). Much like the dynamic network views outlined above, in which neural circuits (commonly associated with cognition or emotion) function by collaborating in larger-scale patterns of activation, appraisal and affective processes become similarly entangled.

Colombetti highlights how both Pessoa and Lewis argue against the vertical modularity of emotion and cognition. Emotion and cognition, according to dynamic network models, refute easy localization to specific neural regions and, as expressed above, the subpersonal processes constituting emotion and cognition relate in dynamic and non-linear ways. Colombetti concludes, "[i]f this is the appropriate way to conceptualize the cognition-emotion relationship, then to characterize appraisal as a separate cognitive process not overlapping with (other) emotional components (as the CPM, for one, does) is misleading, because it does not do justice to the real complexities of the neural level" (Colombetti 2013: 100). In contrast to ATE, a cognitive evaluation is not the only cause of an affective reaction, nor is it the role of cognition simply to appraise the body's reactions. Instead, these elements co-evolve, working to elicit and sustain each other over time.

In the following section, we will apply these insights to the interoceptive predictive processing theory. We will argue, based on neuroscientific evidence on AIC function (and a comparison with other PP theories of emotion), that IP is better suited to a dynamic model of emotion than an appraisal model of emotion.

2.6 Interoceptive predictions are complex, self-organizing and embodied

Seth proposes that IP can extend ATE by detailing some of the neurocomputations that link affective changes and cognitive appraisals. Seth is right to make this connection insofar as both

models are clearly interested in explaining how emotions emerge from the interaction between more basic appraisal and affective processes. However, as we will see in the following sections (which build further on the neuroscientific evidence discussed earlier), what the subcomponents of emotion are and how they are thought to relate is significantly different according to the two theories. IP emerges as a much better fit with more dynamical models of emotion.

IP characterizes an emotional experience as emerging from the tight coordination of bottom-up bodily information and top-down cognitive and perceptual processes. In fact, a hallmark of PP models is this interdependent relationship between bottom-up and top-down information (Dayan et al. 1995; Friston 2002, 2012; Corbetta et al. 2008). According to IP, errors signals are explained away by either a modification of the generative model (so as to make a better fit with the interoceptive signal) or by motivating autonomic changes (eg. heart rate, respiratory rate, smooth muscle behaviour) to better fit the model (Seth 2013: 566). Seth's account highlights the fact that error minimization takes place within two neurophysiological loops simultaneously: a first loop stretches between the high-level prediction and changes in the internal milieu (interoceptive), and a second loop from prediction to overt actions and expressions (exteroceptive). These loops between brain and body are described as deeply interdependent: the ebb and flow of each cycle initiating, maintaining and restraining the next cycle. Subjective feelings emerge in Seth's model from the co-evolution of these streams of information. The self-organizing and interdependent nature of these predictions refutes the description popular in ATE of a clear separation of affective and appraisal processes. To make this point more clearly, it will help to look at processing in the anterior insula cortex (AIC), the central instantiator of IP in the brain.

2.6.a Interceptive predictions and the insular cortex

The strongest evidence for IP comes from research on the AIC (see also Allen, 2014). According to Seth, (2013) the AIC is a central component of the neural instantiation of predictive processing of emotion. In this section, we will review the neuroscientific evidence on the function and connectivity of the AIC. In doing so, we will illustrate that its role in generating emotion fits with the dynamical systems models described in the previous sections. The AIC is what Pessoa calls a connector hub - it is richly connected with myriad other brain regions, it is functionally diverse, and it does not play a specifically cognitive or affective role in emotional processing.

AIC function was previously thought to be limited to feelings of disgust (e.g. Jabbi et al. 2008; Wicker et al. 2003). This hypothesis emerged from the observation that diseases that affect the AIC (such as Huntington's and Parkinson's) are often accompanied by lowered disgust reactions (Mitchell et al. 2005) and difficulty perceiving disgust in others (e.g. Kipps et al. 2007; Sprenglemeyer et al. 1996; Suzuki et al. 2006). As we saw above, such locationist views are increasingly giving way to dynamic models of brain function, in which neural circuits play various roles depending on the larger structural and functional networks they participate in. For example, while there is some evidence that specific neural patterns may instantiate particular emotions (Panksepp 2004; LeDoux 1996), many of the same functional systems of neurons are found to participate in a wide variety of emotional and cognitive processes (Lindquist et al. 2012). The AIC is an ideal case of a brain area that wears many 'functional hats'.

The AIC is believed to be one of the most functionally diverse areas in the brain, activated for a wide range of both emotional and cognitive tasks (Anderson et al. 2013). The AIC has been found to be involved in in almost all emotional processing (Kober et al. 2008; see also Ackermann and Riecker 2010; Craig 2010; Garavan 2010). Moreover, the AIC has also been found to play a key role in many cognitive processes, including predicting what will happen next in the body given the current context (Singer et al. 2009; Damasio 1999). For example, in a study on anxiety, Paulus and Stein (2006) characterized neurons in AIC as computing an *interoceptive prediction error* when anticipated, and experienced bodily responses to aversive stimuli were mismatched. When a similar mismatch was induced using false cardiac feedback by Grey and colleagues (2007) increased dorsal AIC activation was correlated with an increase in emotional intensity/salience to the stimuli. This research characterizes the AIC as an area uniquely capable of integrating interoceptive and exteroceptive signals with past experiences and future expectations. IP has been proposed as an elegant neurocomputational explanation of how these processes might come together: through a continual reducing of error between top-down predictions and bottom-up interoceptive signals (Gu et al. 2013; Seth 2013; Allen et al. 2016).

According to Bud Craig's influential account of the insula (an area heavily referenced in the IP literature; Clark 2015; Barrett & Simmons 2015; Seth & Friston 2016) interoceptive information is processed in a posterior-to-anterior progression in an increasingly complex fashion with various streams of internal, external and contextual information being integrated along the way (Craig 2002, 2009). Feeling states represented in the insula in fact emerge from the collaboration of these various streams. It is important to note that this process does not follow a simple feed-forward progression from sensory signal to representation (like Damasio 2000; Craig 2009; Singer et al. 2009). Seth demonstrates an awareness of this fact when he writes, that such an

assumption is "challenged by evidence of substantial cross-talk between levels of viscerosensory representation, including top-down cortical and behavioural influences to brainstem and spinal centres (Seth 2013: 567; see also Critchley & Harrison 2013). IP relies on a characterization of the AIC as capable of simultaneously integrating various streams of information from within the body with contextual signals from higher-order cortical areas (such as orbitofrontal cortex, cingulate cortex, etc.) *and* continually modifying internal states of the body via massive descending systems (see Gu et al. 2013a). Gu et al. make a similar point in their own account of interoception in predictive processing that also focuses on the AIC. They write:

> "The AIC both responds to and controls the internal milieu or literally "gut feelings". The AIC is perfectly placed anatomically to do this; it is equipped with the anatomical and functional foundation to perform the very important task of inducing transitions in physiological states... neurons in AIC innervate the viscera directly and indirectly for example through projections to the hypothalamic area via the amygdala. In short, the AIC is able to cause changes in the physiological states of the body, in addition to perceiving changes from the body" (2013b: 3382).

As Seth notes, these two directions of interaction (brain-to-body and body-to-brain) unfold "continuously and simultaneously underlining a deep continuity between perception and action" (Seth 2013: 566). According to IP, the brain's best guess at the cause of some perceptual signal reactivates similar somatic patterns associated with such stimuli in the past (see also Damasio 1994). This reactivation prepares the body to respond to the prediction appropriately *and* helps the system predict what will happen next. The reactivation includes both autonomic changes and explicit actions (including gestures, facial expressions and postural changes). These bodily changes provide the basis for the next wave of interoceptive information to be integrated and matched against the evolving prediction (which includes exteroceptive information, memories and predictions), and it is this that will eventually give rise to an emotional experience. Prediction and incoming signals co-evolve in fast back-and-forth succession, attempting to minimize discrepancies between model and signal. The emergent, integrated representation is then re-represented and made accessible to conscious awareness (Harrison et al. 2010; Craig 2002, 2009). Here, the AIC instantiates exactly the sort of emotion-cognition entanglement discussed by Pessoa and Lewis in their arguments for dynamical models of neural and psychological functioning.

ATE models characterize emotion as emerging from the interaction of multiple subcomponents; as we saw above, the components themselves remain distinct, as "modules" subserved by dedicated circuits. However, this characterization of the AIC - weaving together emotional and cognitive information via multiple waves of interaction between brain and body - aligns IP much

more closely with dynamic models of emotion that aim to dissolve the dichotomy between affect and appraisal.

2.6.b A predictive account of seeing with feeling

Although we are still coming to understand how information flows within the brain, we can perhaps look towards other predictive processing accounts that investigate interoception to help underscore the idea that IP is a better fit with the more dynamical and self-organizing views of emotion described above. A parallel argument has recently been made for the influence of interoception on visual perception. Barrett and Bar (2009) argue that our best account of the orbitofrontal cortex (OFC) indicates that bodily affect contributes, in a constitutive way, to visual predictions. This reflects our position that the neural connectivity of the AIC indicates that bodily affect has a role to play in emotional appraisal. Below we review the argument made by Barrett and Bar regarding the OFC and point to its relevance for interoceptive predictive coding.

The orbitofrontal cortex (OFC) is divided into lateral and medial parts. Medial OFC takes rough visual information from subcortical areas and makes an extremely fast guess at what such information could mean. This 'best guess' flows through the hypothalamus, midbrain, brainstem and spinal column into the perceiver's body, recreating the internal affective state previously associated with that stimulus. This affective reactivation loops into pre-motor areas that transform the affective signal into an action plan (e.g. approach or avoid; Damasio 1994; Gallese 2005). Simultaneously, the lateral OFC works to integrate the initial affective response (including autonomic and endocrine changes) with increasingly rich visual information. As the object becomes clearer, the OFC 'updates' its prediction, which in turn invokes subsequent affective reactivations and reactions (which further fills out the percept). This recursive process continues until a detailed representation of the object is finally constructed (Barrett & Bar 2009: 1328; see also Bar et al. 2006). The body in this model plays a key role in the unfolding of the prediction that constitutes our experience. Bodily affect is not merely a signal to be interpreted by cognitive systems, but itself plays an active role in evaluation. As Barrett and Bar write, "the internal world of the body may be one element in the 'context frame' that facilitates object recognition" (Barrett & Bar 2009: 1330). The emergent visual experience, then, is the result of both neural and physiological signals evolving together over time.

Barrett and Bar's take on the relationship between affect and visual processing thus offers a predictive processing version of Colombetti's argument for the non-decomposability of

emotional and cognitive processes: affective changes and cognitive appraisals relate to one another in nonlinear, co-evolving ways, arising from subprocesses that are not themselves cognitive or emotional. The result is that affect is seen as an equal partner in the generation of our experience. Barrett and Bar conclude that they have "laid the foundation for the hypothesis that people do not wait to evaluate an object for its personal significance until after they know what the object is. Rather, an affective reaction is one component of the prediction that helps a person see the object in the first place" (Barrett & Bar, 2009: 1331)

There are important parallels between Barrett and Bar's account of the OFC and Seth's account of the AIC. The OFC and AIC are, in fact, considered to process information in very much the same way, but for external and internal experience respectively (Lewis & Todd 2005: 20). Craig has gone so far as to call the insular cortex the 'sensory cortex of the limbic system', because of the way information is hierarchically processed in multiple waves (2002). What Barrett and Bar's framework reveals, however, is that affective elements arising from physiological states should not be seen as something that occurs in reaction to objects that have already been perceived - rather they are integral to the process that allows perception to occur in the first place. Similarly, we have argued, that visceral states of the body are coupled with appraisals in the formation of internal experience.

Barrett and Bar have presented us with a network depiction of the predictive brain, one in which higher-level predictions and lower-level affective changes are seen as interacting in highly dynamic and nonlinear ways (refuting a view of vertical modularity, as Colombetti described above). Seth's opening bid for a predictive processing account of interoception in fact relies on just this type of rich integration of information (particularly in the AIC). However, we believe that by highlighting those particular brain structures and functions as part of characterizing IP we must also accept a more dynamic and embodied view of interoception.

2.7 Conclusion

One of the most exciting features of the predictive processing framework is its wide applicability to a multitude of cognitive functions. Seth's interoceptive predictive processing model offers an excellent starting place for thinking about how the predictive processing mechanism may apply to feelings and emotions.

In this paper we argue against Seth's suggestion that a predictive processing version of interoception should be thought of as an extension to appraisal theories of emotion. While IP certainly captures some of the main features of ATE, characterizing IP as an appraisal model assumes an overly cognitivist reading of interoceptive processing. We argue that a predictive processing account of interoception is in fact dynamic in ways that run counter to the cognitivist assumptions native to both traditional and contemporary appraisal theories of emotion.

The neurobiology underlying reciprocal top-down and bottom-up processing suggests a substantially more dynamic and embodied PP story than the one due to Seth. We argued for this view by reviewing recent neuroscientific descriptions of the relationship between emotion and cognition, revealing them as essentially non-decomposable at both the neurobiological (Pessoa) and psychological (Lewis) levels. Given the reciprocal interaction between emotional and cognitive processes highlighted by these frameworks, and the fact that IP relies on the very same densely interconnected neural regions highlighted by Pessoa's and Lewis's accounts, we conclude that an interoceptive predictive processing comports best with non-linear dynamic systems approaches to emotion rather than with standard appraisal models.

Chapter 3
Happily entangled: prediction, emotion, and the embodied mind

Abstract

Recent work in cognitive and computational neuroscience depicts the human cortex as a multi-level prediction engine. This 'predictive processing' framework shows great promise as a means of both understanding and integrating the core information processing strategies underlying perception, reasoning, and action. But how, if at all, do emotions and sub-cortical contributions fit into this emerging picture? The fit, we shall argue, is both profound and potentially transformative. In the picture we develop, online cognitive function cannot be assigned to either the cortical or the sub-cortical component, but instead emerges from their tight co-ordination. This tight co-ordination involves processes of continuous reciprocal causation that weave together bodily information and 'top-down' predictions, generating a unified sense of what's out there and why it matters. The upshot is a more truly 'embodied' vision of the predictive brain in action.

3.1 The strange architecture of predictive processing

In 2012 the AI pioneer Patrick Winston wrote about the "puzzling architecture" of the brain—an architecture in which "Everything is all mixed up, with information flowing bottom to top and top to bottom and sideways too." He added, " It is a strange architecture about which we are nearly clueless" (Winston 2012).

It is a strange architecture indeed. But that state of cluelessness is increasingly past. A wide variety of work—now spanning neuroscience, psychology, robotics and artificial intelligence is converging on the idea that one key role of that downward- flowing influence is to enable higher-levels to attempt (level-by-level, and as part of a multi-area cascade) to try to predict

lower-level activity and response.[12] That predictive cascade leads all the way to the sensory peripheries, so that the guiding task becomes the ongoing prediction of our own evolving flows of sensory stimulation. The idea that the brain is (at least in part, and at least sometimes) acting as some form of prediction engine has a long history, stretching from early work on perception (Helmholtz 1860; MacKay 1956; Neisser 1967; Gregory 1980) all the way to recent work in deep learning (Hinton 2007, 2010).

A promising subset of such work is the emerging family of approaches known as 'predictive processing'.[13] Predictive processing plausibly represents the last and most radical step in the long retreat (see Churchland et al. 1994) from a passive, feed- forward, input-dominated view of the flow of neural processing. According to this emerging class of models biological brains are constantly active, trying to predict the streams of sensory stimulation before they arrive. Systems like that are most strongly impacted by sensed *deviations* from their predicted states. It is these deviations from predicted states (prediction errors) that now bear much of the information-processing burden, informing us of what is *newsworthy* within the dense sensory barrage. When you see that steaming coffee-cup on the desk in front of you, your perceptual experience reflects the multi-level neural guess that best reduces visual prediction errors. To visually perceive the scene, if this story is on track, your brain attempts to *predict* the scene, allowing the ensuing error (mismatch) signals to refine its guessing until a kind of equilibrium is achieved.

To appreciate the benefits, first consider learning. Suppose you want to predict the next word in a sentence. You would be helped by knowledge of grammar. But one way to learn a surprising amount of grammar, as work on large-corpus machine learning clearly demonstrates, is to try repeatedly to predict the next word in a sentence, adjusting your future responses in the light of past patterns. You can thus use the prediction task to bootstrap your way to the world-knowledge that you can later use to perform apt prediction. Importantly, learning using *multi-level* prediction machinery delivers a *multi-scale* grip on the worldly sources of structure in the sensory signal. In such architectures, higher levels learn to specialize in predicting events and states of affairs that are—in an intuitive sense—built up from the kinds of features and properties (such

[12] There is a large and growing literature here. Good places to start include Friston (2005, 2010), Clark (2013), Hohwy (2013) and Clark (2016).

[13] For this usage, see Clark (2013).

as lines, shapes, and edges) targeted by lower levels. But all that lower-level response is now modulated, moment-by-moment, by top-down predictions.[14]

To make the best and most flexible use of the flow of prediction error PP architectures simultaneously estimate the so-called "precision" of the prediction error signal itself. Precision is the inverse variance of a prediction error signal—in other words, it sets error bars around an error signal according to its currently estimated importance or reliability. High-precision errors enjoy greater post-synaptic gain and (hence) increased influence. Conversely, even a large prediction error signal, if it is assigned extremely low precision, may be rendered systemically impotent, unable to drive learning or further processing. This enables different circumstances to render differ- ent prediction error signals important, and may mandate different balances between processing in different brain regions and between top-down prediction and incoming sensory evidence.

Action itself is accomplished using the same resources. The core idea here (Friston et al. 2010) is that there are two ways for brains to match their predictions to the world. Either find the prediction that best accounts for the current sensory signal (perception) or alter the sensory signal to fit the predictions (action). Importantly, the flow of action can *itself* be brought about, PP suggests, by a select sub-set of predictions — prediction of the (trajectory of) proprioceptive consequences that would ensue were the desired action to be performed. This turns out to be a computationally efficient way of implementing motor commands (Friston 2011).
A central claim of this 'active-inference' view is that top-down predictions and bodily actions co-evolve in circular and self-organizing ways. Friston and colleagues write, "Crucially, this inference or assimilation is active, in the sense that the internal states affect the causes of sensory input vicariously, through action. The resulting circular causality between perception and action fits comfortably with many formulations in embodied cognition and artificial intelligence; for example, the perception–action cycle (Fuster 2004), active vision (Wurtz et al. 2011), the use of predictive information (Ay et al. 2008; Bialek et al. 2001; Tishby and Polani 2011), and homeokinetic formulations (Soodak and Iberall 1978). Furthermore, it connects these perspectives to more general treatments of circular causality and autopoiesis in cybernetics and synergetics (Haken 1983; Maturana and Varela 1980)" (2014, p. 443).

---

[14] This helps make sense of recent work showing that top-down effects (expectation and context) impact processing even in early visual processing areas such as V1—see Petro et al. (2014) and Petro and Muckli (2016). Recent work in cognitive neuroscience has begun to suggest some of the detailed ways in which biological brains might implement such multi-level prediction machines—see Bastos et al. (2012).

The resulting picture is one in which perception and action are complementary manifestations of a single adaptive regime, whose core operating principle is the reduction of precision-weighted prediction error.

Adaptive predictions cannot, however, take shape in an organismic vacuum. What my brain predicts, moment-by-moment, needs to be delicately geared to what I need, and to what I need to be doing. And what I need, and what I need to be doing, are both matters that depend heavily upon both my current physiological states and the shape and progress of current world-engaging activity. It is here that sub-cortical structures, and especially the thalamus (and within it, the pulvinar) seem posed to play a special and crucial role. Understanding that role requires us to move beyond what Pessoa (2014, p. 11) describes as the 'corticocentric' image in which evolutionary older subcortical structures are dominated and controlled by the more recent cortical overlay. Instead, we will be led to endorse an 'embedded' view (op. cit., p.14) according to which cortical and sub-cortical states and activities change in a co-ordinated fashion characterized by ongoing patterns of mutual influence.

3.2 Continuous reciprocal causation in cortico-subcortical loops

The term *cortical myopi*a was first coined by Parvizi (2009) in reference to a lingering tendency in contemporary neuroscience to under-appreciate or to ignore the rich contributions sub-cortical brain regions make to higher cognitive function and intelligent behavior. The bias comes to us in part as a hangover from 19th century experimental brain research (see LeDoux 1987). At the time human brain evolution was imagined to be a largely linear affair, with phylogenetically newer and more advanced cortical areas overlaying and controlling older more primitive subcortical areas (Herrick 1933; Papez 1937). With higher cognitive functions such as language seen evolving alongside the massive expansion of the neocortex (Barton and Harvey 2000) researchers naturally assumed higher cognition to be localized in the uppermost cortical tip of the neural axis. Together with Charles Darwin's astute observations that basic emotions (eg. fear and rage) are shared across species, this led researchers to look for emotional/instinctual circuits in the older and highly conserved sub-cortex. Their conclusion was that human rationality emerged as the new and improved neo-cortex exerted increasing control over the outdated emotional-instinctual circuitry. As John Hughlings-Jackson wrote, "the higher nervous arrangements evolved out of the lower keep down those lower, just as a government evolved out of a nation con- trols as well as directs that nation" (Jackson 1884, p. 662, quoted by Parvizi 2009, p. 354). This picture of the brain has produced a long-standing tradition in cognitive

neuroscience of investigating cortical and sub-cortical structures as dichotomous sets of regions with "higher" circuits controlling/inhibiting the "lower" circuits (perhaps most dramatically in MacLean's (1990) 'triune' brain model).

The major issue with such corticocentric views is not that the brain might be usefully described in hierarchical terms (see Lewis and Todd 2005), nor is it the claim that processing 'higher' up the neural axis is essential for cognitive functions such as decision making and language (which they most certainly are). What needs to be doubted is the assumption of a clear division of labor between a higher 'cognitive brain' and subordinate 'emotional brain' and the assumption that higher areas influences lower ones, but not the other way around. As neuroscientific techniques improve, it is becoming increasingly clear that such a dichotomous view of neural processing is too simplistic to capture the complex, reciprocal and self-organizing nature of human brain function.

Contrary to the Victorian view, the cortex is not a total newcomer to human brain evolution, but has in fact long been part of the basic mammalian neural floor-plan (Pessoa 2013). Moreover both cortex *and* sub-cortex have continued to change throughout human evolution. For example sub-regions of the human amygdala are believed to be 60% larger than apes' relative to brain mass (Sherwood et al. 2012). Compare this with only a 24% increase in neocortical mass (Rilling and Insel 1999), and surprisingly no significant increase in frontal lobe mass (Semendeferi et al. 2002). And finally there is growing evidence that cortical and sub-cortical areas evolved in a highly coordinated fashion, thereby producing rich looping interdependencies between lower emotional and higher cognitive circuits. A recent proposal by Chareyron et al. (2011) proposes that brain areas which are structurally or functionally interconnected evolve in tandem promoting increases in the convergence and integration of information between the areas. A good example of such coordinated cortical-sub-cortical growth comes from Barton (2012), who suggests that the increased size of the primate cerebellum should be viewed in the context of a three-way co-evolution between the diencephalon, cerebellum and the neocortex (see also Barton and Harvey 2000). Pessoa (2014) makes a similar observation about the massive size increase in the primate amygdala and its remarkable connections (both afferent and efferent) to a wide variety of cortical and subcortical regions.

The result of this coordinated evolution has been the creation of a rich embedding of neural circuitry in which newer circuits are continually woven into older ones producing novel functional landscapes stretching across the entire brain. As Pessoa writes, this interweaving "creates a web of functional and structural couplings in a way that blurs "old" and

"new"" (Pessoa 2015, p. 49). One way to see this is to note that complex sub-cortical dynamics now continuously influence, and are continuously influenced by, complex cortical dynamics. Such processes (of 'continuous reciprocal causation'—see Clark 1997) bind multiple 'components' into unified dynamic wholes in which the state variables of one system are the parameters of the other, and vice versa.[15] Such complex couplings are characteristic of evolved systems in which neural processing, bodily action, and environmental forces are constantly and complexly combined. In the case at hand, we shall see that sub-cortical systems are themselves constantly informed by bodily changes and our own ongoing actions, thus coupling neural predictions and bodily unfoldings in deep and transformative ways.

3.3 Sampling the coupling

To put flesh on these teleo-functional speculations consider the huge number of subcortical structures that target cortical regions either directly or via the thalamus, including areas such as the basal forebrain (Dunnett et al. 1991), hypothalamus (Pessoa 2014), basal ganglia (Clarke et al. 2008), amygdala (Pessoa 2013), cerebellum (Leiner et al. 1986), and brainstem via dopaminergic and noradrenergic systems (Parvizi and Damasio 2000; Mather et al. 2015; Markovic et al. 2015). Many of these ascending systems form important reciprocal loops with the cortex. For example the basal ganglia is connected to the cortex by at least five separate circuits, some of which form closed loops with cortex via the thalamus. This allows information flowing from cortical areas to basal ganglia to return again to the same cortical area (Parvizi 2009). As Parvizi writes, the richness of this looping relationship means that "in reality, there is no cortex versus basal ganglia divide. One does not exist without the other, and there is only an inter- linked network of corticostriatal loops" (op. cit. p.356). From this perspective, online cognitive function cannot be assigned to either the cortical or subcortical component, but instead emerges from their tight coordination.

Work on the hypothalamus provides further examples of the rich suite of interactions weaving cortex and sub-cortex. In the past, researchers primarily focused on the hypothalamus' descending systems (connecting to brainstem and spinal cord), lead- ing to its characterization as a homeostatic controller (Bard 1928; Cannon 1929). The hypothalamus also receives information from the body allowing it to finely tune affective responses to the environment (for a good discussion see Lewis and Todd 2005). Recently however our understanding of hypothalamic

[15] See Clark (2014, chapter 7).

connectivity has expanded to include a rich set of bidirectional pathways connecting to the whole cortical mantle (Risold et al. 1997). Contrary once again to the corticocentric vision, the hypothalamus appears to exert a large influence on cortical function both directly and indirectly via the thalamus, basal forebrain, amygdala and brainstem (all of which are them- selves bidirectionally connected to the cortex and each other Pessoa 2014). This makes the hypothalamus the second largest contributor to cortical inputs after the thalamus (Swanson 2000). The reciprocal connectivity to both cortical areas and the internal body would allow the hypothalamus to synchronize huge flows of information within the brain and body (Pessoa 2013, pp. 230–231). As Kiverstein and Miller have recently noted, "this coordination [facilitated by the hypothalamus] allows for cognitive and affective processes to be mobilized together allowing the animal to behave flexibly, and in ways that are adapted to the particularities of a context of activity" (Kiverstein and Miller 2015).

Finally, consider the profound reciprocal relationship that exists between prefrontal cortex and brainstem nuclei. The brainstem (and striatal) circuits are believed to play a central role in generating rapid emotional responses (the so called 'action tendencies' in Frijda's work; 1986) and so have been called the "seat of emotions" (Panksepp 1998). Meanwhile the prefrontal cortex uses highly processed information from a variety of cortical areas to directly inhibit brainstem reactions thereby allowing time for more sophisticated, and context appropriate, behaviors to emerge. But once again, this is not a one-way relationship. Instead, systems in the brainstem also modulate the frontal lobes by way of neuromodulatory chemicals. Neuromodulatory systems producing dopamine, norepinephrine and acetylcholine within the brainstem, basal forebrain, hypothalamus have terminals in a huge portion of both the sub-cortex and prefrontal areas (Arnsten and Li 2005; Joels et al. 2006). Such neuromodulatory systems are believed to attune cortical processing to signals from the body and environment that are important for survival (Lewis and Todd 2005). As they have recently written:

> "if not for the bottom-up flow, the brain would have no energy and no direction for its activities. If not for the top-down flow, recently evolved mechanisms for perception, action and integration would have no control over bodily states and behavior. It is the reciprocity of these upward and downward flows that links sophisticated cognitive processes with basic motivational mechanisms" (op. cit. p. 20).

During complex behaviors, elements of emotion and cognition are thus so intermixed that a significant decomposition becomes impossible at the level of the brain (Pessoa 2013). If a meaningful decomposition of emotion and cognition is indeed impossible, then processes

considered to be emotional will be poised to play a much richer role than previously proposed (eg. influencing vision). Just such an account has recently been proposed, within the predictive processing framework, by Barrett and Bar (2009)—see also Barrett and Simmons (2015) and Chanes and Barrett (2016).

With massive ascending and descending circuits the PFC becomes an important center of vertical integration of information (Pessoa 2015). PFC along with related areas such as the orbitofrontal, cingulate and insula cortex are all richly interconnected with one another and with amygdala and hypothalamus both of which have huge influence over internal (affective) processes. This collection of areas has also recently been highlighted as part of the so called 'rich club': a community of highly interconnected neural hubs that serve as the backbone for brain wide (cortical and subcortical) synchronizations (van den Heuvel and Sporns 2011, 2013a, b; see also Chanes and Barrett 2016 for a good discussion on the relationship between rich clubs and predictive processing). This tight vertical integration of neural processing suggests that cognitive and emotional processes are strongly interdependent (Lewis 2005; Stapleton 2013; Pessoa 2013; Colombetti 2013). In contrast to the corticocentric vision, cognition and behaviour are better seen as emerging from numerous systems stretching the entire neural axis and dynamically interacting via feed-forward and feed-backward loops (see Lewis 2005 for a richly detailed account of this ebb and flow).

3.4 A new look at the thalamus

This picture of dense cortical-sub-cortical coupling is further enriched by new under- standings of the thalamus itself. While traditionally described as a byway through which information was shuttled into the cortex, today the thalamus is now being re- described as an important neural nexus point capable of orchestrating huge flows of cortical communication. As Pessoa writes, "corticothalamocortical information transfer may represent an important addition to, or even replacement of, the current dogma that corticocortical transfer of primary information exclusively involves direct corticocortical pathways" (Theyel et al. 2010).

