A Compatibilist Computational Theory of Mind

by

Marcus Erik Vestberg

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

Oxford Brookes University

School of History, Philosophy and Culture

March 2018

Abstract

This thesis defends the idea that the mind is essentially computational, a position that has in recent decades come under attack by theories that focus on bodily action and that view the mind as a product of interaction with the world and not as a set of secluded processes in the brain. The most prominent of these is the contemporary criticism coming from enactivism, a theory that argues that cognition is born not from internal processes but from dynamic interactions between brain, body and world. The radical version of enactivism in particular seeks to reject the idea of representational content, a key part in the computational theory of mind. To this end I propose a Compatibilist Computational Theory of Mind. This compatibilist theory incorporates embodied and embedded elements of cognition and also supports a predictive theory of perception, while maintaining the core beliefs pertaining to brain-centric computationalism: That our cognition takes place in our brain, not in bonds between brain and world, and that cognition involves manipulation of mental representational content. While maintaining the position that a computational theory of mind is the best model we have for understanding how the mind works, this thesis also reviews the various flaws and problems that the position has had since its inception. Seeking to overcome these problems, as well as showing that computationalism is still perfectly compatible with contemporary action and prediction-based research in cognitive science, the thesis argues that by revising the theory in such a way that it can incorporate these new elements of cognition we arrive at a theory that is much stronger and more versatile than contemporary non-computational alternatives.

Contents

Chapter I - Introduction

1.0 - An Introduction to the Computational Mind.

The core of this thesis is a defence of the computational theory of mind, or rather it is a defence of the idea that our mind operates in essentially computational ways; through manipulation of internal symbols that we call mental representations. What I will be doing in this thesis is essentially two things: First, I will defend the existence and importance of mental representations in philosophy of mind and cognitive science, this in the face of proponents of radical enactivism, who seek to deny the existence of mental content. Second, I will suggest a compatibilist computational model of the human mind, fusing together classical aspects from Fodor - like modularity - and Kahneman's System 1 and 2, then show how these are perfectly capable of accomodating newer theorised aspects of cognition – like embodied and predictive theories, which have seen a rise to prominence in the last couple of decades – without losing the core of what makes computationalism such a comprehensive approach to explaining the mind. My motivation for this move is that I believe that the classical version of the computational theory of mind that grew up in the 70s and 80s is simply not fit to handle the new way we understand cognition through mirror neurons,[1] backward connections and corollary discharge in the brain. Yet, it is also rigorously equipped to deal with many of the nuances that appear in the way we think, perceive and interact with the world. All that is needed, I argue, is to take out the 'engine', to use an analogy, and place it within a new frame that is better suited for the modern landscape, yet also still *compatible* with the theory's central claim; that thinking is essentially the manipulation of mental symbols, some of which is done modularly and some of which is done in central processing. This is also accompanied with a defence of the idea that the mind is brain-centric, and that within the mind itself a central processing system is the essential cornerstone of thought, and not merely a mediator. The latter is an idea presented by enactivism, a theory holding the position that cognition is a trinity of brain, body and world, and that it is the interaction between these three that constitutes mind and cognition.

The computational theory of mind (CTM) was originally put forth by Hilary Putnam in his article 'Brains and Behaviour' (1961). The theory is grounded in the idea that the mind works in many ways like a digital computer; the mind is parsing internal representations (symbols) in algorithmic ways,

[1] Mirror neurons were first discovered by di Pellegrino, Fadiga, Fogassi, Gallese and Rizzolatti (1992). They revealed that we had similar events going on in the brain while performing an action as we did when observing that same type of action. For more in-depth analysis, see Rizzolatti & Craighero (2004).