In just this vein, Sherman and Guillery (2013) argue that large amounts of thalamic connectivity are not simple way-stations or 'first order' relays, conveying information to the cortex from some sub-cortical source such as the retina. Instead, most of the thalamus is said to be composed of 'higher-order relays': circuits that transmit information between cortical areas—specifically, from layer 5 of one cortical area to some other cortical area. This intriguing feature of the

connectivity matrix is directly suggested by impressive bodies of physiological and anatomical evidence, reviewed and summarized by Sherman and Guillery. It suggests that the primary role of much thalamic connectivity may be to mediate intra-cortical communication. If so, the question naturally arises, what are the differences in the kinds of information being carried by these various pathways? Here, it is notable that "the extra-thalamic targets of drivers[16] to the thalamus seem to be involved in motor control" (Sherman 2007, p. 420). This opens up an intriguing possibility, which is that the information conveyed in cortico-thalamic- cortical circuits may be profoundly entangled with unfolding motoric commands and activity. This possibility has been defended and explored by Guillery (2003, 2005), and is further refined by Sherman and Guillery (2013). On this speculative account, transthalamic cortico-cortico pathways tend to transmit information about the motor consequences of current processing in that area. This means that: "at every level of sensory processing, perception is inextricably linked to ongoing instructions for action, prior to the action itself." Sherman and Guillery (2011, p. 1073).

Sherman and Guillery go on to suggest that the thalamus may thus play a role in establishing and maintaining transient action-specific dynamic couplings between cortical areas, and in alerting cortical areas to any unexpected motor instructions being computed by other areas (see Sherman and Guillery 2011, p. 1074). The resulting picture is one in which "sensorimotor processing is unified throughout all levels of thalamo-cortical function" (op cit p. 1075). Processes of continuous reciprocal causation coupling cortical and thalamic sources here put higher-level prediction machinery in direct contact with unfolding bodily actions. But more importantly, they do so in ways that thus defy easy decomposition into 'leader' and 'led'. Instead, bodily actions and complex top-down predictions co-evolve, delivering just the kinds of 'circular causality' between perception and action imagined by proponents of 'active inference' (see e.g. Friston et al. 2012).

This emerging vision of the densely woven cortico-sub-cortical economy is essential, we will now argue, if we are to flesh out key aspects of the predictive processing story described earlier. In particular, it will help us understand both the power, and the implementation, of a key component of that story—the variable precision-weighting of prediction error.

---

[16] A driver is traditionally distinguished from a modulator. Drivers, as the name suggests, are seen as primary transmitters of information whereas modulators alter the impact of that information. Driver inputs to a thalamic relay are thus diagnostic of the function of that relay, whereas modulator inputs are not—see Sherman and Guillery (2011). Within PP, precision-weighting acts a kind of universal modulator.

3.5 Precision and the threat of magic modulation?

PP accounts are distinguished, in part, by their pervasive reliance upon 'precision estimations' to perform a variety of key tasks and functions. Precision estimates, as mentioned earlier, track the inverse variance of the prediction error signal. In other words, they set error bars around different aspects of that signal according to their estimated reliability, given the task and context. Precise prediction error signals result in increased post-synaptic gain, driving further processing more strongly than their less precise cousins.

There are two main (but deeply related) roles played by precision within the PP architecture. First, and most generally, variable precision weighting alters the balance between top-down prediction and the incoming sensory signal, allowing us to rely on specific chunks of sensory evidence to a greater or lesser degree depending upon task and context. For example, on a clear but windy day, for many tasks, visual information remains highly accurate and reliable and should be given more weight than (say) smell. By varying the impact of specific aspects of prediction error upon further processing, PP systems achieve a remarkable degree of flexibility in their use of long-term stored knowledge about the world. In the case of the McGurk effect (see McGurk and MacDonald 1976) for example, we allow visual information from a mismatched, overdubbed video of a speaking face to overwhelm some aspects of the auditory signal, resulting in our mishearing 'ba' as 'da'. This makes ecological sense since lip movements are normally an excellent cue to speech sounds, and we must often rely upon them in situations of noise or uncertainty.

Second, precision determines the nature of control. For example, Pezzulo et al. (2015) leverage the precision estimation mechanism as a means of 'flipping' between habitual and more reflective means of control. Habitual control emerges when sensory prediction error is rapidly resolved at lower levels of the processing hierarchy. More reflective means of control emerge when prediction error is resolved at higher levels— levels that contextualize lower-level responses. In this way: "the ontology of behavioural paradigms in associative learning can be seen as a successive contextualisation of more elemental sensorimotor constructs, within generative models of increasing hierarchical depth" Pezzulo et al. (2015, p. 18)

Higher-level representations here entrain systemic response over longer time-scales, by predicting—and hence helping to bring about—more complex sequences of responses and environmental (or bodily) states. Influential work on 'multiple controllers' for habitual (model-

free) versus goal-directed (model-based) response is here accommodated within a single precision-modulated inferential schema in which: "it is the precision or reliability of alternative controllers that arbitrates their relative contribution" (op cit p. 19)

High precision predictions exert greater force, and when those predictions originate from much higher levels, they entrain prospective forms of control—forms of control that anticipate and help bring about extended sequences of inputs so as to implement choices and policies concerning future actions. This is the domain that is sometimes referred to as 'counterfactual prediction'—prediction that is truly oriented towards the future, and concerns states of affairs that do not yet obtain. Control is thus:

> "not dichotomized into two discrete systems [habitual and model-based], but viewed as distributed along a graded continuum going from the highest levels of abstract, prospective and conscious reasoning to more concrete, short-sighted unconscious levels of reasoning down to the arc reflex". (op cit p.24)

This picture can be enriched in various ways, for example by noting that habitual control (here, the resolution of high-precision prediction errors using only lower-levels of the processing regime) may sometimes itself extend over larger time-scales, as in the case of highly skilled, over-learnt sports performances. But for present purposes, what matters is simply the pervasive role of precision estimation in supporting flexible, context-sensitive responses that seamlessly negotiate a smooth continuum between more-or-less 'automatic' and goal-directed modes. Behaviour, if these accounts are on track, is contextualized by different hierarchical levels in ways that are arbitrated (op cit p.27) by precision dynamics. Precision here performs two distinct yet inter- related functions. It determines which areas and hierarchical levels currently exert most control. And it "reports opportunities to achieve a goal" (op cit p.28) by reflecting current confidence in those opportunities, and also by responding to signs of progress or failure.

In order to perform these functions adequately, variations in precision-estimation must be delicately responsive to an interacting medley of exteroceptive, interoceptive, and proprioceptive sensory signals. For what goals we pursue, what actions we perform, which aspects of behavior can safely be left to habitual control, and which demand higher-level contextualization, are all matters that require the simultaneous satisfaction of multiple kinds of constraint. Are we in physical danger? Are we hungry, or likely to become hungry if we do not take precautionary action? Is there a sudden opportunity to fulfill a long-standing goal? Is our body currently able to reach some desired target? Do we have enough information to make a good call on whether or not to pursue a certain goal, or should we instead act so as to harvest additional information?

Delicate waves of precision-engineered influence must reflect our brain's best task-and-context reflecting guesses about all these matters, modulating the impact of every aspect of the prediction error signal so as to soft-assemble neuronal resources into temporary webs that keep us viable and that enable us to achieve our goals.

Such spectacular fluidity might raise suspicions. It can sometimes seem as if precision-variation is playing the role of a 'magic modulator', putatively altering the balances of internal power so as to allow the PP framework to accommodate every conceivable form of adaptive behavior, from simple reflexes, to the most complex goal- driven unfoldings. Have we inadvertently imported an un-explanatory free variable into our explanatory schema? We believe that the threat of magic modulation can be averted once we better appreciate the role of sub-cortical processing in the estimation, orchestration and distribution of precision.

3.6 Sub-cortical contributions to precision estimation

Implemented by multiple means in the brain (such as neurotransmitter-based modulation, and temporal synchronies between neuronal populations) flexible precision- weighting renders these architectures spectacularly fluid and context-responsive. Sub-cortical contributions affords further opportunities to incorporate constantly updated information about the state of the body and its readiness for action, and about the uncertainties associated with the bodily information itself.

For example, Kanai et al. (2015) suggest that thalamic nuclei weight precision in the visual cortex. More specifically, their account focuses on the pulvinar. The pulvinar complex is the largest mass in primate thalamus and connects to a wide variety of cortical and subcortical areas via thalamocortical loops. It has extensive bidirectional connections with visual, temporal, parietal, cingulate, frontal and insular cortices, as well as the amygdala. As Pessoa writes, "at a gross level, it is as if the entire convoluted cortex were 'shrink-wrapped' around the pulvinar" (2014, p.11). This massive connectivity provides the pulvinar with ample opportunity to modulate the flow of information across much of the brain.

Kanai et al.'s proposal is that a key task of the pulvinar is to alter the influence (gain) of specific cortically-computed prediction errors so as to reflect their estimated precision. Such a sub-cortical contribution would be a prime instance of deep (cortico-sub-cortical) neural embedding. Core computations performed by the cortex would here be sensitively and constantly modulated

by information about the state of the body and unfolding actions, as registered by the sub-cortical nexus. These gain alterations would, in turn, impact the flow of moment-by-moment cortico-cortico communication, routing and re-routing flows of information and control as task and context unfold. The pulvinar, Kanai et al suggest, is both architecturally well-suited and anatomically well-situated to perform this role.

This has quite radical implications. Because many subcortical circuits are tightly coordinated with internal bodily processes (vascular, visceral, endocrine, autonomic) information from the body turns out to play a much more important role than that assumed by the corticocentric vision of the brain. A particularly important function that requires this integration is the evaluation of (and motoric response to) sensory information. This is the bodily-sub-cortical-cortical weave that "addresses the question: given the present sensory information and the organism's present internal state, how should it act?" (Pessoa 2008 p.150).

Consider vision. What has been called the standard account of vision assumes a division of labor between a 'high-road' ascending from the retina through the visual cortex in a (mostly) hierarchical manner, and a 'low road' that fast-tracks affectively salient visual information from retina through the superior colliculus and pulvinar to the amygdala thereby helping to produce quick survival behaviors. The subcortical pathway is typically referred to in order to help explain the fast reaction time emotionally salient information produces in the brain and body (Pessoa 2013). Many such dual-systems models exist, proposing two competing (or sometimes cooperating) routes: a faster, automatic, emotional, subcortical route, and a slower, controlled, cognitive, cortical route (e.g. Kahneman 2003; Stanovich and West 2000). These models remain both 'cognitivist' and corticocentric insofar as they assume emotional process- ing takes place entirely sub-cortically, and often in a way that is completely insulated from so 'higher' processes such as awareness and attention (believed to be processed cortically). They are also myopic in so far as they are blind to the wealth of subcortical contributions to perception. To help map this more complex process Pessoa describes in detail six circuits (not meant to be exhaustive) that work to fold value into cognition and behaviour by biasing cortical processing towards patterns that are important for the organism's survival. They include the amygdala, cortical valuation regions such as the OFC and insula, fronto-parietal attentional networks, basal forebrain and reticular nucleus, and the pulvinar.

Of particular importance for present purposes is the novel role Pessoa and col- leagues propose for the pulvinar nucleus. Citing a wealth of experimental evidence and neuro-anatomical data, Pessoa suggests abandoning the standard view of pulvinar as a simple subcortical by-way by

which affectively salient signals are fast tracked from retina to amygdala (Pessoa 2013). He offers in its place a detailed account of how the pulvinar biases visual processing towards signals that have emotional or motivational significance in part by way of its rich looping relationship with multiple cortical and sub-cortical areas.

Importantly, the pulvinar is connected with the entire cortical mantle (Shipp 2003). Kanai and colleagues focus primarily on the inferior portion of the pulvinar which is connected with striate and extrastriate cortex (targeting all 20–30 visual areas). However the lateral and medial pulvinar are richly connected to many other cortical and subcortical regions. The lateral pulvinar connects to temporal and parietal lobes (as well as portions of extrastriate cortex), and the medial pulvinar connects to the parietal, frontal, orbital, cingulate and insular cortex and amygdala. Many of these areas in turn have rich bidirectional relationships with visual cortex as we saw above (including the OFC and amygdala). The medial pulvinar here modulates the flow of multimodal information between a huge collection of cortical and sub-cortical areas including OFC, AIC, ACC, and amygdala believed to be of central importance in determining the affective value of signals and preparing the organism to respond appropriately (Pessoa and Adolphs 2010). In this way the medial pulvinar is proposed to amplify weak or fleeting signals that have biological value thereby producing greater behavioural responses (Pessoa 2014, p.72). This optimal positioning and rich interconnectivity allows the pulvinar to fold value into the action-perception system in ways that respect these affective valuations.

This may be a good moment to respond briefly to an important pair of issues raised by an anonymous referee. The first is: why do we take predictive processing (rather than some other theory) to be good candidate for a theoretical account of sub- cortical-cortical connectivity? The second—closely related—is whether the complex dynamical story we favor, replete as it is with complex looping influence and couplings, is actually consistent with the fundamental tenets of predictive processing anyway. Both these issues resolve once it is appreciated that our fundamental claim is that sub-cortical processing plays a major role in delivering the evolving flow of precision estimation essential to fluid and task-optimized cortical processing. Such precision estimations lie at the very heart of the predictive processing machine, sculpting the moment-by-moment flow of information in the brain. Their role, recall, is to enable contextual information to reconfigure the impact of any area or level upon any other area or level according to the changing details of task and of inner and outer context. Our suggestion, in line with Kanai et al. (2015) is that these precision estimations are, to a surprising extent, sub-cortically mediated. This is what keeps them in touch (as they need to be) with both the ongoing flow of embodied action and the changing physiological state of the organism. It was not our aim,

however, to present evidence that predictive processing is the best story about cortico-sub-cortical connectivity. Rather, we assume (for the purposes of the paper) that the predictive processing story is worth pursuing in general, and ask how sub-cortical influence might fit into the story. The fit, we argued, is excellent—such connectivity is in fact ideally placed to carry out the important task of embodied-action-based precision modulation. The upshot is that we should expect to find subtle looping dynamics whereby precision-weighted pre- diction error both reflects and enables ongoing action—the kinds of circular dynamics rightly foregrounded in traditional dynamical systems approaches. To fully appreciate the potential significance of this complex interaction, we next locate affect where it belongs—as a reflection of changing states of organismic readiness for action.

3.7 Affect and action-readiness

Cisek (2007) was the flagship treatment of the so-called 'affordance competition' hypothesis, according to which:

> "the brain processes sensory information to specify, in parallel, several potential actions that are currently available. These potential actions compete against each other for further processing, while information is collected to bias this competition until a single response is selected" Cisek (2007, p. 1585).

The brain, so the story goes, is constantly computing, or starting to compute, a large set of possible actions. These possible actions (which are essentially states of partial 'action readiness') are computed constantly and in parallel. They are also, as Cisek and Kalaska (2010, p. 279) put it, 'pragmatic' insofar as "they are adapted to produce good control as opposed to producing accurate descriptions of the sensory environment or a motor plan". All this makes good ecological sense, allowing time-pressed animals to partially 'pre-compute' multiple possible actions, any one of which can then be selected, completed, and deployed at short notice and with minimal further processing.

In line with such a view, Hoshi and Tanji (2007) found activity in monkey premotor cortex correlated with the potential movements of either hand in a bimanual reaching response task in which the monkey had to wait upon a cue signaling which hand to use. Similar results have been obtained for the preparation of visual saccades (Powell and Goldberg 2000) and using behavioral and lesion studies of reaching behavior in human subjects (Humphreys and Riddoch 2000; Castiello 1999). Decision-making seems to be folded right into these densely interanimated loops so that, to a certain degree at least, "decisions about actions emerge within the same populations

of cells that define the physical properties of those actions and guide their execution" (Cisek and Kalaska, 2011, p. 282).

Emotion belongs at the very heart of this embodied nexus. As processing proceeds, affect and content must be co-computed: intertwined (Barrett and Bar 2009) within the process of settling upon a coherent, action-guiding interpretation of the scene. Sub-cortical mechanisms that assign precisions to cortically computed quantities seem ideally suited to the implementation of such affectively-informed affordance competition in the brain. The sub-cortical loops here keep 'higher-level' prediction systems constantly in touch with our evolving actions, Consistent with such a picture, Frijda (1986, 2007) proposes that affect itself reports on embodied action-readiness programs that simultaneously indicate the relationship between the organism and the environment, and motivate the organism to improve that relationship. As Frijda writes, "emotion, by its very nature, is change in action readiness to maintain or change one's relationship to an object or event" (Frijda 2004, p. 158). Frijda's account thus slots neatly into place with the work of Barrett and Bar (2009), Lewis and Todd (2005) and Pessoa (2015) discussed earlier. The common theme is that affect reports on action- readiness, revealing cognition, emotion, and action as inextricably entwined.

Our speculative story—or better, story sketch—is now complete. The broad connectivity of the medial and lateral pulvinar allows it to integrate various streams of information including affect (insula), action (cingulate), value (OFC) and cognition/attention (frontoparietal). Such thalamocortical loops work to amplify weak or fleeting signals that have biological value thereby producing greater behavioral responses (Pessoa 2014, p.72). Sub-cortically orchestrated precision weighting thus emerges as a potent (and notably non-magical) tool for modulating bodily response, affect, and action. By fully weaving in sub-cortical contributions, we arrive at a vision of a brain permeated by affect, constantly preparing the organism for action.[17]

3.8 Conclusions: coupling the active body and the predictive brain

If cortico-sub-cortical weave plays the roles we are suggesting, the consequences for our understanding of prediction, perception, and action are profound. On the one hand, attention to this delicate multi-dimensional weave should help allay a major worry about the PP approach—the worry that too many puzzles and problems are being solved by the blanket appeal to context-

[17] This is a contemporary version of the profoundly 'motocentric' vision of the brain suggested in the classic work by Churchland et al. (1994).

variable precision assignments. For that blanket appeal, it may now be hoped, may be cashed out in many different ways, that make the most of these (relatively newly-discovered) properties of multiple interacting cortico-sub-cortical loops. In particular, we argued that reflection of the role of the medial pulvinar highlights the role thalamocortical loops play in directing various flows of information towards what is affectively salient.

The resulting picture is profoundly 'embodied' insofar as mutual couplings (with the full dynamical signature of continuous reciprocal causation) obtain between body, sub-cortex, and cortex, with sub-cortical (specifically thalamic) nuclei serving to bring bodily information constantly into the mix. These sub-cortical loops help influence precision estimations in ways that reflect bodily states and unfolding actions, allowing value (to the organism) and affect (relating to interocepted bodily states) to exert a continuous influence on high-level predictions, that themselves exert a continuous influence on bodily states and unfolding actions.

What begins to emerge is a richer vision of the predictive brain itself. Our neural prediction engines are fundamentally in the business of preparing the organism for action, courtesy of constant sub-cortically mediated two-way communication with bodily unfoldings. In this way we construct an affect-laden experiential world that is as much about our own changing needs as it is about the state of a mind-independent reality.[18] Mind like these are thoroughly permeated by emotion and by readiness for action. Emotion, cognition, and preparation for action here form a single whole, self- organized around precision weighted, prediction-error minimizing interactions that span cortical and sub-cortical circuits. These interactions couple the active body to the predictive brain.

[18] Such a perceptual realm is constructed in a fashion that is deeply 'narcissistic' in exactly the sense of Akins (2006).

Chapter 4
The feeling of grip: novelty, error dynamics, and the predictive brain

Abstract

According to the free energy principle biological agents resist a tendency to disorder in their interactions with a dynamically changing environment by keeping themselves in sensory and physiological states that are expected given their embodiment and the niche they inhabit (Friston 2010). Why would a biological agent that aims at minimising uncertainty in its encounters with the world ever be motivated to seek out novelty? Novelty for such an agent would arrive in the form of sensory and physiological states that are unexpected. Such an agent ought therefore to avoid novel and surprising interactions with the world one might think. Yet humans and many other animals find play and other forms of novelty-seeking and exploration hugely rewarding. How can this be understood in frameworks for studying the mind that emphasise prediction error minimisation? This problem has been taken up in recent research concerned with epistemic action—actions an agent engages in to reduce uncertainty. However that work leaves two questions unanswered, which it is the aim of our paper to address. First, no account has been given yet of why it should feel good to the agent to engage the world play- fully and with curiosity. Second an appeal is made to precision-estimation to explain epistemic action, yet it remains unclear how precision-weighting works in action more generally, or active inference. We argue that an answer to both questions may lie in the bodily states of an agent that track the rate at which free energy is being reduced. The recent literature on the predictive brain has connected the valence of emotional experiences to the rate of change in the reduction of prediction error (Joffily and Coricelli 2013; Van de Cruys 2017). In this literature valenced emotional experiences are hypothesised to be identical with changes in the rate at which prediction error is reduced. Experiences are negatively valenced when overall prediction error increases and are positively valenced when the sum of prediction errors decrease. We offer an ecological-enactive interpretation of the concept of valence and its connection to rate of change of prediction error. We show how rate of change should be understood in terms of embodied states of affordance-related action readiness. We then go on to apply this ecological-enactive account of

error dynamics to provide an answer to the first question we have raised: It may explain why it should feel good to an agent to be curious and playful. Our ecological-enactive account also allows us to show how error dynamics may provide an answer to the second question we have raised regarding how precision-weighting works in active inference. An agent that is sensitive to rates of error reduction can tune precision on the fly. We show how this ability to tune precision on the go can allow agents to develop skills for adapting better and better to the unexpected, and search out opportunities for resolving uncertainty and progressing in its learning.

4.1 Introduction

Cognitive neuroscience is on the brink of formulating an elegant unifying theory that shows how the principles that define living systems, also explain the workings of the human mind. The foundations of this theory come from a mathematically complex principle—the so-called "free energy principle" (FEP), which can be applied to every biological system that resists a tendency to disorder (Friston 2009, 2010, 2013; Kirchhoff and Froese 2017).[19] Friston has proposed that everything that can change in the brain will change so as to maintain the adaptive fit of the agent to its dynamically changing environment. The brain (as an integrated part of larger agent-environment system) should steer the agent's interactions with the world so as to maximize the probability that it stays in the physiological and sensory states that are expected given its embodiment and the niche it inhabits. For example, the human body has a high probability of having a temperature of around 37°C. Homeostatic processes in the brain should then regulate body temperature so that the thermal states of the body stay as close as possible to this expected value. In other words, the brain should be organized in such a way as to suppress "surprise", which will remain low when the organism maintains itself in physiological and sensory states that are expected, and will increase should the organism find itself in states that are improbable and hence unexpected. "Surprise" is a technical term and relates to predictions of sensory and physiological states over time—it is future oriented. More precisely, surprise is associated with trajectories or sequences of sensory input; thereby lending it a dynamic and anticipatory aspect. In this treatment, we will be concerned with the surprise of extended sensory outcomes,

[19] Elsewhere we have provided a detailed philosophical overview and analysis of the free-energy principle (FEP), showing how the FEP should be interpreted in ecological-enactive terms (Bruineberg and Rietveld 2014; Bruineberg et al. 2016). Here we restrict our focus to dealing with a problem that seems to arise for FEP when it comes to accounting for behaviours that are motivated by curiosity, exploration and playfulness. We provide only as much theoretical background in this paper as is needed for generating the problem. For readers interested in learning more about FEP we recommend they consult the papers by Friston we cite here and the recent special issue on predictive brains in this journal.

consequent upon the pursuit of action policies. The brain contributes to ensuring that the agent avoids surprise by anticipating how the agent's sensory and physiological states will change over time as it moves through its environment. So long as the brain succeeds in minimizing the divergence between the change in sensory states that it anticipates and the changes in sensory states that actually ensue, it will succeed in keeping the agent away from surprising outcomes, and maintain the agent's adaptive fit to its environment.

Friston provides a precise mathematical framework for quantifying the value of this divergence between the change in sensory states the brain anticipates and the change that actually occurs, using the information-theoretic concept of free-energy. Friston claims that self-organising adaptive systems will avoid surprising states by having a functional organization that continuously minimizes free-energy over the long run.[20] Free energy is related to entropy, it is a measure of the biological agent's order. In thermodynamics and statistical mechanics it refers to the amount of energy that can be extracted from a system and put to work (McEvoy 2002; Clark 2013: p. 186), which is roughly the "difference between the energy and the entropy of the system" (Friston and Stephan 2007: p. 419). The concept of free energy at work in the free energy principle is variational free energy, a "measure of statistical probability distributions" (Friston and Stephan 2007: p. 420). More precisely, what free energy measures is the divergence between a probability distribution typically interpreted as encoding "prior beliefs" about the hidden statistical structure in data, and current sensory evidence. This divergence provides a means of quantifying the information that is available for use in the current sensory evidence. The lower the free energy, the better the system's "beliefs". This is to say that the system's cognitive resources are being put to work in ways that are maximally useful for adapting it to the environment. Free energy increases when the biological agent finds itself in states (potentially life-threatening) that are unexpected relative to its beliefs about the world. The more free energy, which is to say the more often the biological agent finds itself in unexpected sensory and physiological states, the less useful work the biological agent's "beliefs" about the world do.[21]

---

[20] When these dynamics are instantiated in an embodied organism the natural result of this ongoing reduction of error at all levels is the generation and maintenance of homeostasis (Seth 2015; Bruineberg et al. 2016).

[21] Free energy is for this reason sometimes defined in terms of accuracy minus complexity (Friston 2010). Accuracy is a function of how much prediction error the model produces over time—how well does the model do at reducing surprisal over the long-term. Complexity is measured technically in terms of Kubler–Leibler (K–L) divergence. It refers to the divergence between the prior probability of the hypothesis and the hypothesis selected based on the evidence. Complexity is high when many changes were made to the priors to fit the evidence (i.e. the divergence is large). Complexity leads to models that may fit current evidence very well but end up doing worse over time, generalising poorly to new situations (they are overfitted) (Hohwy 2015: p. 5).

Wherever there is free energy there is room for improvement—there is sensory prediction error that is not currently accommodated by one's model. This prediction error can then be accommodated through action, or by improving one's existing model. In this paper we will be concerned with the prediction of the temporally-extended sensory consequences of action, or expected free energy. We say more about the latter concept below. The FEP therefore claims that biological systems are organised in such a way as to minimise free energy continuously over time, which is equivalent to minimising uncertainty in an agent's active engagement with its environment.

If free-energy minimisation is a fundamental organising principle of the brain why is it then that our brains don't steer us towards environments in which sensory states can be easily predicted such as empty dark rooms? In the next section we outline the dark room problem and the solution that has been proposed. We agree with others that the dark room problem is in some ways a red herring (e.g. Friston et al. 2012b), however we will argue that a significant problem remains. The real problem is that of explaining why a biological system that acted based on the imperative to resist a tendency to disorder would be curious, motivated to explore its environment and seek out novelty. The free energy principle would seem to imply that valuable states are the ones the agent expects to be in. Yet curiosity and playfulness will more often than not lead an agent into states that are unexpected. Thus it looks at first glance as if a free energy minimising (FEM) agent ought not to be a curious and playful agent.

In Sect. 1 we outline this challenge in a little more detail. Section 2 shows how the problem has been addressed in recent work on "epistemic action". Epistemic actions are actions an agent engages in to reduce uncertainty. They allow the agent to "disclose information" through exploration "that enables pragmatic actions in the long run" (Friston et al. 2015: p. 2).[22] The research on epistemic action leaves two questions unanswered. First, it fails to explain why it feels good to the agent to engage the world playfully and with curiosity. Pleasure is a part of the value of curiosity and play for agents like us. Existing accounts of epistemic action do not fully explain how value works in epistemic action. Second, an appeal is made to precision-estimation to explain epistemic action, yet it remains unclear how precision-weighting works in active inference. We argue that the answer to both questions may be found in the bodily states of an agent that track the rate at which free energy is being reduced. In Sect. 3 we take up the first of these questions and show how the agent can be sensitive to rates of free energy minimisation

[22] Pragmatic actions are actions that minimise free energy directly leading the agent to occupy the states they expect to be in based on past learning. We discuss Friston and colleagues' distinction between pragmatic and epistemic actions in more detail in the next section.

(FEM). This information about rate of change is given corporeally as states of affordance-related action readiness that are simultaneously affective and behavioural (Bruineberg & Rietveld, 2014; Rietveld, Denys & van Westen, 2017). We show how felt states of action readiness can account for the positive and negative hedonic tone that is often a feature of novel experience. In Sect. 4 we turn to the second question about precision-weighting and show how sensitivity to rate of change may play a role in tuning precision on the fly. This can ensure that the agent is steered towards opportunities for reducing uncertainty. We finish up in Sect. 5 by showing how an agent that is sensitive to error dynamics (rate of FEM) will be a curious agent, motivated to explore and play in its environment.

4.2 Worries About dark rooms

In FEP, agents act so as to keep themselves within expected sensory states given their embodiment and the niche they inhabit. The value of a sensory state is a function of how surprising it is. We are using the term "surprise" here in the technical sense introduced above that makes a conceptual connection between "surprise" and sensory states that are highly probable and thus expected given an agent's embodiment and the niche it lives in. Unsurprising states (states highly frequented) are expected and are thus highly valued. Surprising states are not expected (they are improbable), and thus negatively valued. Positively valued states are often associated with reward (sometimes understood in terms of pleasure), while negatively valued states are typically associated with punishment and are consequently aversive. It follows that unsurprising states should be associated with pleasure, according to FEP and surprising states should be aversive (Friston et al. 2012a). A novel outcome of this perspective is that while it might feel as though we seek pleasures and avoid pains and so end up frequenting pleasurable states more often than painful states, according to FEP *highly frequented states are themselves the rewards*. Through a process known as "active inference" the agent acts to keep itself in states that are expected. A consequence of minimising free energy is that some states are occupied more than others. These are the states that are positively valued by an agent (Friston et al. 2014: p. 2). This means that we are not so much drawn to rewards, but are instead rewarded for reducing errors between expected and actual states. Traditional reinforcement learning models describe goal- directed behaviour as a product of the agent working out how best to maximize an expected reward (Schutz et al. 1997; Sutton and Barto 1998). In active inference the rewarding states are the states the agent learns to expect to occupy through a process of approximate Bayesian inference. Subsequent behaviour unfolds as the system attempts to reduce the discrepancies between the current state and the expected reward state (Schwartenbeck et al.

2014).

This view of decision-making as the outcome of active inference comes with a puzzle. If highly frequented (more expected) states are themselves rewarding and less frequented states (more uncertainty) are aversive, why then should agents ever be motivated to seek out novelty? A good strategy for guaranteeing that one remains in sensory states that are expected might seem to be to seek out a simple, static environment such as a dark empty room in which very little ever changes (Friston et al. 2012b). The agent adopting such a strategy would be pretty much guaranteed to only occupy the sensory states they expected to occupy given their model of the dark room. Nothing unexpected happens in a dark empty room. Thus, once one has learned a good model of such an environment, one is pretty much guaranteed to remain in the states one expects.

The problem of why free energy minimising agents tend not to retreat and hide away in dark rooms has already been well answered by Friston et al. (2012a). On the whole embodied creatures expect to stay warm, well fed and healthy. Dark rooms are not the kinds of environments that allow living agents to meet these basic biological needs. An agent that "felt the pull of the dark room" (Clark 2017a) would be an agent that would after a while experience dehydration and hypoglycaemia, bodily conditions that are highly surprising.

While this response is clearly correct it only takes us so far. What is missing still is an explanation of the adaptive importance of things such as curiosity, play, and the spirit of adventure. Why would an agent that aims to occupy only those sensory states that are expected ever engage in behaviours that lead to novel and surprising discoveries, as happens when we play and explore? More specifically, why would agents ever be motivated to engage in such behaviours? If positively valued states are sensory states that are expected while negatively valued states are surprising, an agent whose actions are the result of active inference should only act to bring about sensory states that are expected. Novel sensory states such as those that occur as a result of exploration should be negatively valued (i.e. highly aversive). Yet this is not the case: many valued experiences are discovered by us through exploration.

4.3 The exploit/explore dilemma

In recent work, Friston and colleagues have shown that in active inference agents don't only act to keep themselves in states that are expected, but also act so as to minimise uncertainty about

future outcomes (Friston et al. 2014, 2015, 2017; Schwartenbeck et al. 2013). They show how curiosity and novelty-seeking, and exploratory behaviour more generally, allow agents to reduce or resolve uncertainty about the world. Consider a scenario in which the agent is uncertain about which outcome to prefer, such as a mouse in a maze that needs to find its way to an unknown location of a reward while avoiding harm along the way. Recall that in active inference, preferences take the form of sensory states the agent expects to occupy or regularly frequent over time. The mouse is uncertain about where the dangers lie, and where the rewards are to be found. Its priors therefore tell it to keep its options open. It should resolve its uncertainty by further exploring the maze.

Friston and colleagues call priors that inform agents about which states they can expect to frequent over the long run "policies". An "action policy" as Friston and colleagues use this term, can be thought of as a rule for the selection of a sequence of actions. We could think of policies in terms of paths of activity some of which are more probable than others to lead you to be in the states you expect to be in. Policies should serve to minimise future free energy, or what Friston and colleagues understand in terms of "expected free energy" (Friston et al. 2015, 2017). Expected free energy is the free energy an agent expects to receive for each of its different policies, were it to pursue them (i.e. the trajectory or sequence of sensory states it expects in the future as a consequence of its actions).[23] The only prior belief about a policy that is consistent with an agent's continued existence is that it will pursue policies that reduce expected free energy. An agent that didn't select policies based on such a prior would be unable to stay well adapted to its niche, and would eventually cease to exist.[24] This implies that policies that have the highest prior probability are rules for generating action that will most likely help the agent to attain what they want in the future.[25] This is because free energy is the divergence between the states an agent predicts it is likely to occupy (its posterior predictions) and the state it believes it should occupy if it is to satisfy its preferences (i.e. the states it expects to be in over time). The higher

---

[23] Technically, the expected free energy also includes an ambiguity term; in other words, the expected free energy or uncertainty comprises the pragmatic value of behaviour in relation to prior preferences plus an epistemic term that resolves ambiguity about the causes of sensations. We thank an anonymous reviewer for drawing our attention to this important detail.

[24] Based on this assumption that the probability of a policy is proportional to expected free energy, Friston and colleagues have built simulations of economic decision making and foraging behaviour (Friston et al. 2014, 2015, 2017).

[25] It should be noted that decision-making is always playing out in a hierarchically organised biological architecture in which there are multiple policies in play operating over multiple time scales. This reflects the individual's simultaneous openness to multiple relevant affordances (Bruineberg and Rietveld 2014).

the prior probability of a policy, the less free energy an agent can expect in the future. When a policy has a high prior probability, the agent can allow the policy to drive action and be confident it will attain what it expects. We see this, for instance in the case of habitual behaviours. These are behaviours we have performed on many occasions that reliably lead to comfortable and familiar outcomes. They are policies that for this reason have a high probability because the free energy we can expect from following them is low.

According to Friston and colleagues, FEP mandates that an agent's choices should reflect their beliefs about which policies have the highest prior probability of causing them to occupy the states they expect in the future. These beliefs inform the agent about the probability of reaching the states it expects from the states it currently occupies. An action policy will be assigned a value based on how well it is predicted to do at minimising the divergence between the states an agent is likely to occupy, and the state it believes it should occupy (i.e. its desired states such as securing a maximum payoff in a game). When the pursuit of an action possibility does not lead to the consequences an agent expects, she will engage in an epistemic action and explore her environment. By contrast, when the pursuit of action possibilities does lead to the consequences an agent expects, and a policy allows an agent to predict a clear path from her current states to the states she wants to occupy, she will treat the policy as highly probable. Policies that the agent believes have a high probability (such as habits) drive an agent's actions since they stand the best chance of minimising free energy.

Friston and colleagues characterise the agent's beliefs about the probability of a given policy in terms of the "precision" of a policy (Schwartenbeck et al. 2014). Precision weighting adds a second-order layer to active inference. In addition to estimating the probability distributions of outcomes the predictive brain must also track the reliability (or precision) of its own estimates given the state of the organism and the current context. It uses this estimation of reliability to flexibly adjust the gain (or the "volume") of particular error units: increasing the impact those units will have on the unfolding process (Friston 2010).[26]

The agent's overall confidence in a policy is reflected in the *precision* of its policies. This precision

[26] Much of the argument in this paper is based upon formulations of active inference in terms of expected free energy using discrete state space (Markov decision) processes (MDP). The concept of prediction error units appeals to a slightly different formalism; namely, predictive coding and Bayesian filtering. However, it is possible to formulate the belief propagation in discrete (MDP) schemes in terms of prediction errors. We can therefore borrow the notion of attention and precision as it is used in predictive coding, when referring to active inference in choosing discrete policies, with no loss of generality. We thank an anonymous reviewer for drawing our attention to this difference.

is updated as a function of the predictive success of the consequences of actions. When the consequences of acting on a policy are correctly predicted (i.e. the potential outcomes are clearly consistent with prior preferences), precision increases and it decreases when new sensory input is surprising. In other words, if things are unfolding as expected, we become increasingly confident in the policy that we are pursuing. The precision of our beliefs about our own behaviour will increase as expected free energy decreases.

Low precision skews the prior over policies to a flat distribution, which reflects the likelihood that policies will be explored with roughly equal probability. Low precision causes the agent to keep its options open, and explore different options (as in the maze example sketched above). High precision by contrast, skews the prior to those policies that have the lowest expected free energy.[27] It is the precision of beliefs about competing policies that Friston and colleagues hypothesise will decide whether an agent continues to follow a well-trodden path or departs from this path to actively explore the world.[28] We can think of epistemic actions as being selected based on a recognition of the current state of the world as offering what one might call epistemic possibilities for action or "epistemic affordances". The world is recognised as offering epistemic affordances when to put it in Friston's terms (1) there is uncertainty to be resolved and (2) there is a clear and precise way forward that is driven by our beliefs about the policies that will best minimise expected free energy.[29] An example might be the uncertainty one experiences about the colour of an item of clothing due to artificial shop lighting. One might resolve this uncertainty by say asking the shop assistant if one can take the item of clothing out of the shop to view it under natural light, or by comparing it with other items whose colour one is more certain about.

While Friston and colleagues have provided an elegant set of formal tools for explaining exploratory behaviour and how agents resolve the so-called "explore- exploit" dilemma, their proposal nevertheless leaves us with two questions unanswered. It is these questions we take up

---

[27] It should be noted that epistemic actions are not always associated with low precision. If there is a clear and obvious uncertainty-reducing sequence of actions available, precision will increase to effectively render that epistemic course of action more likely to be selected. We thank an anonymous reviewer for emphasising this point to us.

[28] We thank Jelle Bruineberg for clarifying discussion of Friston and colleagues' complex work on epistemic value and for helping us fine tune some of our formulations of these ideas in the text.

[29] We thank an anonymous reviewer for this suggested characterisation of epistemic action in terms of recognition of epistemic affordances. For previous discussion of the concept of epistemic affordances, though not in the terms of Markov decision processes we are employing in this paper (see Rietveld & Kiverstein 2014).

in the remainder of our paper.

First, play and curiosity result in surprising and unexpected discoveries which are often experienced as having positive hedonic value. The novel experiences we have as a result of exploring our environment often feel pleasurable. Think of visiting a new culture as an example. The positive feelings of pleasure are part of what motivate us to engage in this type of exploratory activity rather than sticking with the comfortable familiarity of what is already known. Friston and colleagues suggest that this feeling of pleasure should be a consequence of recognising the action opportunities the world offers to reduce or resolve uncertainty (i.e. an epistemic affordance of the agent's situation). Recall how the states an agent values are unsurprising states the agent expects to occupy. These states are the consequences of the behaviours the agent performs in seeking to minimise expected free energy. In other words the states an agent expects to occupy given its priors should be states with positive hedonic value. This however leaves it unexplained why novel experiences that reduce uncertainty should *feel good*. Why should there be a positive phenomenology that comes with exploring and making progress in reducing uncertainty? We enjoy being curious. This is part of the value we assign to these activities. Insofar as current work on epistemic action doesn't account for the positive feelings that characterises curiosity and play, it doesn't yet fully account for how value works in epistemic action.

Second, the success of Friston and colleague's account depends on their explaining how it is that agents are able to optimise the precision of their policies. They say that precision estimation is a consequence of free energy minimisation. They show how the optimisation of precision estimations has dynamical properties that closely resemble those of dopaminergic systems in the brain.[30] This provides a candidate mechanism for the implementation of precision in active-inference. However it leaves us with questions still about how precision is weighted in a given context. We will show how these two questions may turn out to share a common answer. Bodily feelings in the form of affordance-related states of action readiness turn out to be a part of what allows an agent to estimate precision on the fly.

4.4 The feeling of action readiness

[30] It has been suggested that precision weighting is the result of neuromodulators (such a dopamine).Kanai and Friston have suggested that, "this may explain why superficial pyramidal cells have so many synaptic gain control mechanisms such as *N*-methyl-d-aspartate (NMDA) receptors and classical neuromodulatory receptors like D1 dopamine receptors" (Kanai et al. 2015).

We've seen in the previous section how expected free energy gets to decide whether the agent exploits familiar solutions or explores the environment to reduce uncertainty. Precision influences how evidence is accumulated and thus how "beliefs" are formed by an agent. Precision estimates are made more generally by the brain to separate out organism-important error signals from the surrounding unimportant noise (Feldman 2013). For example, while walking home on a familiar busy street one might barely notice the general buzz of the people around you. Day to day the buzz is inevitably different in shape—different people interacting in different ways. Nevertheless, while the exact shape of the buzz is exceptionally unpredictable it draws almost no attention: no further processing is allocated. Now if someone were to unexpectedly fall close by and seem to require assistance this bit of error would suddenly be pertinent. If we think of error signals as broadcasting newsworthy information then we must also explain how the brain decides which channels it should "listen to" (Kanai et al. 2015; c.f. Kwisthout et al. 2017). Estimating the precision of a policy is a special case of a more general phenomenon in which the agent is continuously engaged in monitoring its own level of confidence in its predictions about the world.

The foregoing concerns the precision of prediction errors in relation to states of the external environment. However we have been discussing precision in relation to action-selection, and have therefore been discussing the precision of beliefs about policies. These notions of precision turn out to be closely related and intertwined on our account.[31] Policies consist of interrelated, and nested states of action readiness, which are patterns of readiness for the sensory consequences or outcomes of action. We therefore suggest that in assigning precision the brain isn't *only* concerned with the reliability of the PE signal. It is more accurate to say that precision relates to how well the agent is doing at engaging with expected uncertainty in relation to the sensory consequences of temporally extended sequences of actions. Precision doesn't just concern the agent here and now and its momentary state of uncertainty with regards to some current prediction error. We've seen above how expected or future uncertainty is also important when it comes to assigning precision to policies. FEM agents actively seek a means of managing uncertainty over time. The rollercoaster of continual increases and decreases of errors that accompany life become expected and are folded into our expectations. For such a system it becomes important not only to track the constantly fluctuating instantaneous errors, but also to pay attention to the dynamics of error reduction over longer time scales. This means paying attention to *the rate* at which those errors are being reduced or increasing. Rate of change is an important (but largely overlooked dimension) of free energy minimisation. If we compare two

[31] See our earlier footnote 8.

agents both of which succeed in dealing with prediction error the agent that does it faster will do better in the long run than the agent that takes longer.[32]

We can think of the rate of change of prediction error reduction by analogy with velocity (Joffily and Coricelli 2013: p. 3). The velocity of an object is the rate of change in the position of an object relative to a frame of reference over time. So velocity is equivalent to the speed of an object moving in a particular direction. Rate of change in relation to prediction error reduction thus refers to how fast or slow prediction error is being reduced relative to the states of the whole agent-environment system. If the speed of error reduction increases, this equates to decrease in free energy over time (relative to what was expected). If speed of error reduction decreases, this equates to an increase in free energy over time.

Each agent's performance in reducing error can be plotted as a slope that depicts the speed at which errors are being accommodated relative to their expectations. The steepness of the slope indicates that error is being reduced over a shorter period of time and so faster than the agent expected: the steeper the slope, the *faster* the rate of reduction. Think of mastering a second language and finding it easy to take part in a conversation with a stranger in this second language. A gentle slope by contrast indicates that error is being reduced at a *slower* than expected rate. The agent has encountered an error that is proving difficult to deal with. This has the result that they reduce fewer errors over time. Suppose for instance that a person has broken a leg and they now need to get around their environment on crutches. They are now much slower to move around —the rate at which they are able to get into the states they expected to be in has slowed dramatically. This is typically the source of some frustration for the person.

Once we have the idea of rate of change in play, we see that even large instantaneous errors (sudden spikes in the slope) could be experienced as positive as long as the error takes place within a more general reduction of error over time. This is to say that a large but resolvable error signal informs the agent that the environment offers the opportunity to resolve uncertainty. An epistemic action then becomes an attractive option.

This goes a long way to helping make sense of why certain errors are accept- able and even highly desirable. The environment offers an epistemic affordance, an opportunity for

---

[32] Technically, what we are saying here is that the path or time integral of prediction error (or expected free energy) is the important thing. Crucially, because these time integrals or averages are known as Hamiltonian Action, we are saying is that the best policies will conform to Hamilton's Principle of Least Action. We thank an anonymous reviewer for the formulation of this technical point, and for making the connection to Hamilton's principle.

information gain that allows one to pick up on a free-energy lowering policy. Information about rate of change is thus highly important for updating one's expectations, changing the course of action entirely or rather continuing on the same course of action. It informs the agent whether, given what they know already, there is still room to improve. Perhaps there is no room for improvement because error rate just keeps increasing in which case what you should do is just try something differ- ent. You shouldn't change what you anticipate happening when the result of making such a change will just be more error. You should instead explore and look for new opportunities that help you to learn to grip better.

In the recent literature a number of authors have begun to relate information about rate of change in error reduction to the valence of full-blooded emotional experience like happiness, disappointment, hope and fear (Joffily and Coricelli 2013; Van de Cruys 2017; Van de Cruys and Wagemans 2011). These authors hypothesise that the valence of different emotional experiences is a reflection of *unexpected rate of change* in prediction error reduction. When free energy is increasing at a greater than expected rate this can feel bad, it can for instance be experienced as frustrating. Conversely when free energy is decreasing at a faster rate than expected this can feel really good. For example, a positive emotion like happiness is a reflection of an unexpected reduction of prediction error (e.g. error being reduced at a rate faster than expected), while a negative emotion like disappointment reflects an unexpected decrease in prediction error reduction (e.g. error being reduced at a rate slower than expected).

In this section we've been applying the notion of rate of change to explain why an agent might be moved to explore rather than exploit. Assuming these authors are right to tie valence to rate of change, our proposal is that it is valence that plays this role of motivating an agent to to engage in exploratory actions rather than exploit.[33] The concept of "valence" is used by these authors to refer to the positive or negative (approach or avoid) character of an emotional experience that inform the organism about its current relationship with the environment. Pleasurable states have positive hedonic value and are associated with approach behaviours. Aversive states have negative hedonic value and are in turn associated with avoidance behaviours.

[33] A closely related proposal can be found in Van de Cruys (2017). He emphasises that valence as error dynamics helps guide the organism in its exploration to niches that have maximum gain in prediction-error minimisation. Van de Cruys doesn't explicitly apply this hypothesis to the case of epistemic action as we have been doing in this paper. However, we can suppose that sometimes increasing PE may lead an agent to shift into an exploratory mode of engagement with their environment, while on other occasions it may lead the agent to shift to another reliable exploit mode. Exactly what the agent does next will depend on the precision of the policies it is enacting as discussed above. Differences between our ecological-enactive account of valence and the view that can be found in Van de Cruys are discussed below.

Their claim is not so simple as to say we feel good when our predictions fit and bad when they do not. They claim (rightly) that the predictive organism is in a constant state of error management at all levels of the hierarchy, and yet clearly there is not a felt experience of each and every fluctuation. Errors are reviewed on a background of learned expectations concerning how fast or slow (the rate) such errors have been reduced previously (Joffily and Coricelli 2013).

We suggest thinking of valence differently in terms of multiple states of affordance- related action readiness that are simultaneously affective and behavioural (Rietveld 2008).[34] At the same time as emotional experiences feel good or bad they also prepare or make us ready to act on relevant affordances (possibilities for action offered by the environment). It is the relevant affordances of the environment that have valence. This valence consists in solicitations or invitations to act: relevant affordances attract or repel the agent's actions (Bruineberg and Rietveld 2014). Think of how the apple sitting next to you as you work can look enticing to eat when you are hungry: it has positive valence. But when you take a bite into and find it is rotten inside it ceases to be enticing in the same way and takes on a negative valence. We thus disagree with Joffily and Coricelli (2013) and Van de Cruys (2017) who treat valence as the avoid/ approach character of full-blooded emotional experiences. They treat valence as a property of emotional experience while we suggest understanding valence in relational terms, and as necessarily environment involving: it is relevant affordances that have valence in virtue of which they solicit or invite some form of action on the part of the agent.

In addition to valence at the level of individual relevant affordances we suggest it also makes sense to think in terms of the field of relevant affordances as a whole as having valence. This is the kind of valence that is made explicit when someone asks how things are going, or how one feels in a situation. One's initial response to this question relates to the situation as a whole, though of course it is possible to zoom in on particular aspects to make things more specific. The field of relevant affordances comprises the multiple relevant affordances that get the agent bodily ready to respond. One is ready to respond to each of the relevant affordances but one

---

[34] Here we are drawing upon the analysis of emotion found in the research of the emotion psychologist, Nico Frijda (see (Frijda 1986, 2007), in which states of action readiness are central. We will have nothing to say here about so-called *full-blooded* emotions that Frijda analyses (e.g. happiness, sadness, anger, etc.). Our interest will be instead in the myriad ways that the brain and body change as part of our meaningful engagement with the world. Our use of felt states of action readiness to refer to the individual agent's openness to a field of relevant affordances closely parallels Matthew Ratcliffe's account of "existential feelings" (2008). Existential feelings are background bodily states of action readiness that orient us in the world. They are what attune us towards or away from relevance. While existential feelings are certainly felt, the affective changes can themselves be more or less explicit, foregrounded objects of attention. So for example, we may be expressly aware of our experienced anxiety, or we may simply perceive a situation (through our anxious state) as worrisome.

also has an overall grip on the situation. It is this valence at the global level of the field as a whole that we propose to understand in terms of rate of change. Global level valence gets the agent ready to either exploit the particular inviting relevant affordances or to seek out epistemic affordances that offer opportunities for information gain, thereby allowing one to pick up on a free-energy lowering policy.

Valence can thus be thought of both at the global level of the agent in relation to the field as a whole, and at the more local level of micro-level states of action readiness elicited by relevant affordances that invite the agent to act. The former ("global valence") is best thought of in the context of what we've earlier described as the tendency towards an optimal grip on the field of relevant affordances (Bruineberg and Rietveld 2014). The relation between multiple micro-level states of action-readiness and macroscopic patterns of activity at the level of the individual agent as a whole is analysed in terms of self-organising dynamics in our earlier work (Bruineberg et al. 2016; Rietveld, Denys & van Westen, 2017).

In earlier work (Bruineberg and Rietveld 2014; Bruineberg et al. 2016), we've proposed an ecological-enactive reading of the FEP, providing an analysis of free- energy in terms of disattunement of internal and external dynamics, a dynamic state of disequilibrium within the agent-environment system as a whole. In active inference, agents prepare actions that will reduce disattunement and thereby lead them closer towards some dynamical equilibrium or grip on the situation. States of action readiness originate in fluctuations of affect that orient us towards affordances that matter to us, preparing us for sensory consequences that arise from responding to inviting possibilities for action. Emerging a few hundred milliseconds after an event, action readiness is the earliest coordinated evaluation of the new situation by the organism as a whole (Klaasen et al. 2010). Positive and negative feeling should be thought of as an integral part of tending towards an optimal grip. As Frijda notes:

> "Emotional feeling is to a very large extent awareness, not of the body, but of the body striving, and not merely of the body striving, but the body striving in the world... emotional experience is to a large extent experienced action tendency, or experienced state of action readiness." (Frijda 2004: p. 161, quoted by Lowe and Ziemke 2011: p. 8)

Felt states of action readiness manifest as a "complex space of polarities and combinations" (Colombetti 2005; Thompson 2007: p. 378). The agent prepares to move towards/away, approach/withdraw, and is receptive/defensive to affordances in the environment that are exerting a pull on, or repelling the agent. These movement tendencies can also be consciously felt as pleasant/unpleasant, positive/negative, and they relate to affordances the

individual likes/dislikes, is attracted/repelled by. We suggest that felt states of action readiness arise when there is an unexpected change in rate of error reduction at the level of the agent-environment system as a whole. Rate of change is an important source of information for the agent because it can help them to always be ready for opportunities for improving grip, and living systems continuously strive to improve grip (Bruineberg and Rietveld 2014).

This ecological-enactive analysis of rate of change can help us to address the first of the questions we raised about epistemic action in the previous section. There we raised the question of why reducing uncertainty through epistemic actions should feel good to the agent. Why should novel experiences have a positive phenomenology? We've just suggested thinking of rate of change as the changes felt in the skilled body (in its relation to the field of relevant affordances as a whole) with a positive or negative hedonic value. When an agent succeeds in reducing error at a faster than expected rate (or recognises the opportunity to do so) this feels good. It is thus not uncertainty reduction alone that the agent cares about, but also the rate at which uncertainty is being reduced. Pleasurable feelings arise in the form of feedback as part of the process of our moving towards, or being drawn towards affordances that are relevant to us. We are drawn towards opportunities for improving grip, and positive feelings arise when we improve grip at a faster than expected rate. This is to say that the slope one could plot describing the rate of error reduction would have a steep incline.[35]

Conversely, negative affect is experienced by the agent in terms of being repelled from a situation. This occurs when we do worse (or anticipate doing worse) than expected at reducing error, and the slope describing error reduction has a gentle incline. Consider boredom as an example. Suppose you find yourself stuck in a seat in a music hall sitting through a boring symphony. The music you are hearing has nothing to offer. It is boring because there are no salient regularities to be harvested relative to the skills you have that attune you to the environment. There is no opportunity to do better than expected at reducing prediction error because the skills you have already attune you to the music without this requiring any effort from you. Alternatively the skills of the individual may be such that the structure of the music is too complex to get a grip on it. This is an inherently frustrating situation for the listener stuck in their seat who nevertheless aims to continuously improve their grip on the world. The felt frustration manifests as boredom for the music and agitation, a drive to get up out of your seat and leave.

---

[35] Again we see Hamilton's Principle of Least Action in play. Our thanks to an anonymous reviewer for this point.

In the next section we argue that rate of change may also play a key role in *precision estimation* as it occurs in active inference. Active inference can be tuned by rate of change in agents that are sensitive to this feedback signal. They can use this feedback signal to allow themselves to be pushed towards or pulled away from aspects of the environment that offer opportunities for making progress in uncertainty reduction. The hypothesised role of rate of change in precision estimation is among the important novel contributions of this paper.[36]

4.5 Using rate of change to tune precision-weighting on the fly

How is precision weighted in active inference? We suggest in line with our ecological- enactive reading of FEP that precision should be understood in the context of tending towards an optimal grip on the affordances available in the ecological niche. Precision is what sets the degree of influence on behaviour (or more precisely, action readiness) of the multiple relevant affordances inviting us to act.[37] Error dynamics (rates of FEM) are "grasped corporeally", we feel the dynamics of error-reduction (and whether it is going well or badly) in our attunement to the world (Patocˇka 1998). The agent doesn't simply act so as to improve grip; it is a part of their acting skillfully that they can do so with sensitivity to how well or badly they are doing. Agents that make use of feelings that arise from rates of change can continuously do better at improving their grip on what is relevant in the landscape of affordance by exploring and seeking out novelty. They can aim to better engage with error, and attune to the unexpected so as to broaden their skills and grip in more and more domains of their ecological niche.

Once we understand FEM as always being enacted in an ecological context as we propose, it makes sense to suppose that in general, FEM agents will be on the lookout for opportunities that are rich in the kinds of error they can manage given their skills (eg. manageable errors). We suggest this is the kind of error that will allow an agent to improve on their level of skilled engagement with affordances. Unexpected improvements or setbacks, in relation to the expected rate of error reduction, then provide a particularly valuable learning signal that can direct resources to opportunities for improvement or speak in favour of task-switching (see Rietveld and Brouwers 2016 for real life examples of this). FEM organisms do *not* try to *maximally* reduce

[36] It is a possibility that has not to our knowledge been recognised in previous work that has appealed to error dynamics to explain the valence of emotional experience.

[37] This hypothesis about the precision mechanism is developed in more detail in Miller and Clark (2017).

error, since sometimes error can be invaluable for learning,[38] and in any case prediction error is unavoidable. When we think about a day that has gone well, we evaluate the day as a whole in part based on the surprising and unexpected things that happened to us and how well we managed to deal with them. A good day is not only one in which we succeeded in reducing prediction error relative to our expectations. It is one in which we were met with all manner of unexpected events and we did well at meeting these challenges.

Agents can be sensitive to different degrees to how well or badly they are doing at gripping to their environment. To put this in terms of rate of change of FEM, they are sensitive to different degrees to the increases and decreases in FEM. Given an advanced level of skill we expect our skilled body (as a model of the world) to do well at attuning to a given context. When we run into more troubles than we anticipated, this slows down the rate at which our current skills succeed in attuning us to unexpected changes in the environment. Think about a difficult day at work in which many problems come up that you fail to solve. The feeling of frustration in this case is feedback that informs us that things are going worse than expected. This is important information for an agent because it can be used to move us to do things differently. When free energy is rapidly increasing, this is a sign that the agent is doing poorly over time at accommodating sensory input. An agent that possesses this information about rate of change should downgrade confidence in their policies. Equivalently, they should assign more confidence to prediction errors relative to their policies.

Now consider a more positive scenario in which prediction error progressively decreases like in the example of the day in which lots of unexpected events occur that we nevertheless manage well. The agent's policies are performing well at attuning them to change in the environment, and thus they have good cause for being highly confident in their policies.

The more an agent takes note of rate of change, the more sensitive they are to how well they are gripping in a given situation. Agents that lack sensitivity to felt states of action readiness will be more likely to get stuck in situations that are frustrating to them. Feelings provide them with the impetus or impulse to switch. When people are not sensitive to these feelings, this can lead them to be overconfident or underconfident in their policies. Overconfidence in one's policies can lead one to overlook unexpected changes in the environment that bodily feelings attune us to. Underconfidence in one's policies can make an agent dissatisfied with what they are doing when

[38] Schouppe et al. (2014) shows that more reward comes from resolving incongruent stimuli than from congruent stimuli in a Stroop task. These findings points to initial errors being conducive to higher rewards.

they are doing just fine at adapting to the unexpected. Sensitivity to rate of change and to the feelings it gives rise to can be used to tune confidence as we go.[39] This is maybe one of the essential ways in which agents are able to stay attuned to what matters to them. Failure to tune confidence as we go along leads to inflexibility and failure to switch activities based on changes in the environment or changes in internal state (e.g. homeostatic needs). An agent might for instance persist with some activity that has been weighted as highly important, but is failing to reduce free energy at the expected rate because of some other pressing need that is being neglected. Sensitivity to rate of change, which reflects how things are going at the global level of the agent as a whole, would tell the agent that they need to assign priorities differently allowing them to do better at reducing free energy overall by continually tuning their confidence in the expectations that are driving their actions on the fly.[40]

Why are people so curious and playful? We can remove some of this mystery we suggest, once we appreciate the role of rate of change in precision estimation. Agents that weight precision based on feedback from the feeling of grip will be attracted to opportunities for continually improving in their skills. Positive and negative hedonic value is felt in the body and can (in the agent that is sensitively attuned to these feelings) provide feedback that moves the agents through the environment leading them to places where they stand to learn the most.