forming an internal computational language that is used to process input data into output. While the theory started with Putnam, it was further developed by Jerry Fodor (1975). Fodor saw the symbol-dependent processing of the mind as a language and referred to this internal syntax as "the language of thought" and "mentalese." This interpretation placed mentalese as essentially a computational language of the mind, physically realised in the brain. By comparison, we could compare this to how a computer language works, and how the symbols and syntax of programming languages are realised in software (thinking), but physically situated in hardware (the brain). In fact, Fodor wants to claim that this is exactly what is going on, suggesting that the mind is a physically realized computer environment where information processing occurs. That is, our minds are a formalized system parsing a language based on information-carrying representations, structured by syntactic and semantic rules. It is a computer environment that manipulates symbols based on a formal rule set (Fodor, 1975). As such, CTM was a very brain-focused account of the mind,[2] the idea being that the mind and the world are connected through our understanding of what is effectively a second simulated world within our mind. Just as there are dogs and parks and buildings in the actual world, we have representations of dogs and parks and buildings *in our heads*. Information about the world enters the mind as sensory data; the things we see, hear, taste, smell and feel all enter our mind as raw data, which is then used to form mental representations, starting simple and building in complexity to form concepts. These representations are the components of our thought processes, which in turn are algorithmic in nature; our thought processes are problem-solving operations using an internal rule set which determines how the symbols (representations) are to be manipulated by the system. Representations within the mind range from simple to complex, in such a way that "(a) there is a distinction between structurally atomic and structurally molecular representations; (b) structurally molecular representations have syntactic constituents that are themselves either structurally molecular or are structurally atomic; and (c) the semantic content of a (molecular) representation is a function of the semantic contents of its syntactic parts, together with its constituent structure." (Fodor & Pylyshyn, 1988, p.8) The representation of the thought "I want a glass of milk, so I will go to the kitchen" is thus built up out of simpler representations like "I want", "milk", "to go", "kitchen" etc. In this way, we can see that mental content within CTM carries meaning in the same way that words in a language carry meaning: When I speak of "glass of milk" and "kitchen" there are particular objects/locations (types and/or tokens) to which these words refer. By having these words, like "kitchen" I am able to speak meaningfully about kitchens. Equally, our thoughts are syntactic processes (just like sentences) built up of representations of various

[2] Just as information processing states in computers are physically realized in hardware states, so is our mental representational processing physically realized in brain states. The neural pathways of the brain compromise the circuitry, the language of thought is the computer language, and our thoughts are the processes that take place.

objects and concepts, allowing me to meaningfully think about these objects and concepts. If I hold the belief that the neighbour is stealing my mail, I have a representation in my head such that "the neighbour is stealing my mail." This mental representation contains meaningful content about things external to me, like "neighbour" which both represents the general concept of what a neighbour is, but also the specific token of neighbour that the sentence is referring to, the one I suspect of stealing my mail. This belief may lead to follow-up thoughts such as "I should ask the postman about it" or "I should confront my neighbour about it." When we are dealing with the world, we are thus dealing with gathered data that has formed a comprehensible and (one would hope) accurate representation of what the world is like outside of our heads. One could liken this to viewing the world on the screen of a digital camera; what is shown on the screen is the gathered data coming from the lens, then displayed to us. The 'us' in the case of the mind would be the Central Processing System which, according to Fodor, in turn would be separate from perceptual what we call 'modules' by limited accessibility (Fodor, 1983). In short, certain parts of the mind are subconscious, in that they handle mental processes (processes that are definitely ours) that we are not directly aware or knowledgeable of. This is not necessarily a phenomenal consciousness we speak of, but rather modules are subconscious in terms of access consciousness (Block, 1995). In this sense, CTM's central processes are *access conscious* in the sense that they are open to interaction in a global workspace.[3] Meanwhile, modular processes are only partially accessible due to encapsulation. These modules operate quickly and automatically, while the central processing system operates more in the way in which we experience our inner monologue or trail of thought (more on this later). CTM became a very popular theory of mind in the 1980s and gave new fuel to cybernetics (which had already been around since the 1930s) which led to the formation of modern cognitive science, as well as the resurgence of artificial intelligence research. As a result of this, throughout the 70s, 80s and 90s, much of the research in cognitive science and AI followed computational models. Despite its popularity, in philosophy CTM has had its fair share of criticisms. Gödel's incompleteness theorem, which was published in 1931 (Gödel, 1931/1992) is an early example of criticism to liken the mind to a mathematical system. By the time Putnam and Fodor were presenting computation as a theory of mind, the worry that Gödel stirred had been mostly moved away from until it was brought back by Penrose in 1994. While Penrose's attack on CTM had a few good points (as I will outline below), it was mostly unsuccessful in striking CTM down (McCullough, 1995; Feferman, 1996). However, the 90s were still a decline in CTM's era as the leading theory of mind. One of the reasons was the rise in

[3] To Block, there are different types and aspects to consciousness. In terms of phenomenal consciousness, we are talking about the experience of what it is like to consciously perceive or be aware of something. This kind of consciousness can be hard to properly explain or integrate in a theory as there is no clear connection between our scientific understanding of the mind and brain, and the 'redness' of a visually experienced tomato. By contrast, access consciousness focuses on the aspect of consciousness where we are to degrees aware or unaware about our own mental states.

popularity of connectionism, and the people adopting it generally went against the idea of a language of thought. Connectionism was built upon neural networks of interconnected nodes rather than the more linguistically inspired CTM. In particular, eliminative connectionism sought to move away from the idea of computationalism and mental representation in thought (Churchland, 1989; Horgan & Tienson, 1996). Thus the 90s was a transformative era where we moved away from the heyday of classical CTM and arrived in the contemporary era where CTM has decreased in popularity. Whenever I refer to CTM in classical terms, it is thus the pre-90s era definition in particular that I refer to. In contrast, whenever I refer to *contemporary* theories, I refer to the theories in post-90s philosophy and science.

1.1 – The relevance of AI.

The history of AI (Artificial Intelligence) research is also indirectly (or sometimes directly) linked to the history of various philosophical theories of mind. While the endeavour of AI research is primarily one of discovering how to artificially create intelligence, discoveries within that field can also indirectly teach us lessons about how the human mind itself is likely to work. After all, should we ever stumble upon a time where an Artificial Intelligence is *perfectly* indistinguishable from a human in cognitive capacities (say a futuristic android beyond most sci-fi proportions) we would likely want to say that we have, in fact, gained some insight into plausible model of how our own mind works. Similarly, should our theorized models of how the human mind works fail as models for creating AI, then there is an equally compelling argument to make that such results discredit said theorized models. The successes and failures in AI research have affected how many philosophers of mind treat the current prevalent theories of mind. In fact, the failures of primarily input-based AI models have caused a similar backlash to the traditional or, as you could call it, Classical Computational Theory of Mind as developed by Putnam and Fodor. Classical CTM has not so much come under attack as it has been left behind for what are perceived as greener pastures. Some of these greener pastures, as we will see later on, involve a move from passivity to more pro-active, engaged and prediction-based frameworks. One of these new frameworks is the collection of theories that fall under the shared label “EEEE” or the E-theories. The EEEE (standing for Embodied, Embedded, Enacted and Extended) are actually four schools of thought that to varying degrees want to put more focus on cognition in terms of interaction with or offloading upon the world outside of the brain. In this way, these theories are looking to some extent to break down the separating barrier between mind and world that classical computationalism has set up. They do not outright deny that there may be *some* computational elements in cognition (apart from enactivism which outright rejects even the brain's

central role in mindedness [Noë, 2004]) but they deny the brain the role of an input-output sandwich that handles all cognition on its own (where cognition is the filling hidden between input and output to and from the brain). One of their main reasons for this turn is indirectly tied to developments in AI and robotics, where recent successes have come from reading and navigating the environment. In contrast to CTM, these more body-and-action-focused theories take inspiration from how the new type of body-and-environment-focused robotics-based AI[4] has generated more success in contemporary times compared to its brain-and-neural-focused cousin and take this as a cue that this points to new potential discoveries in our understanding of the mind. In addition, there are recent developments in neuroscience that seem to go against ideas held by classical CTM, for example the idea that our brains may be prediction-based rather than relying on rich input (Clark, 2013), as well as the above mentioned discoveries of mirror neurons, suggesting that our mental states for action and observing action are closely related (Rizzolatti & Craighero, 2004).