4.6 Novelty seeking and learning progress

We've been arguing it is not only prediction error that the agent seeks to reduce in their skilled engagement with the environment, but also the rate of change in prediction error (i.e. the opportunity to reduce uncertainty in the future). Anticipation isn't just about determining what is most likely to happen next. Optimising one's engagement with a dynamically changing environment requires in addition, sensitivity to how well one is doing at reducing disattunement

---

[39] Although beyond the remit of our current treatment it is a prediction of our arguments that the dopamine system should play a central role in this online tuning of precision in active inference (thanks once again to an anonymous reviewer for suggesting this point to us). This is because, in physiological versions of predictive coding and active inference, precision is thought to be mediated by neuromodulatory effects (see e.g. Friston et al. 2012a). Key among these (in the domain of motivated behaviour) is dopamine—a neurotransmitter strongly implicated in hedonics, reward and emotional learning.

[40] Inflexibility that arises from failure to tune confidence on the fly may go some way towards accounting for the current political climate in which people are falling back on comfortable, familiar certainties from the past in response to the many complex, hard to manage uncertainties of the future. The current political situation presents people with a tremendous amount of uncertainty that is far too complex for them to integrate. One natural response to this is to fall back on outdated but familiar and entrenched patterns of interaction that no longer give a grip on current events, which for interesting sociological reasons people nevertheless overlook.

over time. Sometimes it feels good for the agent to generate more prediction error in their interaction with the environment as a part of their epistemic foraging.

Irreducible error means environmental complexity is too high for the agent—and this feels bad, as in the example of the symphony that one finds boring because its complexity is too high. Too little error in our dealings with the environment means that our model is already fitting well to the environment, there is nothing further to be learned and we feel bored. Value in epistemic action thus seems to be a matter of finding the right balance so that the agent is continuously improving in the speed at which they are reducing prediction error.

Imagine signing up for a one hour swimming lesson to improve one's swimming skills. When one arrives for the first time at the swimming pool for the training one finds out that the group is very large and that it is not a beginners group. Most people in the class are already proficient at doing breast-stroke for instance, something one has not yet mastered. Given the size of the class, the teacher probably notices that one's level of performance is lagging behind other members of the class but is unable to give one the attention needed for acquiring a new skill. One feels that things are going a good deal worse than one expected, and one's swimming ability is not at all improving. Every new exercise during the class is too difficult. The valence of this situation is negative; one feels out of place and has the action tendency of leaving the class, and joining a different one that is better suited to one's level of ability. Rather than being determined to keep trying until one has mastered the exercises sufficiently, one might well decide it would be better to switch and find a class that is targeted at beginners and has more personal attention for the participants.

The richest opportunities for improving in FEM will come from situations that are neither too complex, nor so simple and straightforward that we already know how to deal with them. In finding this balance of complexity and simplicity an agent is able to learn optimally and so is able to do better, while nevertheless always falling short of fully attaining an optimal grip or equilibrium with the environment. This means being sensitive to the felt states of action readiness that attune one to positively or negatively valenced relevant affordances. Negatively charged feelings of action readiness tell one that things are not going well. One is failing to grip in the way one expects given one's level of confidence. This may make switching activities or strategies an enticing option so long as one is sensitive to this feeling. Positive feelings tell you that you are doing well at dealing with, and adapting to unexpected changes. This provides you with valuable feedback for further improving your skills.

Various lines of research support the view that agents seek out environments with optimal amounts of novel complexity (error). For example, Berlyne (1966) argues that organisms actively seek out stimuli that are slightly above the complexity the organism is used to. A few good examples of this come from research on early childhood development in which newborns were found to attend longer to stimuli that are neither too simple nor too complex (Kidd et al. 2012). In non-human studies, rats were found to frequent parts of a maze that were decorated in a slightly more complex fashion than parts they commonly frequented in the past (Dember et al. 1957). Learning environments in which the complexity is just above the abilities of the agent offer the largest accelerations of error reduction. These are the environments the rats *like* to explore the most because they offer the most opportunities to learn relative to what they know already. Exploration of these environments feels good to the rats because it is in such environments that acceleration of FEM is at its greatest. We hypothesise that the rats are motivated to explore because it feels good to explore this kind of environment—doing so offers them the greatest opportunity for acceleration of FEM.

A similar line of thinking comes from recent research on error reduction dynamics in artificial intelligence and robotics (Oudeyer et al. 2007, 2013; Schmidhuber 2010). Kaplan and Oudeyer take the rate of error reduction to be associated with intrinsic rewards in humans (2007).[41] By linking error reduction dynamics and intrinsic rewards they offer a model of learning in which agents are intrinsically driven to investigate particular regions of the environment as long as there are learnable regularities left to harvest given their current skill level. Being sensitive to error dynamics guarantees that the agent avoids wasting time in places where regularities are either already learned or too complex given the agent's skill level. To put this in the terms of our paper, sensitivity to felt states of action readiness tunes the agent to, and draws them to explore, learning-rich places simply by tracking local learning progress. Such systems will naturally and spontaneously move from one stage of development to the next, from one level of complexity to another, as error reduction becomes less available (either all that's left is uninteresting noise or the complexity is yet too high to be managed given the skill level of the organism). A neat outcome of such systems is the self-organization of developmental and learning trajectories that naturally move agents from acquiring simple to more complex skills over time (Oudeyer and Kaplan 2006; Oudeyer et al. 2007; Kaplan and Oudeyer 2011; Moulin-Frier and Oudeyer 2012).

These ideas about learning progress have shown promise in developmental robotics, allowing

---

[41] In the free energy framework, the concept of intrinsic motivation that underlies (self) exploration (e.g. motor babbling) and novelty seeking is understood in terms of *expected free energy;* also referred to as Bayesian surprise, information gain in the literature. Intrinsic value is the value of information or epistemic value (see Sect. 2 above).

robots implementing these routines to "efficiently learn repertoires of skills in high dimensions and under strong time constraints and to avoid unfruitful activities that are either well learnt and trivial, or which are random and unlearnable (Pape et al. 2012; Ngo et al. 2012; Baranes and Oudeyer 2013; Nguyen and Oudeyer 2013)" (Gottlieb et al. 2013: p. 9). There are good reasons to suspect a similar approach to learning takes place in humans. Progress-based systems in robotics simulate closely infant sensorimotor development. This has led to the hypothesis that certain patterns of information-seeking behaviour in humans may emerge from a particular embodied system (morphology, etc.) intrinsically motivated by progress-based learning strategies (Smith 2003; Kaplan and Oudeyer 2007; Oudeyer et al. 2007). To put this point in our own words: the sensitivity to the valence of a situation moves agents to explore situations that maximize learning progress by directing them towards a trajectory of action possibilities whose complexity is neither too simple nor too novel. Oudeyer and colleagues refer to these learning sweet-spots as *progress niches*:

> "Progress niches are not intrinsic properties of the environment. They result from a relation between a particular environment, a particular embodiment [...](sensors, actuators, feature detectors, and techniques used by the prediction algorithms), and a particular time in the developmental history of the agent. Once discovered, progress niches progressively disappear as they become more predictable." (Oudeyer et al. 2007, p. 282)

By tracking their own progress in learning, agents are moved to seek out opportunities that present them with just the right level of complexity that they can make something of these opportunities given their current level of ability.[42] FEM agents don't need to orient towards novelty indiscriminately. Agents that make use of their sensitivity to how well they are doing at tending towards an optimal grip can orient to *the right kind of novelty*—the kind that maximises their own learning rate or the opportunity to do better at improving their grip on what matters in the environment. They will be agents that are intrinsically motivated to explore, seek out novelty and along the way improve their skill at gripping.

To offer a rather extreme example of this, some people are able to forecast events like election results, economic collapses, famines and wars with an accuracy much better than chance. They have been shown to perform on average 65% better than the average person and 30% better than US intelligence agents that forecast these kinds of events for a living (Tetlock and Gardner

[42] We don't mean to suggest that progress niches are discovered by individual agents all by themselves. Often other people from the practice, more experienced practitioners, give the most relevant feedback about how well one is performing and educate one's attention to relevant affordances.

2015).[43] These people are highly skilled at estimating the probable outcomes of counterfactual scenarios. It is a prediction of our arguments that the secret to their success must lie in their ability to form better expectations than the average person about how errors will arise. They do a way better job than the average person of anticipating when the unexpected is likely to arise and adapting their probability estimates accordingly. If our arguments are along the right lines, this is something they are able to do by paying close attention to what their feelings tell them to do so they can continuously tune the confidence they have in their own estimations of what could happen in the future. They are always looking for opportunities to incrementally improve in how they are doing at reducing error, and they don't care so much about the end product. They are as Andy Clark nicely put it to us in conversation "slope-chasers".

4.7 Conclusion

We started the paper by raising a puzzle for FEP about why an agent would be motivated to play and explore if all they ever aim to do is keep themselves in states that are expected relative to its model of the world. We've argued that a FEM agent should naturally engage in epistemic foraging, and seek out epistemic affordances or opportunities to reduce uncertainty in relation to their policies. We've raised two questions for an account of epistemic foraging in terms of FEM and active inference. The first question asks why it should feel good to an agent to engage in exploratory, curiosity- driven behaviours. We've offered a new perspective on recent work on the rate of change in error reduction to address this question. According to this earlier work, it feels good for an agent to increase in the speed of error-reduction, and it feels bad for an agent when they reduce error at a slower than expected rate. We've suggested that what rate of change tracks are affordance-related changes in states of action readiness. Our contribution is thus two-fold. First we have proposed an ecological and enactive interpretation of these relatively recent ideas about what rate of change might be doing in predictive-processing. Second we have put this interpretation to work to explain the positive and negative hedonic value that can motivate an agent to engage in epistemic actions.

The second question we've raised asks how precision-weighting might work in active inference. Precision-weighting plays a crucial role in epistemic action since it is the precision that is assigned to a policy that settles the question for the agent as to whether to exploit what one already knows, or to explore and seek out novelty. Building on our ecological-enactive interpretation of FEM we've proposed that sensitivity to rate of error reduction may play a role

[43] Many thanks to Andy Clark for suggesting this fascinating example of super-forecasters.

in precision estimation. Active inference can be tuned on the fly in agents that use rate of change as a feedback signal, thereby guaranteeing that they continuously make progress in reducing their own uncertainty.

An organism wired to be rewarded for reducing disattunement at a faster rate should naturally and spontaneously orient itself to places in the ecological niche that offer opportunities where disattunement can be managed the best. They should be motivated to seek out opportunities for managing errors that are just above their current level of ability and to develop new skills. This is to say that the FEM agent ought to be a curious agent, motivated to explore and play in their environment, constantly pushing the boundaries of what they know how to do. In their exploratory engagement with their ecological niche, curious agents discover novel affordances that allow them to constantly improve in their skills.

Chapter 5
Desire and the predictive organism

Abstract

People tend to feel good when they attain something they desire, and frustrated when they fail to do so. They tend to value positively feelings of pleasure, and value negatively feelings of frustration. While no doubt true these generalisations leave open the exact nature of the relationship between reward, value and desire. We will address this question by starting from predictive processing theories in cognitive neuroscience that offer explanations of reward and value in terms of probabilistic expectation. The predictive processing theory suggests that desire should be understood in terms of its effects on attentional processes. Attentional processes estimate the significance of prediction error. Desires steer the predictive organism towards or away from things of reward or value in the environment. What this leaves unexplained however is the phenomenology of desire - why should it be that when prediction errors are assigned high precision this exerts a strong push or pull on the agent? We provide an answer to this question, and thus offer an account of the phenomenology of desire in predictive processing terms.

5.1 Introduction

People tend to feel good when they attain something they desire, and frustrated when they fail to do so. They tend to value positively feelings of pleasure, and value negatively feelings of frustration. While no doubt true these generalisations leave open the exact nature of the relationship between reward, value and desire. We will address this question by starting from predictive processing theories in cognitive neuroscience that offer explanations of reward and value in terms of probabilistic expectation. In the predictive processing (PP) theory reward and value map onto processes in the brain that generate predictions of the sensory and bodily states the organism expects to occupy. The brain then processes information in such a way as to keep the error in its predictions to a minimum. Our main aim in this paper is to determine how to think about desire in the PP theory.

Timothy Schroeder (2004) has developed a naturalistic theory of desire that explains desire in terms of its effects on reward learning in the brain. He has shown how desire has three faces which show up in its effects on what an individual finds pleasurable, how the individual is motivated to act, and learning how to frequent states that are valuable or rewarding, and avoid not occupying those states. We make an analogous proposal in relation to PP models that take value and reward to be the brain's predictions. Does it follow then desire should also be thought of as driving prediction-error minimisation (PEM)?

The predictive processing (PP) framework suggests that desire has its effects on feeling, motivation and learning through its effects on attentional processes. We will show how attentional processes estimate the significance of prediction error. Desires thus steers the predictive organism towards or away from things of significance in its environment. We propose desire has these effects by generating in the organism bodily states of affective tension in relation to aspects of its surroundings. We therefore add an important ecological twist to the PP story. We propose a view of desire as states of a larger organism-environment system. Desires manifest as embodied states of action-readiness that move us towards or away from possibilities for action we care about as agents.

Our paper is divided into five main sections. In the first section we briefly explain how desire, reward and value are standardly taken to be related in the reward learning theory of desire (Schroeder 2004). Section two shows how reward learning is understood within PP. We show how the reward learning system is implicated in attentional processing, setting the precision on prediction errors at multiple scales in the brain based on what the organism has found to be rewarding and valuable. The PP account of desire however seems to fail to account for the phenomenology of being a desiring creature. In section three we show how PP may after all be able to explain the phenomenology of desire. We argue that the role of reward and value in modulating attention should be understood as taking place in the organism's body and not only in its brain, as the organism as a whole prepares to act on what matters to it in the environment. In section four we show how the predictive organism weighs precision on the basis of affective tension with the environment. We show how affective tension is rooted in expectations of how fast or slow the organism expects to reduce prediction error. The organism doesn't only track the significance of prediction error but also how fast or slow it is reducing prediction error relative to what it was expecting. We call this rate of change in error reduction, "error dynamics." Section five shows how error dynamics can account for what Schroeder calls the three faces of desire. Desire has its effects on feeling, motivation and learning because of an organism's sensitivity to error dynamics. We conclude that in PP the phenomenology of desire is due to the organism's

being drawn into action by the multiple things it cares about in the environment. Desire guides the organism to seek out spaces in which they can simultaneously maintain grip on and balance the many things that matter to them in their environment.

5.2 Reward, value and desire

If you have felt the pangs of hunger, the heat of romance, or the burning need to do better in a skilled activity, you know what it is to desire. Desire is the force of attraction that leads us through the world, it is the felt urge that gets us moving towards what matters to us, and it is the source of our pleasure and frustration as we succeed or fail to attain what we care about.

While we know what desire is from our first-person experience of being desiring creatures, how does this first-person experience relate to biological processes describable in the third-person terms of the natural sciences? In what follows we will mostly be concerned with intrinsic desires - desires whose target is states of affairs that are wanted for themselves and not as means to other ends. Examples are pleasure, avoidance of suffering, adequate nutrition and hydration, the affection of others, and wellbeing of those one cares about, and so on. Intrinsic desires are distinguished from instrumental desires in which the object of one's desire is a means to attaining some other end that is intrinsically desired. A person might desire money for instance not as an end in its own right but because it buys power and influence, or simply as a means to providing for one's family. Intrinsic desires are arguably an essential part of a person's moral psychology. They help to guide the person towards courses of action that are right, virtuous and praiseworthy (Arpaly & Schroeder 2013). Our concern in this paper is however not so much with the moral work intrinsic desires do for the person. Our concern is with the nature of desire and what it is for people to desire the things they do.

One way naturalistic philosophers have commonly understood desire is by looking at its common effects on cognition and behaviour. To desire a thing has been conceptualized as having a tendency to pay attention to that thing when it is present (Scanlon 1998), behaving in ways that the agent believes will help them attain that thing (at least when appropriate) (Smith 1987), and a tendency to feel good when we attain the object of our desire and suffer when we don't (Davis 1986; Morillo 1990; Schueler 1995). The naturalisation of desire has then proceeded by appeal to neuroscientific research to account for these effects. Timothy Schroeder (2004) for instance developed a neuroscientifically grounded theory of the effects of desire in terms of reward-based learning mechanisms in the brain. According to Schroeder, to desire an object X is

for a representation of X (perceived or imagined) to have the power to contribute to the calculations that produce reward-based learning signals (as described below). In other words, to desire p is for representations of p to drive reward-based learning in part in the form of dopamine production and distribution.

Reward learning is now a well studied model of probabilistic learning in which the agent learns which kinds of actions are to be valued in themselves and which actions one should have an aversion to because the costs of performing them outweigh the benefits (Montague, Dayan & Sejnowski 1996; Schultz, Tremblay & Hollerman 2000). The outcome of this learning is that the agent comes to associate an action with either a positive or negative value. This value is the expected reward or punishment. Reward learning takes place when an action or event elicits a "reward prediction error" (RPE) signal. This signal reports something better or worse than expected just happened. The reward prediction error signal is the expected reward value less the actual reward received. Dopamine neurons in two midbrain structures called the ventral tegmental area (VTA) and the pars compacta of the substantia nigra (SNpc) have been found to fire at rates that directly correlate with an RPE signal. These subcortical midbrain areas are believed to be the core of what's known as the brain's reward system (although reward learning should be thought of as the result of a constellation of areas working in concert).

Consider as an illustration the famous experiments in which a monkey was given a sip of sweet juice one second after a light was shown (Schultz and Romo 1990; Romo and Schultz 1990). The monkey has no response to the light initially, but responds to the juice with a huge spike in dopamine. After a short time of this pairing being presented the dopamine activation changes. Now, it is the light that activates the dopamine spike, while the arrival of the juice ceases to make a difference. By this time the monkey has come to unconsciously expect the juice upon seeing the light, the arrival of the juice is now taken for granted. However, what remains uncertain is when the light will be presented. The light comes to function as a cue that predicts reward. What is important now for the organism is tracking and learning about the occurrence of the light. In the final stage of the experiment, the light is shown (which increased dopamine) but the juice is withheld. The result is a drastic decrease in the production of dopamine when the juice was not received. The monkey has learned the light no longer predicts sweet juice.

Reward learning steers an agent through the world, making it highly likely that they will come into contact with and thus attain the things they intrinsically desire. A natural consequence of this type of learning is that habits solidify - we learn to make more or less automatic and unthinking use of opportunities for reward when they arise. As long as there are predictable and

learnable opportunities for rewards in the environment then the organism will act in ways that will increase the probability of obtaining rewards, while decreasing the probability of failing to attain what is desired. Reward learning tunes the organism to what is valuable motivating the organism to seek out what is rewarding and avoid what is aversive in ways that lead to pleasure and not to pain. These are Schroeder's three faces of desire: pleasure, motivation and learning (Schroeder 2004).

The reward theory of desire takes desire to consist in representations of objects or states of affairs that *drive* reward based learning. In the next section we will see how in the prediction error minimisation (PEM) framework reward and value are explained *as predictions that drive error reduction routines*. Does it follow from this account of reward and value that desire should also be understood in terms of PEM?

5.3 The predictive processing theory of reward

It is intuitive to think that agents are motivated to seek out rewards and to avoid punishments, and that reward learning teaches agents how to frequent rewarding spaces more often. Predictive processing theories of decision-making however reverse this intuitive way of thinking about expectation and reward (Fitzgerald et al 2014; Friston et al 2012; see also Friston, Mattout and Kilner 2011 pp 138). Reward is understood in terms of the sensory states the organism frequently occupies and punishment with sensory states that are infrequently occupied. The frequency with which a location in a state space of possible behaviours is visited (or a signal is encountered) equals the reward value (Moutoussis, Story and Dolan 2015). This is because the states the agent ought to revisit regularly are the states that are highly probable given the environment it inhabits and the kind of being it is. Think of a fish in water: the sensory states associated with an aquatic environment are highly probable for the fish. Being out of water is by contrast highly improbable for most fish. Should they find themselves momentarily out of water they need to take immediate action to return to their familiar environment. The agent thus needs to avoid improbable states that threaten its integrity, or in some other way conflict with the life it leads.

PP claims that agents come to develop or "embody" a "generative model" of their environment that is shaped by a history of reward and punishment, and thus by what the agent finds valuable (Friston 2011). Based on the generative model the organism predicts that it will occupy on average and over time the sensory states that are rewarding to it. These predictions are fulfilled

when they fit with the sensory states the organism tends to occupy on average and over time. In an effort to improve this fit, humans are able to make use of both habitual (pragmatic policies) and goal-directed behaviours (including epistemic policies) (Pezzulo et al 2015). Some of the action possibilities the agent acts on will serve purely pragmatic ends. They will be action possibilities that are already trusted and familiar, and the agent can simply rely upon habitual ways of acting in exploiting them. For example, the action sequence that constitutes my sipping coffee while I work on this paper takes care of itself. Other action possibilities the agent selects will serve epistemic ends of reducing uncertainty. If I want a coffee but am working in a new office today, then action policies are chosen that help fill out our expectations about such spaces so that in the future pragmatic policies may be applied (eg. I wander around the nearby area looking for a cafe). The agent selects exploratory actions in ways that yield new information that can be used to guide action in improved ways reducing its uncertainty about the environment in the future (Kiverstein et al 2017).

The organism will thus select actions that fulfill the predictions of its generative model. This will however only result in the organism occupying rewarding and valued sensory states if the organism can be confident in the predictions of its generative model. To minimise prediction error in the long run the organism must select actions that have a high probability of taking them from the sensory states they currently occupy to the rewarding sensory states they expect to frequent. They must minimise the divergence between the outcomes that are likely given a certain course of action and the outcomes that are desired.[44] The organism will need to maximise their certainty that a policy (a sequence of actions) will minimise the difference between the predicted and valued outcomes in a given context. It will need to evaluate the "precision" of the policies it uses to select actions (Schwartenbeck et al 2015).

"Precision" refers to the agent's confidence in a policy. More precisely, it refers to the degree of confidence that the predicted sensory consequences of action will match as closely as possibly the sensory states the agent values. Thus when hungry the agent expects food. The brain predicts the consequences of the sequence of actions (the policy) that leads to the agent finding and consuming food. The agent must then select the policy it has the most confidence will lead it to satisfy its hunger in its current context.

[44] This is technically referred to as the Kullback-Leibler divergence - the difference between the agent's prior preferences (the desired outcomes) and the posterior beliefs about likely outcomes given current sensory observations (i.e. the agent's current context).

Schwartenbeck and colleagues have shown that changes in expected precision directly correlate with the activity of dopaminergic neurons in the midbrain (Schwartenbeck et al 2015).[45] Organisms should expect different degrees of precision in different contexts. A policy that works in one context may no longer work in a different context. We can thus think of organisms as harboring prior beliefs about the precision of their behavioural policies - beliefs about how probable it is that a policy will lead to desired outcomes in a given context. If the agent is to minimise prediction error in the long run it will prove essential for the precision of a policy to be continuously updated based on the agent's current sensory observations. The agent should only act on a policy when they are confident that doing so is more likely to take them from their current sensory states to the outcomes they desire than acting on some other policy. Based on precision expectations, prediction errors will be assigned different precision weightings. The weighting of prediction errors will determine the influence prediction errors have on downstream processing. Prediction errors that are weighted high will have high impact, while prediction errors that are weighted low because of weak confidence in a policy will be muted.

Desire thus reemerges here as the effect of interactions among precision weighted probabilistic expectations that arise in response to the dynamically changing environment. It has the three characteristic effects that Schroeder (2004) describes on pleasure, motivation and learning as a consequence of prediction-error minimisation. The sensory states the organism expects to regularly frequent just are states with positive hedonic value it feels good to occupy. The agent is motivated to act when predictions errors that arise from mismatch in its predictions and its current sensory states are weighed as precise. It is so motivated because of learning driven by the precision-modulated imperative to minimise prediction error in the long-run.

One key weakness of this proposal is its apparent failure to account for the rich phenomenology of desire, or what it is like to be a desiring creature (e.g. Seth 2013; Clark 2017b). It fails to account for the "oomph" of desire - the way in which desires affect us by energising the body, motivating the agent to seek out and attain valued objects. Precision explains why certain policies are acted on, but what it doesn't so obviously explain is why there should be a tension in the body, an urge or need to take action that can sometimes show up in the agent as a feeling of

[45] Precision expectations closely resemble reward prediction errors. A reward prediction error recall reports the difference between expected and actual reward. Precision expectations behave like reward prediction errors by contributing to processes that maximise expected utility. The reward prediction error in PP reports change in confidence that actions will lead to rewarding outcomes (Schwartenbeck et al 2015: 3442). In PP however decisions are made not only to maximise expected utility but in addition to minimise long term prediction error. Unsurprising states are the rewarding states. So by minimising prediction error (i.e. by avoiding surprise) the agent will in the long run be guaranteed to maximise the probability of occupying rewarding states.

affective tension. We draw out the problem further in the next section, and propose a solution in section 4.

5.4 Being moved towards what matters

Desires arise in the body and manifest for the agent as an affective tension that draws us towards the things we want. As Koffka (1935) put it, "a fruit says, 'Eat me'; water says, 'Drink me'; thunder says, 'Fear me', and woman says, 'Love me'" (p. 7). Both Lewin and Koffka called this the 'demand character' of the environment (Withagen et al 2012: p.251).

Intrinsic desires are embodied affective states that relate to what an organism cares about in its environment. The organism always stands in a particular evaluative relation to its environment because every organism as a living being is simultaneously "in a state of relative equilibrium and in a state of disequilibrium" (Merleau-Ponty 1968/2003: p.149; c.f. Bruineberg et al 2018, Kiverstein et al 2017). The organism is in a state of relative equilibrium insofar as it succeeds in maintaining its biological organisation in its interactions with the environment. It is simultaneously in a state of disequilibrium so long as it remains alive, since there will always be something the organism needs but currently lacks. The organism will thus be moved to act so as to reduce its state of disequilibrium. However it never fully succeeds in attaining equilibrium. To do so would mean death. Both the environment and the organism are dynamical systems that continuously undergo change. To act in a dynamically changing environment, the organism will continually need to adjust its actions to its own changing needs and to unexpected changes as they arise.

The phenomenology of desire can be aptly described, we propose, as an organism's tending towards an optimal grip on what is currently significant in its environment (Bruineberg et al 2018).[46] This bodily stance is something the organism must actively maintain in relation to its current situation, hence our describing it in terms of *grip*. We talk of grip as tending towards *optimality* because the organism is aiming at a relative state of equilibrium with the environment.

[46] In earlier work we have shown how the tendency towards optimal grip is consequence of a more general imperative of the organism to minimise what Karl Friston has referred to as "variational free energy". We showed how the concept of variational free energy can be understood in dynamical systems terms as the disattunement between internal dynamics on the side of the organism and external dynamics on the side of the environment. See (Bruineberg et al 2018, Kiverstein at al 2017).

The environment makes available to the organism many possibilities for action or affordances. Affordances are the possibilities for action provided by the substances, surfaces and other inhabitants of an animal's environment (Gibson 1979). Affordances are relevant for an agent when they contribute in some way to the agent's concern to remain in the bodily states it expects to occupy. Recall that it is in terms of the bodily states the agent expects to occupy that PP understands reward and value. Affordances stand out as relevant for an agent when there is a mismatch between the bodily states the agent predicts and the bodily state it currently finds itself occupying. In other words, affordances stand out to the organism as relevant because of prediction errors: the bodily states the agent expects on average and in the long-run fail to match with those it currently occupies.

Now let us us return to the problem we raised at the end of the previous section. We can now see why the actions that are selected on the basis of precision-estimations of prediction errors should affect the agent to different degrees. Relevant affordances attract or draw the agent towards them or force the agent away from them. They elicit in agents *an affective tension* with varying degrees of intensity or importance. The variability in tension, we propose, can be thought of as precision weighting. Recall how precision weighting is the outcome of the updating of precision expectations. It is not the case that precision is first set and the consequence of this is affective tension in the agent. We propose instead that precision weighting and affective tension are one and the same process. Prediction errors move the agent to act only when they are assigned high precision. The affective tension that is elicited in the agent by relevant affordances is identical with prediction error that is being assigned high-precision. We develop this proposal in more detail in the next section.

5.5 Error dynamics as precision engineering

Agents situated in an environment rich with affordances will typically be affected not just by one affordance but rather by a multiplicity of relevant affordances simultaneously. Relevant affordances can thus be described as forming a field in which each affordance has a varying degree of valence. Affordances differ in the degree of urgency with which they solicit. Some demand we act on them now, while for other affordances in the field the need to act is less urgent. The relevant affordances that demand we act now correspond with expectations for immediate and fast error reduction. While the relevant affordances that are assigned less importance correspond with expectations for slower rate of error reduction. The variations in valence that relate to the urgency of action map onto what we will refer to as "error dynamics".

The term "valence" is standardly used to refer to the felt positive or negative character of our experience (Barrett 2006; for possible doubts see Colombetti 2005). We propose to understand valence in terms of affordance-related embodied states of action-readiness (Frijda 2007). The inviting or soliciting character of relevant affordances can be given a description in phenomenological terms as the agent feeling attracted by certain affordances that she cares about, or being pushed away and repelled by other affordances that present a risk or threat (Dreyfus & Kelly 2007). Relevant affordances thus have a valence.

Error dynamics refers to temporal comparisons of the rate of error reduction. We can think of the rate of change in prediction error reduction by analogy with velocity (Joffily & Corricelli 2013: 3). The velocity of an object is the rate of change in the position of an object relative to a frame of reference over time. So velocity is equivalent to the speed of an object moving in a particular direction. Rate of change of prediction error reduction refers to how fast or slow prediction error is being reduced relative to what the agent expected. The agent's expectations for fast or slow error reduction map onto affective tension. The greater the tension, the more urgency there is to reduce prediction error fast. Relevant affordances that elicit less tension do so because the agent expects to reduce prediction error for those affordances over longer time scales.

The degree of urgency given to a relevant affordance shouldn't be conflated with the importance of a relevant affordance. Many possibilities that are extremely important to us such as attaining a PhD or writing a book can only be achieved over very long time scales. They are assigned high importance perhaps outweighing all the other things in our lives that matter to us. But still we don't expect error relating to these possibilities to be reduced fully and immediately. We do expect however to make progress in reducing error at a particular rate. So on some days when we are stuck on a problem it feels bad because we are not making the progress we expected. This may lead us to change perspective on a problem so that we find once again things begin to improve and we are back on track. The degree of urgency that error reduction has is premised on things that are important to us. There is something we lack that is the source of an increase in disequilibrium in relation to the environment. It is this disequilibrium that is the source of affective tension. As we just explained tension comes in degrees, and it is the degree of tension that influences how fast or slow the agent expects to reduce error.

We have seen above how prediction errors arise when the agent fails to occupy the bodily state it expects (those that are rewarding and valuable). We describe this in terms of an increase in

disequilibrium in relation the environment. On some occasions this mismatch will be a matter of urgent importance, on other occasions less so. The more important it is to the organism to immediately occupy rewarding and valuable bodily states the greater the build up of affective tension.

Precision expectations may be optimised in a given context in part on the basis of error dynamics. Second order error comparisons of rate in error reduction are made available to the organism as affective feelings. Suppose the agent is expecting to reduce error fast but there are no affordances available to help them do this. This will feel bad. You are feeling hungry and you go to the cupboard for the snacks that can normally be found there. But all the snacks have been eaten by your partner. Your confidence in the policy of where to find the snacks when hungry needs to be updated. The failure to reduce error at the rate that was expected gives rise to negatively valenced feelings of frustration. This feeling gets to do part of the work of updating one's precision expectations. One should no longer place so much confidence in the snack cupboard when it comes to the policies one selects for reducing hunger. Now consider what it feels like for things to go better than expected. You are hungry again and you go to the snack cupboard this time with the expectation that you'll find it empty. But your partner has been shopping. Things go better than you expect, and this feels good.

Our claim is not simply that positive or negative feelings are the product of predictive confirmations or violations. It is not the overly simple claim that we feel good whenever our predictions fit the sensory evidence, and bad when they don't. The organism is always managing some error within its hierarchy of expectations, but it is clearly doesn't feel each and every fluctuation. Instead errors are considered relative to expectations about the rate at which errors have been resolved in the past. Over time the organism comes to predict the rate at which error rises and falls in particular circumstances. One result of this is that error could be experienced positively, if it was expected as part of a more general error reducing regimen. Take for example the necessary errors we expect to encounter while acquiring/mastering a new skill. What comes through as a feeling is unexpected change in the rate at which error is being reduced. This sensitivity to rate of error reduction can help guide the predictive organism towards niches that are replete with reducible errors that are neither too complex to manage, and yet not too well learned and so unuseful.

Also important for optimising precision expectations is a sensitivity to changes in *the overall rate* at which errors are being reduced or increasing over time. We suggest that the agent is sensitive to how well or badly it is doing in gripping to the field of relevant affordances as a whole, and that

therefore the field as a whole can be thought of as having a valence. When you ask a person how their day is going they will initially answer by giving a general overall sense of how they are doing. They say they are doing well or badly or perhaps just fine. They are indicating something about their overall grip on the situation as a whole given what matters to them.

Human agents are ready to respond in an *integrated* manner to multiple relevant affordances simultaneously. Their states of action readiness combine in coherent ways so as to open them to a whole field of relevant affordances. By using error dynamics to tune precision expectations, the agent can thereby ensure that she continues to remain attuned to the field as a whole. Precision-expectations that are set in part based on error dynamics can enable agents to create and maintain *metastable* poise in relation to the environment. Metastable poise allows for fast and flexible switching between action possibilities (Kelso 2012). Think of the boxer finding an optimal distance from the boxing bag where she is ready for all the relevant affordances the bag offers (Bruineberg & Rietveld 2014). She is ready to make jabs, uppercuts and hooks based on her distance from the bag. Given this bodily readiness it can then be a random fluctuation of the bag that contributes to the selection of which action unfolds (see Rietveld, Denys & van Westen forthcoming). This is crucial for adapting rapidly to situations in a way that does not need conscious control.

Remaining metastably poised in relation to the field as a whole plays an important role in learning. It allows the agent to strike the right balance between relying on what they already know and exploring in search of new information. This matters because it is necessary for minimising expected prediction error in the long run. There is always the possibility of the agent ending up in surprising sensory states that are punishing precisely because they are unexpected. What the agent already knows won't always be sufficient for dealing with unexpected prediction error in the future, and thus it won't help them in the long run to always return to the familiar and rewarding sensory states. Thus to deal with unexpected increases in uncertainty in the future agents will need to be ready to switch from what they do as a matter of habit. They must be prepared to sometimes favour policies of exploring the environment in search of new action policies that lead to surprising and unexpected events. As we make our way through the world we encounter various stable and persisting opportunities for action (attractors in the language of dynamics systems). These fixed points emerge and dissolve both due to environmental conditions, and changes in our own internal states and behaviours. However, we also have a tendency to actively destroy these fixed points therefore inducing instabilities and creating peripatetic or itinerant (wandering) dynamics (Friston, Breakspear, and Deco 2012). This type of readiness falls naturally out of maintaining metastable attunement. Maintaining metastable

attunement just is equivalent to finding the goldilocks zone with just the right mixture of novelty or uncertainty, and what is already well-predicted because of the current model of the environment (Kidd, et al 2012; Van de Cruys 2017; Kiverstein at al 2017). An agent that is prepared to break with habit in this way will be an agent that is well placed to return consistently to the sensory states it expects to be in (i.e. the ones that are rewarding and valuable).

When all goes well precision weighting works to maintain metastable attunement with the environment. It is by remaining metastably poised in relation to the field of relevant affordances as a whole that we do best at adjusting to a dynamically changing environment and thus at maintaining a grip on it. Agents that are sensitive to error dynamics can use this sensitivity to update their precision expectations thereby ensuring they remain metastably poised.

5.6 Error dynamics and the three faces of desire

We are now in a position to show how Schroeder's three faces of desire - pleasure, motivation, and learning - are a consequences of a sensitivity to error dynamics. We thus show how PP can account for the phenomenology of desire once we have error dynamics in view. *Pleasure* is the result of doing better than expected at attaining the familiar states that are expected. Familiar expected states recall are states that are valued and rewarding precisely because of their familiarity. They are the bodily states the organism should expect to occupy given the life it leads. Pleasure arises in us when unexpected opportunities to reduce errors arise. Error means there is an important bodily state that one is predicting that doesn't match one's current bodily state. The prediction error is given high precision when it is one that the agent expects to reduce immediately through action. Pleasure thus arises when the agent finds an unexpected opportunity that fulfils this expectation.

The individual is *motivated* to act by the relevant affordances of its environment in ways that reduce the affective tension that arises from its being in a state of disequilibrium with its environment. Relevant affordances elicit in the agent multiple embodied states of action readiness. Each of these states of action readiness varies in urgency. If the agent is to maintain metastable attunement these states of action readiness must be integrated in such a way that the agent is able to manage the overall rate of error reduction. Overall rate of error reduction is important because error arises for each of the things in the environment that matter to the agent. Therefore so long as the agent is succeeding in managing *overall* error this will mean they

are maintaining metastable attunement. They are managing to maintain a grip on the many things that matter to them.

The organism that is sensitive to error dynamics will be a creature that *learns* how to return consistently to the sensory states it expects to be in when things change either in the environment or on the side of the agent. It will be able to harness its learning to adapt to change and will be motivated to seek out better than expected opportunities for reducing prediction error. Recall that the agent isn't only interested in reducing current prediction error but in minimising expected future uncertainty. This is essential for an organism that is able to minimise the effects of random fluctuations in the environment. Minimising expected future uncertainty requires the agent to continuously be on the look-out for better than expected opportunities for reducing error. To see this consider the following three scenarios for an agent that is sensitive to error dynamics and uses this sensitivity for setting precision.

In the first scenario precision is set high for a prediction error that relates to a relevant affordance - lets say a door handle that opens a door. The door handle turns, the agent enters the room and everything unfolds as they expected. Thus the agent doesn't learn anything new about the affordances of door handles. They already know what they need to know.

In the second scenario everything is the same initially as in the first scenario. But this time when the agent tries to turn the door handle the door refuses to open. Prediction error is thus not reduced at the rate the agent was expecting. Now the agent can learn something new - something about the direction you turn this type of door handle, or that the door is locked and the meeting they are late for is probably happening elsewhere.

In the third scenario you encounter a door hand handle that has proven really hard to open in the past. You've stood there for many frustrating and embarrassing minutes trying to work out what to do to no avail. This time when you try to your astonishment the door springs open upon your first attempt. Things go much better than you were expecting. You take a moment to check what the secret was, and it works again. You have made progress in your learning.

An agent that acts to minimise its own expected uncertainty will as we've seen above need to be an agent that is ready to break with habit. They will need to be the kind of agent that is motivated to make progress in learning. As we saw above, the agent cannot always do this by ploughing the same ground, relying on what is already known. To deal well with uncertainty in the long run and return to the familiar and expected sensory states continuously, the agent will

need to be curious. They will need to sometimes actively seek out niches that are replete with reducible errors that are neither too complex to manage, nor too well learned to be useless for making further progress in learning.

The arguments of this paper imply that intrinsic desires arise in organisms continuously striving to maintain grip on a dynamically changing environment. Desire manifests as embodied states of action readiness that are elicited by affordances that are relevant given the organism's disequilibrium with its environment. It follows that - at an admittedly high-level of abstraction and generality - everything that an organism cares about will be a reflection of this more basic need to continuously be reducing disattunement with the environment. What organisms intrinsically desire is to remain well-attuned to relevant opportunities and risks the environment furnishes.

5.7 Conclusion

Our aim in this paper has been to use recent developments in cognitive neuroscience to develop a naturalistic theory of desire. An increasingly influential theory in cognitive neuroscience takes the brain to regulate the organism's interactions with the environment in such a way as to guarantee that the organism finds itself in rewarding and valuable sensory states. The brain does this by building up a model of its environment which it uses to predict its own sensory input.

In PP desire is understood as tuning precision expectations. A weakness of this proposal is however its apparent failure to account for the phenomenology of desire and the way in which desires are felt in the body. We have corrected for this failing by showing how precision can be understood as affective tension. Affective tension is elicited in the body by relevant affordances based on prediction errors that reflect a mismatch between the bodily states the agent expects to occupy - the bodily states that are rewarding and valued - and those it currently occupies. The agent expects to reduce prediction error at a given rate. When this expectation is satisfied or violated this gives rise to feelings that can be used to update precision expectations. The effects of desire on pleasure, motivation and learning are thus explained by the sensitivity to error dynamics. Pleasure, motivation and learning are all consequences of desire. It follows that desire has its effects through its contribution to long term prediction error minimisation.

Precision is weighted in such a way as to ensure over time that the agent is able to balance the multiple possibilities for action they care about. Sensitivity to error dynamics tells the organism

how well it is doing at maintaining metastable poise. It guarantees that the agent assigns the right priority to the right states of action readiness at the right time, so as to remain flexibly poised to switch between the many activities they care about.

Is it a consequence of our argument that all the predictive organism wants in the end is to keep the long-term prediction errors of its generative model to a minimum? We don't think so. Instead we suggest organisms desire to keep to a minimum the affective tensions that arise from disequilibrium with the environment. They want to continuously return to a state of relative equilibrium with the environment when disequilibrium begins to increase as it inevitably will so long as they remain alive. Prediction error minimisation is driven by this basic need or concern of the organism to remain metastably poised. Metastable attunement matters to organisms because it allows them to balance the many aspects of the socio-material environment that they care about and remain flexibly open to the field of affordances as a whole.

We've argued that what the agent intrinsically desires is to maintain metastable poise. Doing so allows them to be maximally responsive to all the possibilities for action that matter to them. By setting precision on the basis of error dynamics the agent is able to balance the many things that matter to them. It thus turns out that the effects of desire are best understood in terms of prediction error minimisation as it plays out within a whole agent-environment system in which the agent is continuously aiming to do justice simultaneously to all the things that matter to them.

Chapter 6
Embodying addiction: a predictive processing account

Abstract

Addicts engage in increasingly self-destructive cycles of behaviour. Often they continue to do so long after the behaviours have ceased to bring them any pleasure. They do things they don't want to do at great emotional costs to themselves and those around them. In this paper we show how addiction can be thought of as the outcome of learning. This is a hypothesis widely held in the empirical literature on addiction. We propose however an account of learning as a self-organising process that leads to the progressive entrainment of the behaviour of addicts by the environment. We show further how to understand this self-organising process using the increasingly influential predictive processing theory of the brain. Perhaps counter intuitively, it is a consequence of our argument that while the brain plays a deep and important role in leading a person into addiction, it cannot be the whole story. Predictive processing is best interpreted in the wider context of the agent's dynamic coupling with its environment. The pathological nature of addiction is thus not to be found in the brains of addicts, we will argue, but in the larger organism-environment system of which the brain is an important part. Our predictive processing account unlike other models of addiction is able to do justice to the complex constellation of causal factors - biological, biographical, societal and historical - that lead a person into addiction. A larger agent-environment system is broad enough to include all of these factors. As a dynamical explanation it can also provide tools for explaining how these factors combine in ways that conspire to lead a person into addiction.

6.1 Introduction

Addiction has a devastating effect upon those whose life it afflicts. Addicts find their life increasingly dominated by their addictive behaviours. The other pursuits they care about begin to be crowded out as they devote increasing amounts of time and energy to the pursuit of their addictions. The undesirable outcomes of their addictive behaviours are increasingly ignored, yet at the same time addicts feel compelled to continue acting on their addictions often long after

the addictive behaviour has ceased to bring any pleasure. Addiction can reach a point in a person's life where it seems all that matters to them is doing what their addiction requires, yet at the same time this is something they do not want. As the director of the National Institute on Drug Abuse Nora Volkow has said "I've never come across a single person that was addicted that wanted to be addicted" (Gugliotta 2003).

In what follows we will propose an account of addiction as a self-organising process that spirals out of control due to feedback loops that entrain the behaviour of the agent, locking them into destructive cycles of behaviours. It is in the dynamic interaction between the agent and its environment that addiction is born and endures. Addiction is the outcome of the formation of deep habitual tendencies further enhanced and strengthened by the impact addictive substances have on brain circuitry set up to build (but also to control) habits. Although we focus on substance addiction in what follows, our account generalises to other forms of addiction such as gambling, eating, sex, shopping, exercise, video games, online social media, and even work[47]. We'll argue that addiction in general should be understood as a self-organising dynamical process that unfolds through the agent's coupling with its environment.

Many accounts of addiction in the literature give a central place to the learning of habitual modes of behaviour that override whatever other goals the agent may value (Lewis 2017; Heather et al 2017; Robinson & Berridge 2008). Our account differs from other learning models in that it conceives of addictive behaviour not as a conditioned "response to stimuli" but as an "active engagement (or entrainment) with meaningful aspects of the environment" (Lewis & Shelly under review). The environment is best understood not as made up of cues that passively trigger the addict to act on their habits. It is made up of meaningful possibilities for action, that is to say it is made up of affordances (Gibson 1979). We will suggest that what is pathological in addiction is the narrowing of the agent's field of relevant affordances to just those that contribute in some way to attaining the object of their addiction. To understand how this narrowing occurs, we will argue, requires that we zoom out and take into consideration the wider organism-environment interactions.

Our account provides an integrated framework for understanding the multiple causal factors that lead to addiction - biological, historical, social and cultural. We may therefore be able to reconcile previously opposing perspectives on addiction. All accounts of addiction stress the importance

---

[47] These behavioural forms of addiction share much in common with substance addiction, both in terms of genetic predisposition (Lejoyeux et al 1997; Grant & Kim 2002; Raymond et al 2003) and changes in specific neurochemistry (Reuter et al 2005; Goudriaan et al 2006).

of recognising the complex suite of causes that lead up to addiction. On one side of the debate a prominent proponent of the medical model of addiction, Leshner writes:

"Addiction is not just a brain disease. It is a brain disease for which the social contexts in which it has both developed and is expressed are critically important… If we understand addiction as a prototypical psycho-biological illness, with critical biological, behavioral, and social-context components, our treatment strategies must include biological, behavioral, and social-context elements. Not only must the underlying brain disease be treated, but the behavioral and social cue components must also be addressed, just as they are with many other brain diseases, including stroke, schizophrenia, and Alzheimer's disease." (1997 p.46)

On the other side of the debate, opponents of the medical model also emphasise the importance of the societal and historical causes of addiction (Alexander 2008; Levy 2013; Heilig et al 2016). These models typically emphasize, for example, that there is now converging evidence that physical abuse, economic inequality and injustice, and psychological trauma in early life increases the likelihood of addiction in the future (Sinha 2008; Satel & Lilienfeld 2013).

The account of addiction we go on to develop can do justice to what is right in both these camps without falling on either side of this divide. Medical models of addiction acknowledge that social and environmental factors play a role in the development of addictions. However they have an unfortunate tendency to downplay the agency of the addict, assigning too much importance to the brain, while the contribution of the environment is only to provide stimulation that passively drives the behaviour of addicts. The behaviour of the addict is treated as like a stimulus-response behaviour, the outcome of operant conditioning. On the other side, in the research on the social and historical causes of addiction, there is an unfortunate tendency to downplay the importance of the brain in the development of addiction. Learning models of addiction emphasize the role of the person's environment and life experiences, but in doing so tend not to take into account the important changes that take place within the addicts brain. Interestingly in common with the medical models these accounts also treat the agent as largely passive in the causal history that leads up to their addiction. The addict is passively acted on by their historical and social circumstances. Both these accounts fail to strike the right balance between explaining addiction in terms of its environmental causes and explaining addiction in terms of its biological causes. We will argue by contrast that addiction is best understood as a phenomenon of the whole agent-environment system.

We share much in common with medical accounts in proposing a learning model of addiction. However learning as we will describe it is not a process of operant conditioning as it is often taken to be in medical accounts of addiction. We will show learning is instead best understood in terms of predictive processing. As research is beginning to show the organism as a whole may best be described as a predictive system, facilitating intelligent behavior by predicting the sensory consequences of its behaviour and keeping any resulting prediction errors to a minimum. The brain in this theory is hierarchically organised with each layer attempting to predict changing patterns of activity in the layer below so as to build up the best overall guess as to what is currently causing sensory input in the world. Error signals propagate forwards in the hierarchy, and are used by the system either to improve future predictions, or to generate actions that when all goes well lead to sensory inputs that cancel out errors.

A key component in the predictive processing theory is a mechanism that estimates the uncertainty or "precision" associated with an action policy - a sequence of actions. Precision is here understood as the probability that an action policy will lead from the agent's current states to the outcomes the agent desires. We will show how addiction may be the outcome of precision estimation. So far this may sound very much like we are proposing to explain addiction purely in neurobiological terms. This however is not the case. We take predictive processing to be a self-organising process that needs to be understood in terms of the agent's dynamical coupling to its environment (Bruineberg et al 2018).[48] Precision expectations tune the organism to the possibilities for action that matter in the environment. We show how this process of context-sensitive updating of precision expectations is what goes awry in addiction.

The remainder of our paper is organised as follows. In section 1 we outline the effects of addictive behaviours on reward processing in the midbrain. The medical model of addition has led to advances in our understanding of the effects of repetition of addictive behaviours on the dopaminergic systems in the midbrain. They have interpreted the changes in these brain areas that addictive behaviours induce in terms of reward learning. In section 2 we show how to understand the effects of substance use on reward learning in predictive processing terms. We show how the hijacking of reward learning can be redescribed in terms of precision weighting. In section 3 we draw on our earlier work on the role of what we call "error dynamics" in precision weighting (Kiverstein at al 2017). "Error dynamics" refers to the expectation of the

[48] We call this interpretation of predictive processing (PP) the ecological-enactive account. It is closely related to the treatment of PP found in Clark (2016) though there are also some important differences in how we characterise the situatedness of the organism in its environment, and in our treatment of the generative model. We return to the latter concept briefly in section five.

organism to reduce prediction error at a certain rate. We show how error dynamics are best understood in the larger context of an organism aiming to improve grip on the things that matter in its environment. Section 4 then applies this model of precision weighting to arrive at an account of what is going wrong in addiction, where the problem lies not only in the brain but in the wider brain-body-environment dynamics. The brain is of course a necessary part of this story but it isn't sufficient for understanding what goes wrong in addiction. In the final section, we develop this predictive processing account further. We show that it is the dynamics of the addict's coupling with the environment that makes their behaviour pathological. We finish up by considering the implications of this hypothesis for the treatment of addicts, and what a person would need to do to effect long-lasting change.

6.2 "Mutiny in the mid-brain"[49]

Normally reward processing in mid-brain areas leads us through the world in ways that increase our contact with what we find valuable and rewarding. Agents learn about values (e.g. expected rewards) in the world by minimizing the "reward prediction error" (RPE) which is the expected reward value less the actual reward received. Dopamine neurons in midbrain areas have been found to fire at rates that directly correlate with RPE signals and perceived reward values (Montague et al 1996; Schultz et al 1997). This strongly supports the hypothesis that dopamine functions as a learning signal in the brain. By signalling the mismatch between predicted and actual reward the agent learns to make the most of these rewarding opportunities when they arise. Most animals respond with a similar reward learning signal when presented with objects of intrinsic desires such as food, water and opportunities to mate. In a similar manner humans respond to such things as money, success, favorite songs, and the flourishing of loved ones.

Substances of addiction produce a similar response in the reward learning system - they produce a burst of dopamine as if the organism was encountering something which is intrinsically desired. They induce learning signals that convince the brain that the agent is progressing in the world in ways they intrinsically desire. This by itself is not the problem. The real problem lies in how drugs of addiction (and indeed other addictive pursuits) are able to "hijack" this system, and over time produce pathological levels of wanting, cravings and compulsive acting (Robinson & Berridge 2008). Addictive substances promote unconscious learning in ways that over time progressively become untied from what the agent finds valuable. If we have no desire to hear

[49] We borrow this description of the neural processes underlying addiction from Flanagan (2017).

punk music attending a concert isn't likely to be rewarding, but even if one hates the idea of taking heroin the substance itself acts on the brain as if the substance was highly desired.

RPE signals naturally diminish over time as the agent comes to learn about and anticipate regularities between cues and behaviours that lead to rewarding outcomes. Once a particular reward has become highly expected it ceases to produce the same reward signal (Rescorla & Wagner 1972; Sutton & Barto 1981) This fails to happens with drugs of addiction. Addictive drugs consistently produce strong reward learning signals long after non-drug induced learning signals should have vanished (Robinson & Berridge 2008). No matter how much the system expects the drug to be satisfying, it always registers as much more than that.

Robinson and Berridge have offered a formula for how compulsive behaviour grows in addiction (1993, 2000, 2001, 2008). They call this process "incentive sensitisation".[50] An important benchmark in the development of addiction is the point at which initial hedonic effects, the feelings of pleasure associated with drug use, begin to wane, while the substance seeking and taking continue to become more consuming, more habitual and ultimately compulsive. A wealth of animal studies suggest that at the level of the brain this change from fascination to habit to compulsion occurs as the neural locus of behavioural control shifts from prefrontal cortical areas to striatal subcortical areas, and again within the the striatum itself processing moves from ventral to dorsal domains (Everitt & Robbins 2013). Once the dorsal striatum gets involved there is a an increasingly high probability that the learned behaviour will flow from the cues regardless of perceived changes in value or reward.[51]

This represents the final stage of addiction. This transition from attraction to compulsion, from being attracted to and engaged by some stimuli to unthinking habitual behaviour, is all a natural part of learning. For example, when our decision to wake up and run every morning stops being a conscious effort and becomes an automatic part of our morning routine, it is because repetition and reward has encoded the selection of those behaviours in the dorsal striatum

---

[50] Although their work primarily looks at animal studies, the areas of the brain they are interested in are not so different in mice and humans.

[51] Compulsivity isn't a consequence of this shift from cortical to sub-cortical control alone. This kind of shift takes place with any habitual behaviour, and is a natural part of skill learning but not all habitual or skilled behaviours are compulsive. Addiction hijacks that process in an unexpected way quickly leading to undesirable actions (drug seeking behaviours) becoming more and more powerful (craving) and less and less controllable (habitual). Part of the power of this compulsivity is due to the fact that in addiction the dorsal striatum begins to pull away from the influence of cortical areas (such as the DLPFC) believed to ordinarily regulate and contextualize habitual responses (Yamamoto et al 2015; Lewis 2015). This leads to unthinking behavioural routines that can be driven by environmental cues that elicit powerful urges regardless of their appropriateness.

(Wolfensteller & Ruge 2012). Addiction however is often described as "hijacking" this process, leading to undesirable actions (drug seeking behaviours) becoming more and more powerful (craving) and less and less controllable (compulsive). Once incentive sensitisation has happened addiction is hard if not impossible to escape.

Although the brain clearly plays an important role in the etiology of addiction as the research we reviewed in this section clearly establishes, it would be a mistake to conclude that addiction is just a disorder of the brain. We will argue that addiction is more to do with their bodily orientation towards the world. It isn't just a problem located in their heads. To see this will however require us to better understand the contribution of the reward learning system in the brain to addiction.

6.3 The predictive processing perspective: reward tunes precision expectations

Reward-based learning is standardly understood as the process by which the organism maximises expected utility while minimising costs and avoiding punishment (Sutton & Barto 1998; Delgado et al 2005; Arpaly & Schroeder 2013). In recent years a different account of reward learning has begun to emerge that takes predictions of reward to be part and parcel of the prediction of the sensory consequences of our practical engagement with the world (Friston et al 2009, Friston et al 2012; Fitzgerald et al 2014; Clark 2016). It seems intuitive to think that we are motivated to seek out rewards and avoid punishments. Over time we thereby learn to frequent rewarding spaces (and avoid punishing spaces) more often than not. The predictive processing theory (PP) turns this intuition about expectation and reward on its head (Fitzgerald et al 2014; Friston et al 2012). In contrast to traditional reinforcement learning models where decisions are made based on what set of actions will maximize expected utility, in PP the rewarding sensory states are the sensory states the agent expects to occupy given the model of the environment it has developed.

In models of reward-based learning the dopamine system is taken to track reward prediction errors. In PP dopamine performs a related but different function. Rewarding outcomes are those the organism expects given its phenotype and past-learning. They are the effect of the organism predicting over multiple spatial and temporal scales the sensory consequence of its actions. The organism's choices are based on probabilistic beliefs about its action policies. The organism selects those actions that are most likely to lead to the outcomes it expects, which as we have seen in PP are rewarding outcomes. Beliefs about action policies are associated with confidence or precision. "Precision" refers to the organism's degree of confidence in the predictions that a

certain policy, or set of behaviours, will bring about outcomes the organism values. Technically it refers to the inverse variance in the mean of prediction error.

When precision is expected to be high the "gain" is increased on prediction errors, thereby increasing the impact those error signals have on processing higher in the hierarchy. Error signals are just the sensory and physiological consequences of action the brain expects that fail to match the organism's current sensory and physiological states. By increasing the gain on certain prediction errors this can have the consequence of prioritising certain actions as means of reducing prediction error. Precision thus ensures that the organism is always preparing for actions the effects of which are valued. Dopamine instead of being used for reward learning does the work of updating precision expectations in ways that fit the agent's current context. When all goes well the agent should weigh confidence in its action policies so as to maximise the probability that their actions will lead from their current sensory states to the outcomes they desire (Schwartenbeck et al 2014).

Precision expectations track among other things what the organism assigns relevance to in its environment. There are always many things agents care about, and based on precision weighting they are able to balance the many things that matter to them. Precision expectations should thus be thought of as the agent's bodily stance in relation to what is relevant in its environment. The environment makes available to the organism many possibilities for action or affordances. Affordances are the possibilities for action provided by the substances, surfaces and other inhabitants of an animal's environment (Gibson 1979). Affordances stand out as relevant for an agent when there is a mismatch between the bodily states the agent predicts based on its past history of interaction with the environment, and the bodily state it currently finds itself occupying. The relevant affordances that demand immediate action will give rise to prediction errors that are assigned high precision.

Agents situated in an environment rich with affordances will typically be affected by a multiplicity of relevant affordances simultaneously. Ideally an agent should be ready to respond in an integrated way to what we describe as a "field of relevant affordances" (Rietveld & Kiverstein 2014; Bruineberg et al 2018). Each of the affordances in the field will be relevant to the agent to different degrees. Some will call out for immediate action. The prediction errors that relate to acting upon those affordances will be assigned a high precision because of the need to act immediately. The agent will thus expect fast error reduction for those relevant affordances that are being assigned high precision. Other relevant affordances will be assigned less importance. The agent can afford to delay acting on those possibilities. The agent will thus

expect slower rate of error reduction for those relevant affordances. If the agent expects to reduce error more slowly this may be because error reduction is more complex and will take them longer to deal with, or it may be because immediate error reduction isn't so urgent. The agent can afford to wait. The more precise prediction error are weighted, the more urgency or priority is then assigned to reducing those prediction errors.