The emergence of the E-theories has brought new perspectives into light in philosophy of mind as well as cognitive science, but I still want to argue that CTM has a lot to offer and that the rejection of its classical version should not spell doom for representation-based computationalism as a whole. What's more, I think that even classical CTM is more compatible with these newer ideas than many philosophers would admit, and I aim to show how CTM can be tweaked and updated to a more compatibilistic approach. There are many ways to improve upon the state of CTM, and in this thesis I want to bring the best offers to the table, while at the same time going through why I think the alternatives offered by EEEE ultimately fail to topple CTM from its place as the best model of the human mind available.

1.2 – A historical overview of CTM

The Computational Theory of Mind was born out of the emergence of computing machines, most prominently the abstract Turing machine, invented by Alan Turing (1936) as a model to, among other purposes, find a solution to the 'decision problem' in formal logic. Around the same time, the ancestors of our modern digital computers were developed: The Atanasoff-Berry Computer (ABC) in 1942, The Colossus and ENIAC both in 1943.[5] What followed was the rise of information technology

[4] For example the Mataric robot (Mataric, 1990, 1992) that navigates a maze by combining sensory input and motion-states when mapping the area. Or the passive dynamic walker robots (Collins et al., 2001) that use the natural pendulum motions of legs to aid walking, creating a smoother gait than the computation heavy and precision-demanding ASIMO.

[5] Though more primitive computers existed before, for example Gottfried Wilhelm Leibniz's stepped reckoner, the world's first digital mechanical calculator, which was completed in 1694.

that sparked the computer age. A line of thought came about that if machines can perform calculations and solve problems, could they be considered intelligent? Could a machine think? Turing himself asked these questions (Turing, 1950) and developed what would be known as the Turing test, a test where an interrogator would pass written questions to two other anonymous participants (one human, one machine) and would receive answers from both. The interrogator would then determine which participant is the machine, and which is the human. The purpose of the test was for the machine to exhibit their capacity of intelligent behaviour. If the machine could, even after rigorous questioning, appear to the interrogator as the human participant, then the machine had a claim for being an intelligent thinker. This pursuit of intelligent machines had a reverse side to it: If these machines could solve problems and make decisions in a way that approaches cognition, could cognition itself in fact be computational? Such a thought offered a robust and structured approach to modelling the mind, while providing an analogous relationship with the emerging science of creating better and more complex computers.

The rise of a computational theory of mind took place in this landscape, beginning in the earliest stages with a suggestion by McCulloch & Pitts (1942) but was properly put into theory by Putnam (1961) and was further developed by Fodor (1975). The comparison between human intelligence and the workings of computers offered a solid argument for how the mechanisms of our mind and thoughts were structured, and provided a bridge between the abstract mind models of philosophers with the physical form which these cognitive faculties inhabit. If machines could think, then human thinkers could also be like machines. The 1960s saw the beginnings of the interdisciplinary field of cognitive science, which involved linguistics, psychology, artificial intelligence, philosophy, neuroscience and anthropology. CTM served as a foundational theory in cognitive science, and enjoyed great popularity through to the 1980s.

The Computational Theory of Mind holds two separate core claims: The first is that mental states are made up of representational symbols possessing syntactic and semantic properties. As such, thoughts are constituted by representative states, the contents of which references what the thoughts are about (as I have described above). The second claim is that these symbols possess semantic and syntactic properties and that through these properties the mind as a purely causal system. Thus when we say that human cognition is computational, what we are saying is that our thoughts follow a system of cause and effect, based on the syntactic properties of our thoughts and their semantic contents. The attractiveness of such a theory lay in that reasoning could be explained by the cause and effect of the system, rather than remaining an esoteric process. Human thoughts, and how input causes these thoughts which in turn cause output (in the form of input for new processes or as initiation of action in response to the input), could be explained in detail as