Precision expectations are thus in part expectations the agent has to reduce error with a certain expedience. In the next section we explain how an embodied sensitivity to what we will call "error dynamics" plays an important role in learning precise policies. Error dynamics refers to the rate of change in error reduction, or how fast or slow the sum of prediction error is being reduced relative to the agent's expectations (Kiverstein at al 2017). Later we will show how this embodied sensitivity to error dynamics is what is hijacked in cases of addiction. First we must explain the role error dynamics play in the context-sensitive updating of precision, and why we think this needs to be understood in the larger context of the agent-environment system.

6.4 The role of error dynamics in tuning precision expectations

We've argued that precision expectations can be thought of as tracking relevance - the significance that is given to prediction errors in a given context. When the context changes either because of something on the side of the agent or something in the world, it is important that the agent be ready to adapt what they are doing to this change. This may mean they restructure the precision assigned to multiple prediction errors so that possibilities that were being assigned high weighting are now down-weighted in favour of other possibilities.

We suggest that agents that aim at long-term minimisation in prediction error will benefit from being sensitive to error dynamics. This sensitivity takes the form of positive and negative bodily feedback felt by the person as a whole. The organism isn't only interested in keeping prediction error to a minimum in its sensory exchanges with the environment. It doesn't only aim to get itself into the states it expects given what it cares about. Also important for setting precision is a sensitivity to changes in the overall rate at which errors are being reduced or increasing over time. Each agent's performance in reducing error can be plotted as a slope that depicts the speed at which errors are being accommodated over time. The steepness of the slope indicates that error is being reduced over a shorter period of time and so faster than the agent expected: the steeper the slope, the *faster* the rate of reduction. Rate of change in error reduction refers to how fast or slow the sum of prediction error is being reduced relative to expectations (Kiverstein at al

2017). If the speed of error reduction increases, this equates to a decrease in prediction error over time (relative to what was expected). If speed of error reduction decreases, this equates to an increase in prediction error over time.

These global or summed error comparisons are made available to the system as embodied feelings (Kiverstein et al 2017; Van de Cruys 2017; Joffily & Coricelli 2013). Feelings of positive and negative affect emerge as a reflection the the quality of the organism's engagement with the environment (see also Polani, 2009). These feelings are embodied as part of valuation process that works as a sort of bodily barometer keeping the organism informed about how it is fairing in its practical engagement with the environment and preparing the agent to act so as to improve its overall situation (Barrett 2017). Positive feelings are related to positive rates of change in error reduction relative to what is expected. It works in the opposite way for negative feelings - they provide bodily feedback to the agent that error has been reduced at a slower than expected rate.

Agent's are normally sensitive to the rise and fall in error reduction and will be able to make use of information about how well they are doing overall in reducing error to learn precise policies - policies that in the long run are maximally likely to reduce prediction error. This will allow them to balance how they distribute precision so that they can on the one hand stay in touch with the many possibilities that matter to them. On the other hand they will be able to redistribute how precision is assigned when something changes and they find error beginning to increase in ways that run contrary to what was expected. Sensitivity to error dynamics thus allows agents to stay in touch with and do justice to each of the many things they care about.

Consider what happens when the agent is in a high state of affective tension with the environment. You are a smoker and you find yourself stuck on an airplane. Your cigarettes are inviting you to smoke but you are not allowed to smoke. Your body is thus telling you that there is relevant source of error that you have the means to immediately reduce but this is not permitted. This situation is felt in the body of the agent as an unpleasant feeling of persisting error or tension. The agent is doing worse than they expected at reducing error. They expect to reduce error fast but their expectation remains frustrated so long as they are flying. This negative feeling may lead the agent to explore the environment for other alternative possibilities to smoking that reduce tension for the duration of the flight. Relevant possibilities that might now stand out soliciting them to act may possible distractions such as ordering another drink, or seeking comfort through complaining to your partner about the unfair smoking policies of airlines.

Contrast this situation with one in which the agent comes across an unexpected opportunity to reduce error. You are in a second-hand book shop and there on the shelf is a rare book you've for years been searching for. An opportunity to reduce error has popped up where you were not expecting one. You have suddenly done much better than you were expecting and this feels good to you. Precision expectations are thus updated by feedback from feelings. This good feeling is part of what sets precision on the policies that will end up leading us back to the same bookshop again in the future in the hopes of acquiring similar rewards.

Precision weighting then isn't just a brainy event. Precision expectations are updated or tuned to the context based on bodily feelings that track and leverage opportunities to improve at reducing error in line with our expectations. In the next section we show how it is exactly this sensitivity to error dynamics which substances of addiction tap into and hijack. In PP addiction is understood as a consequence of pathological precision weighting that leads to habits overwhelming other of our goals, and a loss of contextualisation of habits by higher-level goals and desires. We suggest a different way of thinking about this loss of contextualisation by reference to the ideas on embodied error dynamics we've just introduced. We will see that addictive substances act on the brain in ways that result in a shrinking of the space of possibilities to which the agent attends

6.5 Addiction as tending towards a s*ub-optimal* grip

It feels good to the agent when prediction error is reduced more efficiently than expected for the activities the agent cares about. For example if a particular reduction in error is expected, such as the error reduced when we scratch an itch, little if no positive feeling would emerge. The ideal situation for the agent is one in which error is being reduced at a faster rate than was expected. When this happens the organism can sense that it is doing better than expected. It feels good. Now consider what happens in substance addiction as the habit of using the substance gets a grip on the person. The pleasure that the drug elicits in the first stages of addiction signals that using the drug has offered an unexpected opportunity to reduce affective tension with the world. Predictive agents are always expecting to reduce tension, but the drug exceeds the agent's expectation of reducing tension. In addition to alleviating the tensions relating to the drug seeking and taking behaviour, addictive substances also reduce tensions relating to wider disattunement with the environment. Things overall thus seem to go even better than expected. Each time they act on this policy - the seek out and use the drug - the same things happens. Instead of their expectations simply being met, which would normally signal to the brain nothing

new to be learned here, the brain responds by producing dopamine that signals that there is still something new and surprising being learned.

As the policy is repeated, so the agent begins to predict cues that are associated with the pursuit of the policy, such as being in a particular neighbourhood where you can score the drug. These predictions than give rise to prediction errors that are assigned high precision and so get to drive behaviour to actively seek out those cues. The agent will follow the trajectories (action policies) that lead to the drug. Eventually the agent finds more and more of their everyday life being taken over by pursuit of policies associated with drug use.

Repeated use of drugs of addiction can thus be thought of as training expectations for error reduction at a certain rate. Importantly, this is the source of pleasure that comes with drug use (at least in the early days) on our account. Drugs of addiction act directly on the system that is signalling the probability that a policy does a good job of reducing prediction error. Thus it makes it seem there is now something they can do - namely seek out and take the drug - that does a much better job than expected at getting them into the rewarding bodily states they expect. They make it appear to the agent as if error in relation to the many things that matter is being reduced rapidly at a rate that is faster than anything the agent has anticipated.

The production of dopamine caused by the drug makes it seem to the organism as if the policy of seeking and using the drug is the most reliable way of getting itself into the states it expects to occupy. Thus the policy of seeking and using the substances soon comes to be the policy the organism has the most confidence in. As soon as the drug wears off affective tension begins to increase again. Nothing was in fact resolved in the world through taking the drug. There was only the illusion of error reduction. In fact the addict often finds themselves in a worse situation as is reflected in the negative affect associated with feelings of guilt and shame in the short term and loss of health in the long term. Cravings in the addict can be thought of as the affective tension that can only be resolved by pursuing the policy of finding and taking the substance. Thus the cycle of seeking and using takes hold and exerts a tighter and tighter grip on the agent. The addict has now come to expect a certain rate of error reduction - they have come to expect to do better than expected at reducing tension for all of the things that matter in their lives. So long as this fails to happen they feel bad because they are failing to meet their expected slope of error reduction (cf. Koob & Moal 2001, 2005).

What is pathological about this is that it leads to akrasia, which means it leads to pervasive error in the rest of the life of the addict. What might seem like a fool proof way to reduce uncertainty

is in fact no such thing, The agent increasingly loses touch with all the other possibilities that matter to them resulting in more error in the long run. This also has the consequence that the possibility to explore and gather new evidence is down-weighted relative to the option of continuing to exploit the known consequences of using the substance. Addicts choose for the familiar option, and continue to do so even when the outcomes are negative. They don't gather more evidence that might lead them to change their behaviour.

Habits can be thought of as various stable and persisting opportunities for action (or "attractors" in the language of dynamical systems). These fixed point attractors emerge and dissolve both due to environmental conditions and changes in our own internal states and behaviours. However, agents also tend to actively destroy fixed point attractors therefore inducing instabilities and creating peripatetic or itinerant (wandering) dynamics (Friston, Breakspear, & Deco 2012).[52]

Predictive organisms don't only seek to maximize error reduction, but rather are driven to reduce error at a particular rate (Kiverstein at al 2017). They are willing to disrupt their own fixed-point attractors (habitual policies) in order to explore just-uncertain-enough environments that are ripe for long term prediction error minimisation.[53] Predictive organisms are thus are the kinds of agents that desire to continuously do better at reducing uncertainty in their engagement with the environment.

We suggest the reason addicts don't explore and gather new evidence may be that substances of addiction make the person feel (at least temporarily) like they are well-attuned even though they are not. They create an illusion of attunement to the environment. Drugs of addiction 'cheat' the "affordance competition" (Cisek & Kalaska 2010) - they direct the whole predictive organism to track and engage with (i.e. self organize around) signals that register in the brain as highly valuable in terms of adaptive success, but in fact are precisely the opposite. In the worse case scenarios, such as long term opioid addiction, it may be the case that in fact no skills, relationships, or resources are gained or improved during the repetitive drug seeking and taking behaviours. This further perpetuates the cycle that collapses the field of relevant affordances -

---

[52] Clark has recently written, "Friston suggests, our 'neural expectations' may come to include expectations of 'itinerant trajectories' mandating change, exploration, and search. We 'expect' to sometimes engage in random environmental search as a means of entering into adaptively valuable states. To put it crudely, we randomly sample because - qua evolved organisms - we 'expect' to discover food, mates, or water at some point during the expedition" (2017).

[53] Schwartenbeck et al (2013) extend this direction of thinking by proposing that certain policies may be valuable insofar as they open the way the agent to visit multiple other states.

there is more and more error from the many negative effects of addiction, and less and less skillful development and support.

Sensitivity to error dynamics is one way that good habits are woven into our skillful engagement incrementally over time - all directed to what matters to the organism. Drugs of addiction as we have seen hijack this sensitivity leading the system to self-organize in relation to the environment in ways that lead agents to neglect the many other things in their lives that also matter to them in favour of the policy of feeding their addiction. Over time, the pursuit of this policy is expected to a greater and greater degree, while other opportunities for rewarding behaviour are increasingly ignored.

What is pathological we suggest is the crowding out of the other action options that would normally also exert an affective pull on a person because they also matter to the person. But once the addiction has taken hold there is just one strongly alluring possibility that comes to be expected. We can understand this as the gradual collapse, or shrinking, of what we call "the field of relevant affordances". The very same mechanisms that normally produce curiosity and exploration (Kiverstein at al 2017) once hijacked by the addictive substance produce precisely the opposite effect. Instead of being moved to pursue the multiple possibilities we care about, the hijacked learning system leads the person to engage the world in tighter and tighter circles of habitual behaviours.

Addiction is not a case then of desire overwhelming a person's better judgement (see Holton 2009, ch.7), but instead it is a consequence of our sensitivity to error dynamics no longer functioning to keep us in contact with what we desire. In other words, precision expectations are no longer being modulated relative to context in such a way as to allow the agent to maintain grip on the field of affordances as a whole (Kiverstein & Rietveld 2015; Bruineberg et al 2018). Drug related affordances in the environment become increasingly powerful, dominating and silencing other behavioural policies. Instead of maintaining grip on the many things that matter to them in the environment, the addict find themselves increasingly being gripped by the now drug-seeking infused field of affordances.

6.6 Why addiction isn't just a brain disorder, and why it matters

In section 1 we briefly reviewed the substantial evidence that addictive substances have an impact on the brain's dopaminergic circuitry. However, as Lewis points out all rewarding

activities produce changes in dopamine transmission (Lewis 2017). Thus the changes drug use induces in the brain are not in themselves evidence for the claim that addiction is a disorder of the brain. Addiction isn't a consequence of the dysfunction of the prediction error minimizing system in the addict's brain. It is the result of processes of prediction error minimisation that are working optimally being led astray by chemically induced changes in precision weighting (Schwartenbeck et al 2015). The problems stem from the suboptimal generative model of the environment the agent comes to embody, and where the agent expects to find precise prediction error. The "generative model" as we employ this term should be understood as the whole organism in relation to its ecological niche (Bruineberg & Rietveld 2014; Bruineberg et al 2018; Friston et al 2011; Allen & Friston 2016; Kirchhoff 2017). It is not a model inside of the brain that the organism has but is something the organism develops over time. Friston et al (2012) write:

> "We must here understand 'model' in the most inclusive sense, as combining interpretive dispositions, morphology, and neural architecture, and as implying a highly tuned 'fit' between the active, embodied organism and the embedded environment" (2012, p. 6).

It is in this embodied sense of 'model' that Friston claims that an "agent does not have a model of its world – it is a model." (2013, p. 213; see also Bruineberg et al 2018; Kirchhoff 2016, 2017). Every biological agent can be described as a probability distribution – a hierarchically organised probabilistic model conditioned on the sensory, physiological and morphological states that are highly probable given the life it leads and the eco-niche it inhabits (Friston 2010, 2013).

Schwartenbeck and colleagues (2015) hypothesise that the generative model in addicts owes its suboptimality to the precision that is assigned to the addict's habitual behaviours. Habitual behaviours become more highly expected, while goal-directed behaviours become less highly expected. Drugs of addiction create this imbalance by re-tuning the confidence that some policy of behaviours will result in expected sensory states. We've argued that this re-tuning happens through a sensitivity to error dynamics. Sensitivity to the rise and fall in error reduction plays a crucial role in helping an agent to balance the multiple relevant affordances, staying in touch with many of the possibilities that matters to them when something in the agent's situation changes. Addictive substances progressively drive behaviour to the neglect of other possibilities that are also of concern to the agent. They do so because they make it seem to the agent as if an improvement has taken place in how well the organism is gripping the field of affordances as a whole. Drugs of addiction each time they are consumed signal to the organism that something better than expected has taken place. Unfortunately, the temporary state of chemically-

scaffolded grip wears off all too quickly, only to be replaced with more error and uncertainty. Meanwhile the addict is progressively losing touch with the other things they care about.

The organism's brain is not malfunctioning, but is in fact doing what comes naturally, by increasingly honing in on the opportunities it expects to lead it to what is important in its environment. Through their impact on the dopamine system addictive substances signal to an otherwise optimal prediction-error minimising system that it is indeed progressing in reducing error, while in fact the exact the opposite is often taking place. We should therefore resist the temptation to attribute the many problems common to addiction to any single neurochemical alteration. To understand what is pathological in addiction we need to consider how those neurochemical changes are taken up into wider patterns of behaviour, choice and learning. While changes in associative-reward circuitry certainly plays a necessary role in the development of addiction, we have been arguing that these changes are not sufficient to explain addiction.

We can thus avoid the dichotomy of either addiction is a biological disorder or it is a purely social phenomenon whose causes lie for instance in poverty or in the urban environment. In the place of this dichotomy we suggest a predictive processing account of addiction. Such an account is better supported by what we know about the brains of addicts. Our account claims that the pathological behaviours of addicts are the result of *disorganization* within the agent-environment system as a whole. Human agents enter into a circular causal relationship with their surroundings. The organism's perception of its environment, its actions and its feelings are co-determining. It is this dynamic relationship between the organisms and the environment that is disrupted in addiction. From this perspective addiction is best characterized not as a change in particular neural circuitry, but as as a more general loss of attunement of the organism and its environment. With the addition of error dynamics, PP can offer a rich explanation of how neural computational processes contribute to the breakdown in this wider organism-environment system and do justice to its phenomenology in the form of feelings of doing better or worse.

Alva Noë has offered an analysis of addiction that goes in a similar direction to our account. He has suggested that addiction should be understood more globally as a break down in the dynamic interplay between the addict's goals and preferences and their wider behaviours. He writes,

"normally there is a dynamic quality to our actions and preferences, just as there is with those of rats. We enjoy exercising, but we soon get tired or bored. But rest, too, soon loses its appeal. We eat, and then we are sated. And then we are ready for the treadmill again. And so on. Things

have gradually changing and complementary values. In addiction, this dynamic goes rigid. The addicts goal assumes a fixed value, and the value of everything shrinks to zero, and with terrible costs" (2011).

We agree with Noë's analysis, but extend it further to include the dynamics between the organism's embodied affective states and the field of affordances. Ordinarily individuals are responsive to multiple relevant affordances simultaneously. In addiction the multiple concerns of the individual are silenced by the all consuming need that develops with addiction. Drugs of addiction impact the systems that help us stay tuned to the many things that matter to us. While the brain is registering that the agent improving at reducing error in relation to the many things that matter, the opposite is in fact taking place.

Once we view addiction as a phenomenon of the whole agent-environment system, we can do justice to accounts of addiction that emphasise its societal causes (e.g Sullivan 2018). Recall how affective tension builds up in the body when the bodily states one currently occupies do not match the rewarding and valued states one expects to occupy. One expects to be reducing error at a particular rate (to reach a particular slope of error reduction given the context) and when one fails to do so this is experienced by the agent as a negative feeling. Consider a person who is constantly facing hunger because they don't have the money to buy food, and is cold because they are unable to pay to heat their home. They expect to be well-fed and to stay warm but their socio-economic status means that meeting these expectations is a continuous struggle. People faced with such a struggle to meet their expected slope of error reduction might be more attracted to the possibility to "self-medicate", as it is sometimes described. Once they have discovered the possibility to reduce affective tension with the world in ways that otherwise prove a struggle, one can imagine the temptation to do so repeatedly might be high. Marc Lewis makes this point well in relation to the susceptibility of people struggling with PTSD and depression to addiction:

> "Importantly, it's not just attraction or desire that fuels feedback loops and promotes neural habits. Depression and anxiety also develop through feedback. The more we think sad or fearful thoughts, the more synapses get strung together to generate scenarios of loneliness or danger, and the more likely we are to practice strategies—often unconsciously—for dealing with those scenarios. Neural patterns forged by desire can complement and merge with those born of depression or anxiety. In fact, that's a lynchpin in the self-medication model of addiction. Gabor Maté persuasively shows how early emotional disturbances steer us toward an intense desire for the relief provided by drugs (2008), and Maia Szalavitz vividly portrays her experience as a late adolescent trying to brighten her depression with cocaine and ease her anxiety with heroin (2016). So, when we examine the correlation between addiction and depression or anxiety, we should recognize that

> addiction is often a partner or even an extension of a developmental pattern already set in motion, not simply a newcomer who happened to show up one day" (Lewis 2017, p. 10).

As we have argued above what feels good to agents is to be continuously improving in error reduction in relation to the many possibilities that matter to them. So long as the agent is doing this, they are managing to maintain grip on the field of affordances as a whole. Sometimes a person's life however offers only the prospect of more uncertainty - think of soldiers that become addicted to substances while away in a strange land in a war situation. They can make a predictable and somewhat more comforting reality for themselves out of what is otherwise a confusing reality through substance abuse because the substance can be trusted to have certain guaranteed and predictable physiological effects on the body. Once the soldiers return home to the predictable and familiar reality, drugs no longer present the attraction they once held. There are better policies available to the soldiers for improving attunement with the world. This may go some way towards explaining why rates of heroin addiction were high among soldiers stationed in Vietnam but upon returning home addiction rates fell back to their normal rates. The behavior of the soldiers stationed in Vietnam was in this respect somewhat similar to that of the rats in the famous *Rat Park* studies (Alexander, Coambs & Hadaway 1978; Alexander 2010; Ahmed et al 2013; Hari 2015; Solinas et al 2008). One group of rats were placed in simple cages all alone, but with plentiful opportunity to consume as much opioids as they wanted. For such a rat addiction was an inevitable outcome. When the same rats, now addicted to the substance, were moved to a much larger cage with other rats and a variety of games and opportunities for improving they tended to ignore the available opiates altogether. Given the current proposal, we think this could be explainable insofar as the rats were able to now meet their expected slope of error reduction, just like the soldiers returning home from Vietnam (Robins 1993; Robins et al 1975; Granfield & Cloud 1999). In addiction the agent is increasingly gripped by the environment until they cease to be open to the other non-drug related possibilities that matter to them. The way through addiction then is likely to be in part environmental enrichment. We suggest addiction recovery could be facilitated by changing the expected rate (the slope) of error reduction itself through restructuring (relearning) expectations for where to look in the landscape of affordances for error reduction. A key part of undoing such habits we suggest will be developing new and different skills for reducing error more efficiently, such as techniques of emotional regulation and mindfulness (Garland et al. 2014). People may find their way out of addiction by learning to contextualise their processes of habit-selection through emotion regulation, so that they restore openness to the many possibilities that matter to them.

6.7 Conclusion

In this paper we have argued that addiction must be investigated from the much wider vantage point that includes the whole organism-environment system. The brain of the addict is in fact doing what the brain is meant to do when viewed from the standpoint of predictive processing. It is continually optimizing the fit of the organism with its environment relative to what matters to the organism. However, addictive substances make it seem to the organism as if error had been reduced but sadly for the addict this is just an illusion. The result in the long-run is inevitably a greater amount of uncertainty arising from a loss of a sensitivity to the wider concerns of life.

If all predictive organisms care about is reducing error why isn't the life addicts lead at least one viable strategy for prediction error minimisation? Addicts become extremely skilled at organising their lives around the goals of finding and using the addictive substance. They develop models that are optimised to fit an environment in which these are the only things that matter. We've argued that predictive organism's don't only try and reduce error but reduce it at a particular rate. It might be thought however that this is exactly what the addict is doing as they get increasingly skilled at navigating the environment they come to inhabit.

What this misses however is the way in which all the drug can deliver is short-term reduction in error. The life of the addict becomes increasingly chaotic in other regards. As soon as the drug's effect wears off, what they return to is a world offering all of the uncertainty that never really went away. So long as the addict is high, it seems to them as if they are succeeding at maintaining grip on what matters to them. Once the drug wears off, they find reality is very different. Substance addiction has been likened to a single room with many paths that all in the end lead the addict back into the same room again. The room of addiction is however fraught with difficulties and dangers. The progressive loss of touch with the rest of what matters leads long term addicts to inevitably struggle with loss of material possessions and personal relationships, diminished self worth, and physical health problems. Addiction thus leads to long term increases in error in relation to all the other things that matter to the addict. Humans have come to expect over time to maintain relationships that matter to them, and to hold onto their possessions, and to remain healthy. In addiction however they act in ways that frustrate these expectations. The point at which an addict decides to make a change is sometimes referred to as "rock bottom". This is the point at which what the addict has *actually* lost finally outweighs what they *feel* they are gaining.

The model of the environment the addict becomes is a model of an environment that is tailored and built around the all consuming activity of feeding their habit. What counts as an improvement with regards to fitting this environment is dictated not by finding a balance among the many things that matter to the addict. It is instead dictated to the agent by the increasingly wide range of possibilities that lead them back into the same vicious cycle of behaviour. What is pathological about addiction is thus not to be found inside of the brains of addicts but in their wider engagement with life, and with the environment they enact.

Conclusions

The *predictive processing* (PP) framework explored in the chapters of this thesis has offered a number of revolutionary contributions to the study of the mind. However, one description of these developments threatens to revitalize a classical "internalist" perspective regarding the relationship between the brain, body and world. According to one popular characterization of the framework, cognition is now the result of an evolving internal model constituted by a set of hierarchically structured neurons. If cognition is prediction, and prediction is a brainy affair, then we know where to find the mind - in the skull, secluded from the body. As Hohwy summarizes, "PEM should make us resist conceptions of this relation on which the mind is in some fundamental way open or porous to the world, or on which it is in some strong sense embodied" (2016: 259). This depiction of cognition threatens to set us back into the *bad old days* of disembodied neurocentricism (see Anderson & Chemero 2013).

However, this approach to PP has not gone without opposition. There is an emerging group of researchers who take the very same framework as offering a systems level description of some of the core tenets of the embodied cognitive science paradigm. I wholeheartedly agree. However, as we have seen throughout this project, many of these so called "embodied" PP approaches turn out to be only "moderately" embodied, meaning they continue to under-appreciate or ignore completely the role the body plays in cognitive processes.

The focus of this project has been to show how a better understanding of the role of the living body can reveal the intimate relationship between emotion and cognition and help inform the construction of a more fully embodied vision of the predictive mind. To develop this view, I have drawn heavily on recent dynamic network models of the brain. Such models have captured the interest and imagination of the philosophical community in recent years, due in large part to the novel perspective they propose regarding traditional ontological categories of mind such as emotion and cognition (Colombetti 2014; Pessoa 2013; Lewis 2005). Traditional locationist models of the brain characterized emotion and cognition as linearly related, hierarchically structured and easily separable processes in the brain. In contrast, these dynamic network models suggest that emotion and cognition (our familiar names for functionalities that are not cleanly distinguishable at the circuit level) interact in dynamic, co-evolving and deeply interdependent

ways. I take this depiction of emotion and cognition as support for a view of cognition as inevitably affective, and so also deeply embodied.

Throughout this project I have applied lessons from this research on emotion to the PP framework. I have argued that PP and dynamic network models form a powerful synergy. On the one hand, dynamic network models can offer neurobiological explanations of how emotion and motivation influence cognitive, perceptual and behavioural processes in just the ways that embodied models of PP have recently suggested (Clark 2016; Pezzulo et al. 2015; Seth & Friston 2016). As we have seen, interoception is increasingly being given a special role in RPP accounts. Information from inside the organism's body is described as directing the whole error-minimizing regime in ways that keep the predictive organism in contact with what it cares about. To fulfill this role, embodied PP accounts depict interoception as participating in a dynamic and non-linear exchange with cognitive and behavioural processes. It is here that I see dynamic network models helping to put some much needed neurobiological flesh on an otherwise still largely theoretical PP skeleton. On the other hand, PP can provide an elegant systems level account of why these various cognitive-emotional dynamics unfold as they do, as part of an overall drive to continually reduce free-energy during embodied exchanges between the organism and the environment.

While there is indeed an underlying logic in PP that clearly eschews the easy separation of processes such as perception, cognition and action, in favour of a "seamless" dynamical integration of these processes, there remains a perverse and pervasive cognitivism in the field when it comes to including the realm of affectivity into this mix. And so while RPP is happy to help itself to the sorts of entangled neural architectures I have been describing throughout, researchers have not yet fully appreciated the depth of the embodiment those entanglements may entail (good examples of this are Pezzulo et al. 2015; Seth & Friston 2016)

As a means of helping PP to overcome this lingering 'cognitive-cortical myopia', I have proposed a treatment of precision weighting as closely related to affective significance. Building on recent neuroscientific research, I argue that precision weighting should be understood as specifically emphasizing signals that have *value* for the organism. Precision turns out to be much less about reliability in terms of high fidelity, and much more to do with reliably producing the affective and behavioural changes that allow the organism to secure a better more beneficial relationship between itself and the environment. What becomes obvious here is that the predictive brain is not, after all, a rational scientist out to construct an objectively accurate model of the world (Hohwy 2013), but rather a crooked scientist whose only aim is to confirm the

world it predicts will keep it alive and thriving (Bruineberg et al. 2016). This type of view echoes and extends a motor-centric vision of the brain (see Churchland et al. 1994). While such a model would depart from more traditional Helmholtzian takes on predictive processing, it fits perfectly well within more embodied and enactive takes of PP.

In the final half of this project I push this notion of precision-based affectivity and affectivity-reflecting precision to its extreme by linking precision weighting to *embodied error dynamics*. In the picture I develop, precision is set in part based on affective tensions in the body, which arise due to changes between the organism and its ecological niche. These affective tensions, which occur relative to changes in the *rate* of expected error reduction, are interpreted as patterns of action readiness (Frijda). A consequence of my hypothesis is that instead of thinking of precision as a form of top-down constraint operating at the higher-levels of hierarchy, we can think of the contextualization that precision weighting is believed to offer a little differently. I propose it is the larger metastable system of the whole organism that sets precision. It does so in part on the basis of bodily states of action readiness, which I define as the organism's evaluation of what is of current and future relevance in relation to its surrounding environment. I take this view of precision weighting to be supportive of an ecological and enactive view of PP (as also seen in Bruineberg & Rietveld 2014; Bruineberg et al. 2018).

While this proposal remains in many ways theoretical, there are good reasons for taking such an approach seriously. For one, it suggests a vision of PP that can do justice to the rich relationship between emotion and cognition that recent neuroscience has described. Second, it can provide insights into a number of philosophical puzzles. As I have shown, this framework can provide elegant naturalistic accounts of philosophically interesting concepts such as curiosity, desire and value, and addiction. And finally, the framework I have developed can provide new and more satisfying answers to some the more difficult philosophical and scientific challenges levelled against the PP theory of mind – for example, the worry that it marks a return to an outmoded form of internalism.

Generative models, error minimizing strategies and precision weighting get us a long way in making sense of what drives the neuroeconomy. But as I have argued throughout this project, a complete picture of human cognition will require us to also include the many ways that the predictive brain operates as part of a wider system which includes a situated body that acts, feels and cares. This work thus respects Clark's injunction to "confront predictions in the wild" (2015). By exploring the rich entanglements between the prediction brain and the feeling-acting-caring body, I have tried to develop a picture that takes us far beyond the merely modest

embodiment of first wave PP. I hope this helps set the stage for a future science that foregrounds the living body as the systemic anchor point for the predictive brain.

Bibliography

Ackermann H., Riecker A. (2010). The contribution(s) of the insula to speech communication: a review of the clinical and functional imaging literature. *Brain Struct Funct,* 214 (5–6).

Ahmed, S. H., Lenoir, M., & Guillem, K. (2013). Neurobiology of addiction versus drug use driven by lack of choice. *Current opinion in neurobiology*, *23*(4), 581-587.

Ainley, V., Apps, M. A., Fotopoulou, A., & Tsakiris, M. (2016). 'Bodily precision': a predictive coding account of individual differences in interoceptive accuracy. *Phil. Trans. R. Soc. B*, *371*(1708), 20160003.

Akins, K. (2006). Of sensory systems and the "aboutness" of mental states. *Journal of Philosophy*, *93*(7), 337–372.

Alexander, B. (2010). Addiction: The View from Rat Park. *Retrieved July*, *26*, 2015.

Alexander, B. K., Coambs, R. B., & Hadaway, P. F. (1978). The effect of housing and gender on morphine self-administration in rats. *Psychopharmacology*, *58*(2), 175-179.

Allen, M., Fardo, F., Dietz, M. J., Hillebrandt, H., Friston, K. J., Rees, G., Roepstorff, A. (2016). Anterior insula coordinates hierarchical processing of tactile mismatch responses. *Neuroimage*, 127 34-43.

Allen, M., & Friston, K. J. (2018). From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese*, 1-24.

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

Anderson, M. J. (2014). *After Phrenology: Neural Reuse and the Interactive Brain*. Cambridge, MA: MIT Press.

Anderson, M. L., & Chemero, T. (2013). The problem with brain GUTs: conflation of different senses of "prediction" threatens metaphysical disaster. *Behavioral and Brain Sciences*, *36*(3), 204-205.

Anderson, M. L., Kinnison, J., Pessoa, L. (2013). Describing functional diversity of brain regions and brain networks. *NeuroImage*, Jun; 73:50-8.

Apps, M. A., & Tsakiris, M. (2014). The free-energy self: a predictive coding account of self-recognition. *Neuroscience & Biobehavioral Reviews*, *41*, 85-97.

Aristotle (1991). The art of rhetoric (Transl. H. C. Lawson-Tancred). London: Penguin.

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

Arnsten, A. F., & Li, B. M. (2005). Neurobiology of executive functions: Catecholamine influences on prefrontal cortical functions. *Biological Psychiatry*, *57*(11), 1377–1384.

Arpaly, N., & Schroeder, T. (2013). *In praise of desire*. Oxford University Press.

Ay, N., Bertschinger, N., Der, R., Güttler, F., & Olbrich, E. (2008). Predictive information and explorative behavior of autonomous robots. *The European Physical Journal B-Condensed Matter and Complex Systems*, *63*(3), 329–339.

Balderston, N. L., Schultz, D. H., and Helmstetter, F. J. (2011). The human amygdala plays a stimulus specific role in the detection of novelty. *Neuroimage* 55, 1889–1898.

Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. *Trends in cognitive sciences*, *11*(7), 280-289.

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., et al., (2006). Top-down facilitation of visual recognition. *Proc. Natl. Acad. Sci.* 103 (2), 449–454.

Baranes, A., & Oudeyer, P.-Y. (2013). Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robotics and Autonomous Systems*, *61*(1), 49–73.

Barbas, H., & Rempel-Clower, N. (1997). Cortical structure predicts the pattern of corticocortical connections. *Cerebral cortex (New York, NY: 1991)*, *7*(7), 635-646.

Bard, P. (1928). A diencephalic mechanism for the expression of rage with special reference to the sympathetic nervous system. *American Journal of Physiology*, *84*, 490–515.

Barrett, L. F. (2006a). Are emotions natural kinds?. *Perspectives on psychological science*, *1*(1), 28-58.

Barrett, L. F. (2006b). Solving the emotion paradox: Categorization and the experience of emotion. *Personality and social psychology review*, *10*(1), 20-46.

Barrett, L. (2011). *Beyond the Brain: How Body and Environment Shape Animal and Human Minds*. Princeton, NJ: Princeton University Press.

Barrett, L. F. (2013). Psychological Construction: The Darwinian Approach to the Science of Emotion. *Emotion Review*. Vol. 5 No. 4: 379-389.

Barrett, L. F. (2017). *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt.

Barrett, L. F., Bar, M. (2009). See it with feeling: Affective predictions during object perception. Theme issue: Predictions in the brain: Using our past to generate a future (M. Bar Ed.) *Philosophical Transactions of the Royal Society B*, 364: 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., & Satpute, A. B. (2013). Large-scale brain networks in affective and social neuroscience: towards an integrative functional architecture of the brain. *Current opinion in neurobiology*, *23*(3), 361-372.

Barrett, L. F., Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature Reviews Neuroscience,* 16, 419–429.

Barsalou, L. (2003). Situated simulation in the human conceptual system. *Language and cognitive processes*, *18*(5-6), 513-562.

Barsalou, L. W. (2009). Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society of London: Biological Sciences*, 364, 1281–1289.

Barton, R. A. (2012). Embodied cognitive evolution and the cerebellum. *Phil. Trans. R. Soc. B*, *367*(1599), 2097-2107.

Barton, R. A., and Harvey, P. H. (2000). Mosaic evolution of brain structure in mammals. *Nature* 405, 1055–1058.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*, 695–711.

Beer, R. D. (2003). The dynamics of active categorical perception in an evolved model agent. *Adaptive Behavior*, *11*(4), 209-243.

Belin, D., Jonkman, S., Dickinson, A., Robbins, T. W., & Everitt, B. J. (2009). Parallel and interactive learning processes within the basal ganglia: relevance for the understanding of addiction. *Behavioural brain research*, *199*(1), 89-102.

Berlyne, D. E. (1966). Curiosity and exploration. *Science*, *153*(3731), 25–33. Jul 1.

Bialek, W., Nemenman, I., & Tishby, N. (2001). Predictability, complexity, and learning. *Neural Computation*, *13*(11), 2409–2463.

Blackford, J. U., Avery, S. N., Cowan, R. L., Shelton, R. C., & Zald, D. H. (2011). Sustained amygdala response to both novel and newly familiar faces characterizes inhibited temperament. *Social cognitive and affective neuroscience*, *6*(5), 621-629.

Bressler, S. L., & Menon, V. (2010). Large-scale brain networks in cognition: emerging methods and principles. *Trends in cognitive sciences*, *14*(6), 277-290.

Brooks, R. A. (1991). Intelligence without reason. *Artificial intelligence: critical concepts*, *3*, 107-63.

Brosch, T., & Sander, D. (2013). Comment: the appraising brain: towards a neuro-cognitive model of appraisal processes in emotion. *Emotion Review*, *5*(2), 163-168.

Bruineberg, J., Kiverstein, J., & Rietveld, E. (2018). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese*, *195*(6), 2417-2444.

Bruineberg, J., & Rietveld, E. (2014). Self-organization, free energy minimization, and optimal grip on a field of affordances. *Frontiers in human neuroscience*, *8*, 599.

Bubic, A., von Cramon, D. Y., Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in Human Neuroscience,* 4(25):1–15.

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. The brain's default network: anatomy, function, and relevance to disease Ann NY Acad Sci 2008; 1124: 1-38.

Byrge, L., Sporns, O., & Smith, L. B. (2014). Developmental process emerges from extended brain–body–behavior networks. *Trends in cognitive sciences*, *18*(8), 395-403.
Cannon, W. B. (1929). *Bodily changes in pain, hunger, fear and rage* (2nd ed.). New York: Appleton.
Castiello, U. (1999). Mechanisms of selection for the control of hand action. *Trends in Cognitive Sciences*, *3*(7), 264–271.

Chanes, L., & Barrett, L. F. (2016). Refining the role of limbic areas in cortical processing. *Trends in Cognitive Sciences*, *20*(2), 96–106.

Chareyron, L. J., Banta Lavenex, P., Amaral, D. G., & Lavenex, P. (2011). Stereological analysis of the rat and monkey amygdala. *Journal of Comparative Neurology*, *519*(16), 3218-3239.

Chemero, A. (2009). *Radical Embodied Cognitive Science*. Cambridge, MA: MIT Press.

Chareyron, L. J., Banta, Lavenex P., Amaral, D. G., & Lavenex, P. (2011). Stereological analysis of the rat and monkey amygdala. Journal of Comparative Neurology, 519, 3218–3239

Churchland, P. S., Ramachandran, V., et al. (1994). A critique of pure vision. In C. Koch & J. Davis (Eds.), Large-scale neuronal theories of the brain. Cambridge, MA: MIT Press.

Cisek, P. (2007). Cortical mechanisms of action selection: The affordance competition hypothesis. Philosophical Transactions of the Royal Society B: Biological Sciences, 362(1485), 1585–1599.

Cisek, P., & Kalaska, J. F. (2010). Neural mechanisms for interacting with a world full of action choices. Annual Review of Neuroscience, 2010(33), 269–298.

Cisek, P., & Pastor-Bernier, A. (2014). On the challenges and mechanisms of embodied decisions. *Phil. Trans. R. Soc. B*, *369*(1655), 20130479.

Clark, A. (1997). *Being there: Putting brain, body and world together again*. Cambridge, MA: MIT Press.

Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. OUP USA.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and Brain Sciences*, *36*(03), 181–204. doi:10.1017 S0140525X12000477.

Clark, A. (2014). *Mindware: An introduction to the philosophy of cognitive science* (2nd ed.). Oxford, NY: Oxford University Press.

Clark, A. (2015). Radical predictive processing. *The Southern Journal of Philosophy*, *53*(S1), 3-27.

Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

Clark, A. (2017a). Busting out: Predictive brains, embodied minds, and the puzzle of the evidentiary veil. *Noûs*, *51*(4), 727-753.

Clark, A. (2017b). A nice surprise? Predictive processing and the active pursuit of novelty. *Phenomenology and the Cognitive Sciences*, 1-14.

Clarke, H. F., et al. (2008). Lesions of the medial striatum in monkeys produce perseverative impairments during reversal learning similar to those produced by lesions of the orbitofrontal

cortex. *Journal of Neuroscience*, *28*, 10972–10982.

Cole, J. (1999). On 'Being Faceless': Selfhood and facial embodiment. In S. Gallagher & J. Shear (Eds.), Models of the self (pp. 301–318). Charlottesville: Imprint Academic.

Cole, J. (2010). Agency with impairments of movement. In D. Schmicking & S. Gallagher (Eds.), Handbook of phenomenology and cognitive science (pp. 655–670). Dordrecht: Springer.

Cole, J., & Spalding, H. (2009). *The invisible smile: Living without facial expression*. Oxford: Oxford University Press

Colombetti, G. (2005). Appraising valence. *Journal of Consciousness Studies*, *12*(8–9), 103–126.

Colombetti, G. (2007). Enactive appraisal. *Phenomenology and the Cognitive Sciences* 6:527–546.

Colombetti, G. (2014). *The Feeling Body: Affective Science Meets the Enactive Mind.* Cambridge, MA: 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 Perspective on Embodiment and Consciousness*. New Jersey: Lawrence Erlbaum Associates.

Colombo, M. (2013). Moving forward (and beyond) the modularity debate: A network perspective. *Philosophy of Science*, *80*(3), 356-377.

Corbetta M., Patel G., Shulman G.L. (2008). The reorienting system of the human brain: from environment to theory of mind. *Neuron* 58:306–324.

Craig, A. D. (2002). How do you feel? Interoception: the sense of the physiological condition of the body. *Nature reviews neuroscience*, *3*(8), 655.

Craig, A. D. (2003). Interoception: the sense of the physiological condition of the body. *Current opinion in neurobiology*, *13*(4), 500-505.

Craig, A.D. (2009). How do you feel – now? The anterior insula and human awareness. *Nat. Rev. Neurosci.*, 10, 59–70.

Craig, A.D. (2010). "The insular cortex and subjective awareness," in The Study of Anosognosia, ed G. P. Prigatano (New York, NY: Oxford University Press), 63–88.

Critchley, H. D. (2005). Neural mechanisms of autonomic, affective and cognitive integration. *The Journal of Comparative Neurology*, 393, 154–166.

Critchley, H. D., Harrison, N. A. (2013). Visceral influences on brain and behavior. *Neuron,* 77, 624–638.

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain.* New York: Putnam.

Damasio, A. R. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. New York: Harcourt Brace.

Damasio, A. R. (2010). *Self Comes to Mind: Constructing the Conscious Brain*. London: Heinemann.

Dayan P., Hinton G. E., Neal R. M., Zemel R. S. (1995). The Helmholtz Machine. *Neural Comput* 7:889–904.

Davis, W. A. (1984). The two senses of desire. *Philosophical Studies*, *45*(2), 181-195.

Dehaene, S., Charles, L., King, J.-R., & Marti, S. (2014). Toward a computational theory of conscious processing. *Current Opinion in Neurobiology*, *25*, 76–84.

Delgado, M. R., Miller, M. M., Inati, S., & Phelps, E. A. (2005). An fMRI study of reward-related probability learning. *Neuroimage*, *24*(3), 862-873.

Dember, W. N., Earl, R. W., & Paradise, N. (1957). Response by rats to differential stimulus complexity. *Journal of Comparative and Physiological Psychology*, *50*(5), 514–518.

Diano, M., Celeghin, A., Bagnis, A., & Tamietto, M. (2017). Amygdala response to emotional stimuli without awareness: facts and interpretations. *Frontiers in psychology*, *7*, 2029.

Dreyfus, H. L. (1991). *Being-in-the-world: A commentary on Heidegger's Being and Time, Division I*. Mit Press.

Dreyfus H, Kelly SD (2007) Heterophenomenology: heavy-handed sleight-of-hand. Phenomenology Cognitive Science 6(1–2):45–55

Duclos, S. E., Laird, J. D. (2001). The deliberate control of emotional experience through control of expressions. *Cognition and Emotion*, 15, 27–56.

Duclos, S. E., Laird, J. D., Schneider, E., Sexter, M., Stern, L., Van Lighten, O. (1989). Emotion-specific effects of facial expressions and postures on emotional experience. *Journal of Personality and Social Psychology*, 57, 100–108.

Dunnett, S. B., et al. (1991). The basal forebrain-cortical cholinergic system: Interpreting the functional consequences of excitotoxic lesions. *Trends in Neurosciences*, *14*, 494–501.

Edelman, B. (1984). A multiple-factor of body weight control. *Journal of General Psychology*, 110, 99–114.

Edelman, G. M. (1998). *The remembered present: A biological theory of consciousness*. New York: Basic Books.

Edelman, G. M., & Gally, J. A. (2001). Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences*, *98*(24), 13763-13768.

Ekman, P. (1999). "Basic emotions," in *Handbook of Cognition and Emotion*, eds T. Dagleish and M. Power (Chichester: John Wiley & Sons Ltd), 45–60.

Ekman, P., Davidson, R. J., & Friesen, W. V. (1990). The Duchenne smile: Emotional expression and brain physiology: II. *Journal of personality and social psychology*, *58*(2), 342.

Ekman, P., & Friesen, W. V. (1982). Felt, false, and miserable smiles. *Journal of nonverbal behavior*, *6*(4), 238-252.

Ekman, P., Levenson, R. W., & Friesen, W. V. (1983). Autonomic nervous system activity distinguishes among emotions. *Science*, 221(4616), 1208-1210.

Everitt, B. J., & Robbins, T. W. (2013). From the ventral to the dorsal striatum: devolving views of their roles in drug addiction. *Neuroscience & Biobehavioral Reviews*, *37*(9), 1946-1954.

Fazelpour, S., & Thompson, E. (2015). The Kantian brain: brain dynamics from a neurophenomenological perspective. *Current opinion in neurobiology*, *31*, 223-229.

Feldman, J. (2013). Tuning your priors to the world. *Topics in Cognitive Science*, *5*(1), 13–34.

Figdor, C. (2010). Neuroscience and the multiple realization of cognitive functions. *Philos. Sci.* 77, 419–456.

FitzGerald, T. H., Dolan, R. J., & Friston, K. J. (2014). Model averaging, optimal inference, and habit formation. *Frontiers in human neuroscience*, *8*, 457.

Flack, W. F., Jr., Laird, J. D., Cavallaro, L. A. (1999). Additive effects of facial expressions and postures on emotional feelings. *European Journal of Social Psychology*, 29, 203–217.

Flanagan, O. (2017). Addiction doesn't exist, but it is bad for you. *Neuroethics*, *10*(1), 91-98.

Frijda, N. H. (1986). *The Emotions*. Cambridge: Cambridge University Press.

Frijda, N. H. (2004). Emotions and action. In A. S. R. Manstead, N. Frijda, & A. Fischer (Eds.), *Feelings and emotions* (pp. 158–173). Cambridge: Cambridge University Press.

Frijda, N. H. (2007a). *The laws of emotion*. Mahwah, New Jersey: Lawrence Erlbaum Associate Publishers.

Frijda, N. H. (2007b). What might emotions be? Comments on the Comments. *Social Science Information*, *46*(3), 433-443.

Friston, K. (2002). Beyond phrenology: what can neuroimaging tell us about distributed circuitry?. *Annual review of neuroscience*, *25*(1), 221-250.

Friston K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society London B: Biological Sciences*, 360(1456), 815–836.

Friston, K. J. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301.

Friston, K. J. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Friston, K. (2011). What is optimal about motor control?. *Neuron*, *72*(3), 488-498.
Friston, K. (2011b). Embodied inference: Or I think therefore I am, if I am what I think. *The implications of embodiment (Cognition and Communication)*, 89-125.

Friston, K. (2013). Life as we know it. *Journal of the Royal Society Interface*, *10*(86), 20130475.

Friston, K., Adams, R. A., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology*, *3*, 151.

Friston, K., Breakspear, M., & Deco, G. (2012). Perception and self-organized instability. *Frontiers in computational neuroscience*, *6*, 44.

Friston, K. J., Daunizeau, J., & Kiebel, S. J. (2009). Reinforcement learning or active inference?. *PloS one*, *4*(7), e6421.

Friston, K., Daunizeau, J., Kilner, J., & Kiebel, S. J. (2010). Action and behavior: A free-energy formulation. *Biological Cybernetics*, *102*(3), 227–260.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1211-1221.

Friston, K. J., Fitzgerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: A process theory. *Neural Computation*, *29*(1), 1–49.

Friston, K., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. *Biological cybernetics*, *104*(1-2), 137-160.

Friston, K. J., Parr, T., & de Vries, B. (2017). The graphical brain: belief propagation and active inference. *Network Neuroscience*, *1*(4), 381-414.

Friston, K. J., & Price, C. J. (2003). Degeneracy and redundancy in cognitive anatomy. *Trends Cogn. Sci.* 7, 151–152.

Friston, K. J., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience*, *6*(4), 187–214.

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2014). The anatomy of choice: dopamine and decision-making. *Phil. Trans. R. Soc. B*, *369*(1655), 20130481.

Friston, K. J., Shiner, T., FitzGerald, T., Galea, J. M., Adams, R., Brown, H., et al. (2012a). Dopamine, affordance and active inference. *PLoS Computational Biology*, *8*(1).

Friston, K. J., & Stephan, K. E. (2007). Free-energy and the brain. *Synthese*, *159*(3), 417–458.

Friston, K., Thornton, C., & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in psychology*, *3*, 130.

Fuchs, T., Koch, S. (2014) Embodied affectivity: on moving and being moved. Frontiers in Psychology. *Psychology for Clinical Settings*, Article 508, p.1-12.

Fuster, J. M. (2004). Upper processing stages of the perception-action cycle. *Trends in Cognitive Sciences*, *8*(4), 143–145.

Gallagher, S., & Allen, M. (2018). Active inference, enactivism and the hermeneutics of social cognition. *Synthese*, *195*(6), 2627-2648.

Gallese, V. (2005). Embodied simulation: from neurons to phenomenal experience. *Phenomenology and the Cognitive Sciences*, 4, 23–48.

Garavan, H. (2010). Insula and drug cravings. *Brain Struct. Funct.,* 214, 593–601.

Garland, E., Froeliger, B., & Howard, M. (2014). Mindfulness training targets neurocognitive mechanisms of addiction at the attention-appraisal-emotion interface. *Frontiers in psychiatry*, *4*, 173.

Gibson, J. J. (1979). The theory of affordances The Ecological Approach to Visual Perception (pp. 127-143).

Gottlieb, J., Oudeyer, P. Y., Lopes, M., & Baranes, A. (2013). Information seeking, curiosity and attention: Computational and neural mechanisms. *Trends in Cognitive Science*, *17*(11), 585–596.

Granfield, R., & Cloud, W. (1999). *Coming clean: Overcoming addiction without treatment*. NYU Press.

Gray, M.A., Harrison, N.A., Wiens, S., and Critchley, H.D. (2007). Modulation of emotional appraisal by false physiological feedback during fMRI. PLoS ONE 2, e546.

Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society of London* B: Biological Sciences, 290(1038), 181–197.

Gu, X., Liu, X., Van Dam, N. T., Hof, P. R., & Fan, J. (2013). Cognition–emotion integration in the anterior insular cortex. *Cerebral cortex*, *23*(1), 20-27.

Gu, X., Hof, P. R., Friston, K. J., & Fan, J. (2013). Anterior insular cortex and emotional awareness. *Journal of Comparative Neurology*, *521*(15), 3371-3388.
Gugliotta, G. (2003, August 21). Revolutionary Thinker. *The Washington Post*.

Guillery, R. W. (2003). Branching thalamic afferents link action and perception. *Journal of Neurophysiology*, *90*, 539–548.

Guillery, R. W. (2005). Anatomical pathways that link perception and action. *Progress in Brain Research*, *149*, 235–256.

Haken, H. (1983). Synergetics: An introduction. Non-equilibrium phase transition and self-organisation in physics, chemistry and biology (3rd ed.). Berlin: Springer.

Hari, J. (2015). *Chasing the scream: The first and last days of the war on drugs*. Bloomsbury Publishing USA.

Harrison, N. A., Gray, M. A., Gianaros, P. J., Critchley, H. D. (2010). The Embodiment of Emotional Feelings in the Brain. *J. Neurosci.* 30, 12878–12884.

Heather, N., Best, D., Kawalek, A., Field, M., Lewis, M., Rotgers, F., ... & Heim, D. (2017). Challenging the brain disease model of addiction: European launch of the addiction theory network.

Helmholtz, H. (1860/1962). *Handbuch der physiologischen optik* (J. P. C. Southall, Ed., English trans.), Vol. 3. New York: Dover.

Herrick, C. J. (1933). The functions of the olfactory parts of the cerebral cortex. *Proceedings of the National Academy of Sciences, USA*, *19*, 7–14.

Herry, C., Bach, D. R., Esposito, F., Di Salle, F., Perrig, W. J., Scheffler, K., et al. (2007). Processing of temporal unpredictability in human and animal amygdala. *J. Neurosci.* 27, 5958–5966.

Hilgetag, C. C., Burns, G. A., O'Neill, M. A., Scannell, J. W., & Young, M. P. (2000). Anatomical connectivity defines the organization of clusters of cortical areas in the macaque and the cat. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *355*(1393), 91-110.

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences*, *11*, 428–434. Hinton, G. E. (2010). Learning to represent visual input. *Philosophical Transactions of the Royal Society B*, *365*, 177–184.

Hohwy, J. (2010). The hypothesis testing brain: some philosophical applications.

Hohwy, J. (2013). *The predictive mind*. Oxford, NY: Oxford University Press.

Hohwy, J. (2015). The neural organ explains the mind. In T. Metzinger & J. M. Windt (Eds.), Open MIND: 19(T). Frankfurt am Main: MIND Group.

Hohwy, J. (2016). The self−evidencing brain. *Noûs*, *50*(2), 259-285.

Hohwy, J., Roepstorff, A., Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition* 108(3):687–701.

Holland P. C., Gallagher, M. (1999). Amygdala circuitry in attentional and representational processes. *Trends in Cognitive Sciences* 3: 65-73.

Holton, R. (2009). *Willing, wanting, waiting*. Oxford University Press.

Hoshi E, Tanji J. (2007). Distinctions between dorsal and ventral premotor areas: Anatomical connectivity and functional properties. *Current Opinion in Neurobiology, 17*(2), 234–242.

Humphreys, G. W., & Riddoch, J. M. (2000). One more cup of coffee for the road: Object-action assemblies, response blocking and response capture after frontal lobe damage. *Experimental Brain Research*, *133*, 81–93.

Hurley, S. L. (1998). Vehicles, contents, conceptual structure, and externalism. *Analysis*, *58*(1), 1-6.

Hurley, S. (2008). The shared circuits model (SCM): How control, mirroring, and simulation can enable imitation, deliberation, and mindreading. *Behavioral and Brain Sciences*, *31*(1), 1-22.

Hurley, S. L. (2010). "Varieties of externalism," in *The Extended Mind* ed R. Menary (Cambridge: MIT), 101–154.

Hutto, D., and Myin, E. (2013). Radicalising Enactivism. Basic Minds Without Content. Cambridge, MA: MIT Press.

Izard, C. E. (2007). Basic emotions, natural kinds, emotion schemas and a new paradigm. Perspect. Psychol. Sci. 2, 260–280.

Izard, C. E. (2011). Form and functions of emotions: matters of emotion- cognition interactions. *Emot. Rev.* 3, 371–378.

Jabbi, M., Bastiaansen, J., and Keysers, C. (2008). A common anterior insula representation of disgust observation, experience and imagination shows divergent functional connectivity pathways. *PLoS ONE 3*, e2939.

Jackson, J. H. (1884). The Coronian Lecture on evolution and dissolution of the nervous system. *Br. Med. J.* 1, 660–663.

James, W. (1884). What is an emotion? *Mind*, 9, 188–205.

Joels, M., Pu, Z., Wiegert, O., Oitzl, M. S., & Krugers, H. J. (2006). Learning under stress: How does it work? *Trends in Cognitive Sciences*, *10*(4), 152–158.

Joffily, M., & Coricelli, G. (2013). Emotional valence and the free-energy principle. *PLoS Computational Biology*, *9*(6), e1003094.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*, 697–720.

Kanai, R., Komura, Y., Shipp, S., & Friston, K. (2015). Cerebral hierarchies: Predictive processing, precision and the pulvinar. Philosophical Transactions of the Royal Society of London. Series B Biological. *Science*, *370*(1668), 11–16.

Kaplan, F., & Oudeyer, P. Y. (2007). In search of the neural circuits of intrinsic motivation. *Frontiers in Neuroscience*, *1*(1), 225–236.

Kaplan, F., & Oudeyer, P.-Y. (2011). From hardware and software to kernels and envelopes: A concept shift for robotics, developmental psychology, and brain sciences. In J. L. Krichmar & H. Wagatsuma (Eds.), *Neuromorphic and brain-based robots* (pp. 217–250). Cambridge: Cambridge University Press.

Kauffman, S. A. (1993). *The Origins of Order: Self-Organization and Selection in Evolution*. New York: Oxford University Press.

Kelso, J. A. (1995). Scott (1995): *Dynamic patterns: The self-organization of brain and behavior*. Chicago

Kelso, J.A.S. (2012). Multistability and metastability: understanding dynamic coordination in the brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1591), 906-18.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLoS ONE*, *7*(5), e36399.

Kipps, C. M., Duggins, A. J., McCusker, E. A., & Calder, A. J. (2007). Disgust and happiness recognition correlate with anteroventral insula and amygdala volume respectively in preclinical Huntington's disease. *Journal of cognitive neuroscience*, *19*(7), 1206-1217.

Kirchhoff, M. D. (2016). Autopoiesis, free energy, and the life–mind continuity thesis. *Synthese*, 1-22.

Kirchhoff, M. (2018). The body in action: predictive processing and the embodiment thesis. Oxford handbook of cognition: embodied, extended and enactive. Oxford University Press, Oxford (in press) Google Scholar.

Kirchhoff, M., & Froese, T. (2017). Where there is life there is mind: In support of a strong life-mind continuity thesis. *Entropy*, *19*(4), 169.

Kiverstein, J., & Miller, M. (2015). The embodied brain: towards a radical embodied cognitive neuroscience. *Frontiers in Human Neuroscience*, *9*, 237.

Kiverstein, J., & Rietveld, E. (2015). The primacy of skilled intentionality: on Hutto & Satne's the natural origins of content. *Philosophia*, *43*(3), 701-721.

Klaasen, P., Rietveld, E., & Topal, J. (2010). Inviting complementary perspectives on situated normativity in everyday life. *Phenomenology and Cognitive Sciences*, *9*(1), 53–73.

Kober, H., Barrett, L. F., Joseph, J., Bliss-Moreau, E., Lindquist, K., and Wager, T. D. (2008). Functional grouping and cortical-subcortical interactions in emotion: a meta-analysis of neuroimaging studies. *Neuroimage* 42, 998–1031.

Koffka, K. (1935). Principles of Gestalt Psychology, International Library of Psychology, Philosophy and Scientific Method.

Koob, G. F., & Le Moal, M. (2001). Drug addiction, dysregulation of reward, and allostasis. *Neuropsychopharmacology*, *24*(2), 97.

Koob, G. F., & Le Moal, M. (2005). Plasticity of reward neurocircuitry and the'dark side'of drug addiction. *Nature neuroscience*, *8*(11), 1442.

Kwisthout, J., Bekkering, H., & Van Rooij, I. (2017). To be precise, the details don't matter. On predictive processing, precision and level of detail of predictions. *Brain and Cognition*, *112*, 84–91.