algorithmic processes. Putnam's (1961, 1967) account of CTM offered several advantages over two previous theories about the nature of mental states: logical behaviourism (which proposes that having a certain mental state is enacting a certain pattern of behaviour) and type-identity theory (in which specific mental states are realised in specific physical states). To put into an example, being in a state of pain, to the behaviourist, would be to have a certain pattern of behaviour (i.e. we act like we are in pain). In contrast, to the type-identity theorist, being in pain is to be in a certain specific brain state which causes us to feel pain. Putnam (1967) instead proposes a type of functionalism, where being in pain is a functional state of a cognitive system. This has the advantage over the behaviourist in that I can be said to be in pain even if I am incapable of showing any outward behaviour of being in pain. What is more, this allows CTM to describe mental states that do not have any direct relation to behavioural patterns. For example, I may possess mental states that causally only relate to other mental states, and not to any specific sensory input or behavioural output. The functionalistic approach has an advantage over type-identity theory in that any mental state, being functionally rather than physically realised, can be realised in a multitude of ways. Putnam argues that defining a state of pain is a much smoother task for a functionalist than it is for an identity theorist:

"[The brain-state identity theorist] has to specify a physical-chemical state such that *any* organism (not just mammal) is in pain if and only if (a) it possesses a brain of a suitable physical-chemical structure; and (b) its brain is in that physical-chemical state. This means that the physical-chemical state in question must be a possible state of a mammalian brain, a reptilian brain, a mollusc's brain (octopuses are mollusc, and certainly feel pain), etc." (Putnam, 1967, p.164)

CTM thus allowed for an approach that, rather than looking at behaviour or physical states, looked at the mind as a functionalistic system where, as long as the representational structure and computational processing remained the same, could be found in any kind of organism, or intelligent machine. While breaking the barrier between understanding the brain and understanding thought processes would prove a difficult task (a task that is still underway), the comparison to computer systems allowed for the creation of hypothetical mind models that could be tested in the realm of artificial intelligence. This, in turn, was why the theory was so attractive for artificial intelligence research, which in turn made it attractive for the rest of the fields involved in cognitive science.

Fodor further built upon CTM, turning it into the version that I in this thesis consider as Fodor's 'classical' Computational Theory of Mind. Fodor (1975) took the semantic and syntactic aspects of CTM and properly defined cognition as a language of thought, involving the syntactic causal processing of representational symbols. This hypothesis dubbed the language of thought as

'mentalese'. Fodor (1983) also introduced modularity of mind, proposing that the mind was not a singular system, but a central system connected to several peripheral sub-systems. The existence of these sub-systems were to explain how certain cognitive tasks are easier to perform within certain contexts, or how certain input and learning systems were able to process information so efficiently without apparent cognitive strain. This involved sensory systems like vision, touch and smell, as well as language. The idea was that within these systems there were specialized modules for certain tasks in both 'high' and 'low' levels of cognition (in vision, for example, higher-level tasks like facial recognition, and lower-level tasks like colour perception). Thus with the introduction of the language of thought, as well as modularity, Fodor had considerably expanded the scope of CTM, creating what I would claim to be the defining classical version of the theory.

Even though CTM enjoyed a lot of popularity in the 70s and 80s, it was not without objections raised against it. Some of the more prominent traditional arguments (to which I refer to arguments stemming from before the EEEE era) that directly criticised the theory came from Lucas (1961), Searle (1980) and Penrose (1994). Both Lucas and Penrose's arguments stem from the claim that Gödel's incompleteness theorem posed a problem for the view of the mind as a computational system. The core claim was that according to Gödel a system cannot verify the truths of its own claims, and thus an external (human) third party was needed. Exactly what this means and how such objections could be tackled I will address later on in this chapter, and will therefore not dwell on it much further here.