Laird, J. D. (2007). *Feelings: The Perception of Self*. Oxford: Oxford University Press.

Lazarus, R. S. (1966). *Psychological stress and the coping process*. New York: McGraw-Hill.

LeDoux, J. E. (1987). Emotion. In F. Plum (Ed.), Handbook of physiology. 1: The nervous system. Higher *functions of the brain* (Vol. V, pp. 419–460). Bethesda: American Physiological Society.

LeDoux, J. E. (1996). *The emotional brain*. New York: Simon and Schuster.

LeDoux, J. E. (2012). Evolution of human emotion: a view through fear. *Prog. Brain Res.* 195, 431–442.

Leiner, H. C., et al. (1986). Does the cerebellum contribute to mental skills? *Behavioral Neuroscience*, *100*, 443–454.

Lepora, N. F., & Pezzulo, G. (2015). Embodied choice: how action influences perceptual decision making. *PLoS computational biology*, *11*(4), e1004110.

Levy, N. (2013). Addiction is not a brain disease (and it matters). *Frontiers in Psychiatry*, *4*, 24.

Lewis, M. (2005). Bridging emotion theory and neurobiology through dynamic systems modelling. *Behavioral and Brain Sciences*, 28:169-245.

Lewis, M. (2015). *The biology of desire: why addiction is not a disease*. Hachette UK.

Lewis, M. (2017). Addiction and the brain: development, not disease. *Neuroethics*, *10*(1), 7-18.

Lewis, M. & Shelly, S. (under review). The neurobiology of addiction reveals learning, not disease.

Lewis, M., & Todd, R. (2007). The self-regulating brain: Cortical-subcortical feedback and the development of intelligent action. *Cognitive Development*, *22*(4), 406-430.

Lewis, M. (forthcoming). The neurobiology of addiction reveals learning, not disease

Lewis, M. D., & Todd, R. M. (2005). Getting emotional: A neural perspective on emotion, intention, and consciousness. *Journal of Consciousness Studies,* 12(8–10), 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 (2007) 406–430.

Lewis, M. D., & Liu, Z. X. (2011). Three time scales of neural self-organization underlying basic and nonbasic emotions. Emotion Review, 3(4), 416-423.

Lindquist, K. A., & Barrett, L. F. (2012). A functional architecture of the human brain: emerging insights from the science of emotion. *Trends in cognitive sciences*, *16*(11), 533-540.

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., & Barrett, L. F. (2012). The brain basis of emotion: a meta-analytic review. *Behavioral and brain sciences*, *35*(3), 121-143.

Lovero, K. L., Simmons, A. N., Aron, J. L., Paulus, M. P. (2009). Anterior insular cortex anticipates impending stimulus significance. *Neuroimage,* 45, 976–983.

Lowe, R., & Ziemke, T. (2011). The feeling of action tendencies: On the emotional regulation of goal- directed behavior. *Frontiers in Psychology*, *2*, 346.

MacKay, D. (1956). The epistemological problem for automata. In C. E. Shannon & J. McCarthy (Eds.), *Automata studies* (pp. 235–251). Princeton, NJ: Princeton University Press.

MacLean, P. D. (1952). Some psychiatric implications of physiological studies on frontotemporal portion of limbic system (visceral brain). *Clinical Neurophysiology*, *4*(4), 407-418.

MacLean, P. D. (1990). The Triune Brain in Evolution: Role in Paleocerebral Functions. New York: Plenum Press.

Markovic, J., Anderson, A. K., & Todd, R. M. (2015). Tuning to the significant: Neural and genetic processes underlying affective enhancement of visual perception and memory. *Behavioural Brain Research*, *259*, 41–229.

Maté, G. (2010). In the realm of hungry ghosts: Close encounters with addiction. North Atlantic Books.

Maturana, H. R., & Varela, F. J. (1991). Autopoiesis and cognition: The realization of the living (Vol. 42). Springer Science & Business Media.

McEvoy, P. (2002). *Classic theory: The theory of interacting systems*. San Francisco: Microanalytix.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748.

Menon, V., and Uddin, L. Q. (2010). Saliency, switching, attention and control: a network model of insula function. *Brain Struct. Funct.* 214, 655–667.

Merleau-Ponty, M. (1942/1963). *The Structure of Behaviour.* Translated by A. Fisher Boston, MA: Beacon Press.

Merleau Ponty, M. (1968/2003). *Nature: Course Notes from the Collège de France.* (Evanston, IL: Northwestern University)

Mesulam, M. M. (1990). Large-scale neurocognitive networks and distributed processing for attention, language, and memory. *Annals of Neurology*, 28, 597–613.

Mather, M., Clewett, D., Sakaki, M., & Harley, C. W. (2015). Norepinephrine ignites local hot spots of neuronal excitation: How arousal amplifies selectivity in perception and memory. *Behavioral and Brain Sciences*, *1*, 1–100.

Maturana, H. R., & Varela, F. (1980). Autopoiesis: The organization of the living. In F. V & H. R. Maturana (Eds.), Autopoiesis and cognition. Dordrecht: Reidel.

Montague, P.R., Dayan, P., & Sejnowski, T.J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of Neuroscience*, 16(5),1936-1947.

Moriguchi, Y., Negreira, A., Weierich, M., Dautoff, R., Dickerson, B. C., Wright, C. I., & Barrett, L. F. (2011). Differential hemodynamic response in affective circuitry with aging: an FMRI study of novelty, valence, and arousal. *Journal of cognitive neuroscience*, *23*(5), 1027-1041.

Moulin-Frier, C., & Oudeyer, P. Y. (2012). Curiosity-driven phonetic learning. In *2012 IEEE international conference on development and learning and epigenetic robotics (ICDL)* (pp. 1-8). IEEE.

Miller, M., & Clark, A. (2017). Happily entangled: prediction, emotion, and the embodied mind. *Synthese*, *195*(6), 2559-2575.

Mitchell I. J, Heims H, Neville E. A, Rickards H. (2005). Huntington's disease patients show impaired perception of disgust in the gustatory and olfactory modalities. *J Neuropsychiatry Clin Neurosci.* 17:119-21.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of neuroscience*, *16*(5), 1936-1947.

Morillo, C. R. (1990). The reward event and motivation. *The Journal of Philosophy*, *87*(4), 169-186.

Moutoussis, M., Story, G. W., & Dolan, R. J. (2015). The computational psychiatry of reward: broken brains or misguided minds?. *Frontiers in psychology*, *6*, 1445.

Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts.

Ngo, H., Luciw, M., Forster, A., & Schmidhuber, J. (2012, June). Learning skills from play: artificial curiosity on a katana robot arm. In *Neural Networks (IJCNN), The 2012 International Joint Conference on* (pp. 1-8). IEEE.

Nguyen, M., & Oudeyer, P.-Y. (2013). Active choice of teachers, learning strategies and goals for a socially guided intrinsic motivation learner. *Paladyn Journal of Behavioural Robotics*, *3*(3), 136–146.

Niedenthal, P. M. (2007). Embodying emotion. *Science* 316, 1002–1005.

Niedenthal, P. M., Barsalou, L. W., Winkielman, P., Krauth-Gruber, S., & Ric, F. (2005). Embodiment in attitudes, social perception, and emotion. *Personality and social psychology review*, *9*(3), 184-211.

Niedenthal, P. M., Mermillod, M., Maringer, M., & Hess, U. (2010). The Simulation of Smiles (SIMS) model: Embodied simulation and the meaning of facial expression. *Behavioral and brain sciences*, *33*(6), 417-433.

Noe¨, A. (2004). Action in perception. MIT press.

Noë, A. (2009). Out of Our Heads: Why You Are Not Your Brain and Other Lessons from the Biology of Consciousness. New York: Hill and Wang.

Noe, A. (2011, September 9). Addiction is not a disease of the brain [Blog post]. Retrieved from https://www.npr.org/sections/13.7/2011/09/09/140307282/addiction-is-not-a-disease-of-the-brain

Öhman, A., and Mineka, S. (2001). Fear, phobias and preparedness: toward an evolved module of fear and fear learning. *Psychol. Rev.* 108, 483–522.

Oosterwijk, S., Lindquist K. A., Anderson, C., Dautoff, R. Moriguchi, Y. Barrett, L. F. (2012). States of mind: Emotions, body feelings, and thoughts share distributed neural networks. *NeuroImage,* 62: 2110–2128.

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and brain sciences*, *24*(5), 939-973.

O'Regan, J. K., Myin, E., & Noë, A. (2004). Towards an analytic phenomenology: the concepts of "bodiliness" and "grabbiness". In *Seeing, thinking and knowing* (pp. 103-114). Springer, Dordrecht.

Oudeyer, P. Y. (2014). Socially guided intrinsic motivation for robot learning of motor skills. *Autonomous Robots*, *36*(3), 273–294.

Oudeyer, P.-Y., & Kaplan, F. (2006). Discovering communication. *Connection Science*, *18*(2), 189–206.

Oudeyer, P.-Y., Kaplan, F., & Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. IEEE Transactions on Evolutionary Computation, 11(2), 265–286.

Oudeyer, P. Y., Baranes, A., & Kaplan, F. (2013). Intrinsically motivated learning of real-world sensorimotor skills with developmental constraints. In *Intrinsically motivated learning in natural and*

artificial systems (pp. 303-365). Springer, Berlin, Heidelberg.

Palmer, S., and Kimchi, R. (1986). "The information-processing approach to cognition," in *Approaches to Cognition: Contrasts and Controversies*, eds T. J. Knapp and L. C. Robertson (Hillsdale, NJ: Erlbaum), 37–77.

Pape, L., Controzzi, C. M. O. M., Cipriani, C., Foerster, A., Carrozza, M. C., & Schmidhuber, J. (2012). Learning tactile skills through curious exploration. *Frontiers in Neurorobotics*, *6*(6), 11–16

Papez, J. (1937). A proposed mechanism of emotion. *Archives of Neurology and Psychiatry*, *79*, 217–224.

Patočka, J. (1998). Body, community, language, world. Open Court Publishing.

Paulus, M.P., Stein, M.B. (2006). An insular view of anxiety. *Biological Psychiatry*, 60(4), 383–387.

Panksepp, J. (1998). Affective Neuroscience: The Foundations of Human and Animal Emotions. New York: Oxford University Press.

Panksepp, J. (2004). Affective Neuroscience: The Foundations of Human and Animal Emotions. OUP USA.

Panksepp, J. (2012). "In defence of multiple core affects", in Categorial Versus Dimensional Models of Affect: A Seminar on the Theories of Panksepp and Russell, eds P. Zachar and R. D. Ellis (Amsterdam, NL: John Benjamins Publishing Company), 31–78.

Parvizi, J. (2009). Corticocentric myopia: old bias in new cognitive sciences. *Trends Cogn. Sci.* 13, 354–359.

Parvizi, J., & Damasio, A. R. (2000). Consciousness and the brainstem. *Cognition*, *79*, 135–160.

Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature Reviews Neuroscience*, 9(2), 148–158.

Pessoa, L. (2009). How do emotion and motivation direct executive control?. *Trends in cognitive sciences*, *13*(4), 160-166.

Pessoa, L. (2013). *The cognitive emotional brain: From interactions to integration*. Cambridge: MIT Press.

Pessoa, L. (2014). Understanding brain networks and brain organization. *Phys. Life Rev.* 11, 400–435.

Pessoa, L. (2015). Précis on the cognitive-emotional brain. *Behavioral and Brain Sciences*, 38, e71.

Pessoa, L. (2017). A network model of the emotional brain. *Trends in cognitive sciences*, *21*(5), 357-371.

Pessoa, L., & Adolphs, R. (2010). Emotion processing and the amygdala: From a "low road" to "many roads" of evaluating biological significance. *Nature Reviews Neurosciences*, *11*(11), 773–783.

Petro, L. S., & Muckli, L. (2016). The brain's predictive prowess revealed in primary visual cortex. *Proceedings of the National Academy of Sciences, 113*(5) , 1124–1125.

Petro, L. S., Vizioli, L., & Muckli, L. (2014). Contributions of cortical feedback to sensory processing in primary visual cortex. *Frontiers in Psychology*, *5*, 1223.

Pezzulo, G., Rigoli, F., & Friston, K. (2015). Active Inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology*, *134*, 17–35.

Pfeifer, R., & Bongard, J. (2006). How the body shapes the way we think: a new view of intelligence. MIT press.

Phelps, E. A. (2006). Emotion and cognition: Insights from studies of the human amygdala. *Annual Review of Psychology*, *57*, 27–53.

Plato (1992). Republic (Transl. G. Grube & C. Reeve). Indianapolis, IN: Hackett Publishing.

Polani, D. (2009). Information: currency of life?. *HFSP journal*, *3*(5), 307-316.

Posner, M. I., Petersen, S. E., Fox, P. T., and Raichle, M. E. (1988). Localisation of cognitive operations in the human brain. *Science* 240, 1627–1631.

Powell, K. D., & Goldberg, M. E. (2000). Response of neurons in the lateral intraparietal area to a distractor flashed during the delay period of a memory-guided saccade. *Journal of Neurophysiology*, *84*(1), 301–310.

Ploghaus, A., Tracey, I., Gati, J. S., Clare, S., Menon, R. S., Matthews, P. M., Rawlins, J. N. (1999). Dissociating pain from its anticipation in the human brain. *Science,* 284, 1979–1981.

Prinz, J. (2003). Emotions Embodied. Chapter in R. Solomon (ed.) Thinking about Feeling,. New York: OUP.

Raichle, M. (2010). Two views of brain function. *Trends Cogn. Sci.* 14, 180–190.

Rakic, P. (2009). Evolution of the neocortex: perspective from developmental biology. *Nat. Rev. Neurosci.* 10, 724–735.

Ratcliffe, M. (2008). Feelings of being: Phenomenology, psychiatry and the. sense of reality. Oxford: Oxford University Press.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, *2*, 64-99.

Rietveld, E. (2008a). Situated normativity: the normative aspect of embodied cognition in unreflective action. *Mind* 117, 973–1001.

Rietveld, E. (2008b). The skillful body as a concernful system of possible actions: Phenomena and neuro-dynamics. *Theory & Psychology*, *18*(3), 341–363.

Rietveld, E., & Brouwers, A. A. (2017). Optimal grip on affordances in architectural design practices: an ethnography. *Phenomenology and the Cognitive Sciences*, *16*(3), 545-564.

Rietveld, E., Denys, D., & Van Westen, M. (2016). Ecological-enactive cognition as engaging with a field of relevant affordances: The skilled intentionality framework (SIF). *The Oxford handbook of E*, *4*.

Rietveld, E., & Kiverstein, J. (2014). A rich landscape of affordances. *Ecological Psychology*, *26*(4), 325–352.

Rilling, J. K., & Insel, T. R. (1999). The primate neocortex in comparative perspective using magnetic resonance imaging. *Journal of Human Evolution*, *37*, 191–223.

Risold, P. Y., Thompson, R. H., and Swanson, L. W. (1997). The structural organization of connections between hypothalamus and cerebral cortex. *Brain Res. Brain Res. Rev.* 24, 197–254.

Robins, L. N. (1993). Vietnam veterans' rapid recovery from heroin addiction: A fluke or normal expectation?. *Addiction*, *88*(8), 1041-1054.

Robins, L. N., Helzer, J. E., & Davis, D. H. (1975). Narcotic use in Southeast Asia and afterward. *Archives of General Psychiatry*, *32*(8), 955-961.

Robinson, T. E., & Berridge, K. C. (1993). The neural basis of drug craving: an incentive-sensitization theory of addiction. *Brain research reviews*, *18*(3), 247-291.

Robinson, T. E., & Berridge, K. C. (2000). The psychology and neurobiology of addiction: an incentive–sensitization view. *Addiction*, *95*(8s2), 91-117.

Robinson, T. E., & Berridge, K. C. (2001). Incentive−sensitization and addiction. *Addiction*, *96*(1), 103-114.

Robinson, T. E., & Berridge, K. C. (2008). The incentive sensitization theory of addiction: some current issues. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1507), 3137-3146.

Rolls, E. T. (2005). *Emotion Explained*. Oxford: Oxford University Press.

Romo, R., & Schultz, W. (1990). Dopamine neurons of the monkey midbrain: contingencies of responses to active touch during self-initiated arm movements. *Journal of neurophysiology*, *63*(3), 592-606.

Rupert, R. (2009). Cognitive Systems and the Extended Mind. Oxford: Oxford University Press.

Russell, J. A. (1980). A circumplex model of affect. *J. Pers. Soc. Psychol.* 39, 1161–1178.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychol. Rev.* 110, 145–172.

Sander, D., Grafman, J., and Zalla, T. (2003). The human amygdala: an evolved system for relevance detection. *Rev. Neurosci.* 14, 303–316.

Sarinopoulos, I., Grupe, D. W., Mackiewicz, K. L., Herrington, J. D., Lor, M., Steege, E. E., and Nitschke, J. B. (2010). Uncertainty during anticipation modulates neural responses to aversion in human insula and amygdala. *Cereb Cortex,* 20, 929–940.

Satel, S., & Lilienfeld, S. O. (2014). Addiction and the brain-disease fallacy. *Frontiers in psychiatry*, *4*, 141.

Scanlon, T. (1998). *What we owe to each other*. Harvard University Press.

Schachter, S., Singer, J. (1962). Cognitive, Social, and Physiological Determinants of Emotional State. *Psychological Review,* 69: 379–399.

Scherer, K. R. (1984). On the nature and function of emotion: A component process approach. Approaches to emotion, 2293, 317.

Scherer, K. R. (1999). Appraisal theory. *Handbook of cognition and emotion*, 637-663.

Scherer, K. R. (2001). Appraisal considered as a process of multilevel sequential checking. *Appraisal processes in emotion: Theory, methods, research*, *92*, 120.

Scherer, K. R. (2004). Feelings integrate the central representation of appraisal-driven response organization in emotion. In *Feelings and emotions: the Amsterdam Symp.* (eds Manstead A. S. R., Frijda N. H., Fischer A. H.), pp. 136–157. Cambridge, UK: Cambridge University Press

Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition and emotion*, *23*(7), 1307-1351.

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, *2*, 230–247.

Schouppe, N., Braem, S., Houwer, J. D. A., Verguts, T., & Ridderinkhof, K. R. N. (2014). No pain, no gain: Affective valence of congruency conditions changes following a successful response. *Cognitive, Affective & Behavioural Neuroscience*, *15*(1), 251–261.

Schroeder, T. (2004). *Three faces of desire*. Oxford University Press.

Schueler, G. F. (1995). *Desire: Its role in practical reason and the explanation of action*. MIT Press.

Schultz, W., Dayan, P. & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. Science, 275(5306), 1593–1599.

Schultz, W., & Romo, R. (1990). Dopamine neurons of the monkey midbrain: contingencies of responses to stimuli eliciting immediate behavioral reactions. *Journal of neurophysiology*, *63*(3), 607-624.

Schultz, W., Tremblay, L., & Hollerman, J. R. (2000). Reward processing in primate orbitofrontal cortex and basal ganglia. *Cerebral cortex*, *10*(3), 272-283.

Schwartenbeck, P., Fitzgerald, T., Dolan, R. J., & Friston, K. (2013). Exploration, novelty, surprise, and free energy minimization. *Frontiers in Psychology*, *4*, 710.

Schwartenbeck, P., FitzGerald, T. H., Mathys, C., Dolan, R., & Friston, K. (2014). The dopaminergic midbrain encodes the expected certainty about desired outcomes. *Cerebral cortex*, *25*(10), 3434-3445.

Schwartenbeck, P., FitzGerald, T. H., Mathys, C., Dolan, R., Wurst, F., Kronbichler, M., & Friston, K. (2015). Optimal inference with suboptimal models: addiction and active Bayesian inference. *Medical hypotheses*, *84*(2), 109-117.

Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., et al. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. *J. Neurosci.* 27, 2349–2356.

Semendeferi, K., Lu, A., Schenker, N., & Damasio, H. (2002). Humans and great apes share a large frontal cortex. *Nature Neuroscience*, *5*, 272–276.

Sergerie, K., Chochol, C., and Armony, J. L. (2008). The role of the amygdala in emotional processing: a quantitative meta-analysis of functional neuroimaging studies. *Neurosci. Biobehav. Rev.* 32, 811–830. doi: 10.1016/j.neubiorev.2007. 12.002 Spinoza, B. (1677/1894). *Ethics.* Translated by W. H. White London: Fisher Unwin.

Seth, A. K. (2013). Interoceptive Inference, Emotion, and the Embodied Self. *Trends in Cognitive Sciences*, *17*(11), 565-573.

Seth, A. K. (2014). Response to Gu and FitzGerald: Interoceptive inference: from decision-making to organism integrity. *Trends in cognitive sciences*, *18*(6), 270-271.

Seth, A. K. (2015). The cybernetic Bayesian brain from interoceptive inference to sensorimotor contingen- cies. In T. Metzinger & J. M. Windt (Eds.), *OpenMIND: 35(T)*. Frankfurtam Main: MIND Group.

Seth, A. K., & Friston, K. J. (2016). Active interoceptive inference and the emotional brain. Phil. Trans. R. Soc. B, 371(1708).

Seth, A. K., Suzuki, K., & Critchley, H. D. (2012). An interoceptive predictive coding model of conscious presence. *Frontiers in psychology*, *2*, 395.

Shapiro, L. (2007). The embodied cognition research programme. *Philosophy compass*, *2*(2), 338-346.

Sherman, S. M. (2007). The thalamus is more than just a relay. *Current Opinion in Neurobiology*, *17*, 417–422.

Sherman, S. M., & Guillery, R. W. (2011). Distinct functions for direct and transthalamic corticocortical connections. Journal of Neurophysiol, 106, 1068–1077.

Sherman, S. M., & Guillery, R. W. (2013). *Thalamocortical processing: Understanding the messages that link the cortex to the world*. Cambridge: MIT Press.

Sherwood, C., Bauernfein, A. L., Bianchi, S., Raghanti, M. A., & Hof, P. R. (2012). Human brain evolution writ large and small. In M. A. Hofman & D. Falk (Eds.), *Evolution of the primate brain: From neuron to behavior* (pp. 237–257). Oxford: Elsevier.

Shipp, S. (2003). The functional logic of cortico-pulvinar connections. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *358*, 1605–1624.

Singer T, Critchley H. D., Preuschoff K. (2009). A common role of insula in feelings, empathy and uncertainty. *Trends Cogn Sci,* 13:334–340.

Sinha, R. (2008). Chronic stress, drug use, and vulnerability to addiction. Annals of the new York Academy of Sciences, 1141(1), 105-130.

Smith, M. (1987). The Humean theory of motivation. *Mind*, *96*(381), 36-61.

Smith, L. B., & Thelen, E. (2003). Development as a dynamic system. *Trends in Cognitive Sciences*, *7*(8), 343–348.

Soodak, H., & Iberall, A. (1978). Homeokinetics: A physical science complex systems. *Science*, *201*, 18.

Solinas, M., Chauvet, C., Thiriet, N., El Rawas, R., & Jaber, M. (2008). Reversal of cocaine addiction by environmental enrichment. *Proceedings of the National Academy of Sciences*, *105*(44), 17145-17150.

Sporns, O. (2006). Small-world connectivity, motif composition, and complexity of fractal neuronal connections. *Biosystems*, *85*(1), 55-64.

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.

Sporns, O., Tononi, G., & Edelman, G. M. (2000). Theoretical neuroanatomy: relating anatomical and functional connectivity in graphs and cortical connection matrices. *Cerebral cortex*, *10*(2), 127-141.

Sporns, O., & Zwi, J. D. (2004). The small world of the cerebral cortex. *Neuroinformatics*, *2*(2), 145-162.

Sprengelmeyer, R., Young, A. W., Calder, A. J., Karnat, A., Lange, H., Homberg, V.,Perrett, D. I., Rowland, D. (1996). Loss of disgust. Perception of faces and emotions in Huntington's disease. *Brain,* 119:1647–65.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavorial and Brain Sciences*, *23*, 645–726.

Stapleton, M. (2013). Steps to a "Properly Embodied" cognitive science. *Cognitive Systems Research*, *22–23*, 1–11.

Sullivan, A. (2018, Feb 20). The Poison We Pick. *New York Magazine*. Retrieved from http://nymag.com/daily/intelligencer/2018/02/americas-opioid-epidemic.html

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: expectation and prediction. Psychological review, 88(2), 135.

Sutton, R. S., & Barto, A. G. (1998). Reinforcement learning: An introduction. Cambridge, MA: MIT Press.

Suzuki, A., Hoshino, T., Shigemasu, K. & Kawamura, M. (2006). Disgust-specific impairment of facial expression recognition in Parkinson's disease. *Brain,* 129:707–17.

Swanson, L. W. (1983). "The hippocampus and the concept of the limbic system," in *Neurobiology of the Hippocampus*, ed W. Seifert (London: Academic Press), 3–19.

Swanson, L. W. (2000). Cerebral hemisphere regulation of motivated behavior. *Brain Research*, *886*, 113– 164.

Szalavitz, M. (2016). *Unbroken brain: a revolutionary new way of understanding addiction*. St. Martin's Press.

Tetlock, P. E., & Gardner, D. (2016). *Superforecasting: The art and science of prediction*. New York: Broadway Books.

Thelen, E., & Smith, L. B. (1996). *A dynamic systems approach to the development of cognition and action*. MIT press.

Thelen, E., Schöner, G., Scheier, C., & Smith, L. B. (2001). The dynamics of embodiment: A field theory of infant perseverative reaching. *Behavioral and brain sciences*, *24*(1), 1-34.

Theyel, B., Llano, D., & Sherman, S. (2010). The corticothalamocortical circuit drives higher-order cortex in the mouse. *Nature Neuroscience*, *13*, 84–88.

Thompson, E. (2007). Mind in Life: Biology, Phenomenology and the Sciences of Mind. Cambridge, MA: Harvard University Press.

Thompson, E., & Cosmelli, D. (2011). Brain in a vat or body in a world? Brainbound versus enactive views of experience. *Philosophical topics*, 163-180.

Thompson, E., & Varela, F. J. (2001). Radical embodiment: neural dynamics and consciousness. *Trends in cognitive sciences*, *5*(10), 418-425.

Tishby, N., & Polani, D. (2011). Information theory of decisions and actions. In V. Cutsuridis, A. Hussain & J. G. Taylor (Eds.), *Perception-action cycle* (pp. 601–636). New York: Springer.

Todd, R., & Thompson, E. (2015). Strengthening emotion-cognition integration. Behavioral and Brain Sciences, 38.

Van de Cruys, S. (2017). Affective value in the predictive mind. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing* (Vol. 24). Frankfurt am Main: MIND Group.

Van de Cruys, S., & Wagemans, J. (2011). Putting reward in art: A tentative prediction error account of visual art. *I-Perception*, *2*(9), 1035–1062.

van den Heuvel, M. P., & Sporns, O. (2011). Rich-club organization of the human connectome. Journal of Neuroscience, 31(15), 775–786.

van den Heuvel, M. P., & Sporns, O. (2013a). Network hubs in the human brain. Trends in Cognitive Sciences, 17, 683–696.

van den Heuvel, M. P., & Sporns, O. (2013b). An anatomical substrate for integration among functional networks in human cortex. Journal of Neuroscience, 33(14), 489–500.

Verschure, P. F., Pennartz, C. M., & Pezzulo, G. (2014). The why, what, where, when and how of goal-directed choice: neuronal and computational principles. *Phil. Trans. R. Soc. B*, *369*(1655), 20130483.

Wicker, B., Keysers, C., Plailly, J., Royet, J. P., Gallese, V., Rizzolatti, G. (2003). Both of us disgusted in my insula the common neural basis of seeing and feeling disgust. *Neuron,* 40 (3), 655–664.

Wilson, R., Foglia, L. (2011). 'Embodied Cognition'. Stanford Encyclopedia of Philosophy. Available at http://plato.standford.edu/entries/embodiedLcognition.

Wilson-Mendenhall, C. D., Barrett, L. F., Simmons, W. K., and Barsalou, L. W. (2011). Grounding emotion in situated conceptualization. *Neuropsychologia* 49, 1105–1127.

Winston, P. H. (2012). The next 50 years: A personal view. *Biologically Inspired Cognitive Architectures*, *1*, 92–99.

Withagen, R., De Poel, H. J., Araújo, D., & Pepping, G. J. (2012). Affordances can invite behavior: Reconsidering the relationship between affordances and agency. *New Ideas in Psychology*, *30*(2), 250-258.

Wolfensteller, U., & Ruge, H. (2012). Frontostriatal mechanisms in instruction-based learning as a hallmark of flexible goal-directed behavior. *Frontiers in psychology*, *3*, 192.

Wolpert, D. M., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural networks*, *11*(7-8), 1317-1329.

Wundt, W. (1897). *Outlines of Psychology*. Translated by C. H. Judd Oxford: Engelman

Wurtz, R. H., McAlonan, K., Cavanaugh, J., & Berman, R. A. (2011). Thalamic pathways for active vision. Trends in Cognitive Sciences, 15(4), 177–184.

Yamamoto, D. J., Woo, C. W., Wager, T. D., Regner, M. F., & Tanabe, J. (2015). Influence of dorsolateral prefrontal cortex and ventral striatum on risk avoidance in addiction: a mediation analysis. *Drug & Alcohol Dependence*, *149*, 10-17.

Thomas Yeo, B. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., ... & Fischl, B. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*, *106*(3), 1125-1165.

Zajonc, R. B. (1998). Emotions. In Gilbert, Daniel T., Fiske, Susan T., Lindzey, Gardner, (Eds.). (1998). *The handbook of social psychology*, Vols. 1-2, 4th ed., (pp. 591-632). New York, NY, US: McGraw-Hill.