Searle's objection has a similar goal of showing that a computational system is 'not enough' to constitute cognition as we understand it in humans, but his argument comes from a different direction: language. Searle (1980) presented the Chinese Room Argument, a thought experiment devised to prove that no matter how well a computer could simulate thought, it could never be properly described as being a thinking thing. This argument went against the idea of 'strong AI' (the idea that a mind very much like the human mind could be one day artificially created) and, linked with the idea I presented earlier that if computational machines can be thinkers, then thinkers can be computational, also went against one of the core claims of CTM – that the mind could be fully explained through computational processes. Searle invites us to imagine a man sitting in a room with two books. One book contains Chinese script, which the man cannot read nor understand. The second book is a manual in a language that the man *can* understand, describing how to correlate certain Chinese symbols with others. The man is given a third batch of text through a hole in the wall, and is tasked with using the manual to correlate the given text (which is a question in Chinese) with the Chinese symbols in the scripture he already possesses. If done accurately, the man will be able to produce a string of symbols which he can give to the people outside of the room. Thanks to the

instruction manual and the Chinese script, he will have produced an answer in Chinese to the question he was given. Searle argues that this is the same as saying the man correlated one set of formal symbols with another set of formal symbols based on their shapes and with no understanding of actual Chinese. Searle's claim here is that the man is a functional computational system, yet has no understanding in the way CTM and strong AI propose computational processing could give rise to. After all, if all we do when we speak a language is correlate representational symbols with one another, creating output (responses) from input (stimuli) how do we experience an understanding of language unlike the man in the Chinese room? Though a strong claim, Searle's argument ultimately fails when confronted with modularity. Searle fails to recognise that CTM does not claim the mind to be a man in a room (this homunculus perspective will come back later as well, as it is a common misunderstanding of CTM). Rather, CTM claims the mind to be the man *and* the room, i.e. the mind is the system as a whole. The man in the Chinese Room Argument is, if anything, a modular subsystem, and thanks to Fodor's idea of encapsulation in modules, which I will dive into more detail on further on but can for now be summarised as limited access to knowledge, a singular module doesn't have to possess higher-level understanding of its tasks, making Searle's argument about the man's ignorance moot.

In the new millennium CTM faces objections from the relatively new grouping of the EEEE theories. I will more thoroughly introduce in the next chapter and thoroughly explore their arguments throughout the rest of the thesis, but for now I will present a quick summary of the most common criticisms:

The mind is not all in the head. This is a position that all EEEE theories adopt, but some to a greater degree than others. Embodied and Embedded cognition propose that brain-external factors have a greater role to play in cognition than classical CTM lets on, while Extended cognition claims that the mind can fully and functionally extend and involve brain-external objects as actual parts of the mind. Finally, Enactivism embraces action and interaction over location, claiming that the mind exists in the interactions between world, brain and body. In this view, the mind is a trinity of interaction and cannot be said to be 'located in the brain' as the brain is merely a mediator for enactive thought. In later chapters I aim to show that CTM is not entirely brain-bound by necessity. As the title of this paper reads, I aim to present a compatibilist theory of mind where I prove that CTM can indeed be compatible with at least the Embodied and Embedded claims.

Visual systems based on CTM do not accurately describe how visual systems work in the face of new scientific findings, therefore CTM is a bad theory to base cognitive system models on. This argument comes from criticisms directed at Marr's theory of vision (1980, 1982), a very rigorous

model of the visual system presented by David Marr in a way that follows the claims of classical CTM at the time. My argument against this criticism is that such criticisms are outdated, for they rely on an interpretation that a CTM-based visual system relies on an input-heavy model to explain how we perceive and understand objects in the world. In Chapter 5 I will present a counter to this position in the form of Predictive Processing, a new development in CTM that takes the theory forward, allowing it to comfortably coexist with the aforementioned new scientific findings.

The mind does not operate based on content, symbols or representations. This argument comes mainly from Enactivism, and is based on the idea that mental content is irrelevant to the functioning of symbols, which in turn is irrelevant to the functioning of representations. Ultimately, to posit that the mind is at all computational is a foregone conclusion without evidence, as it bears no functional meaning upon the interaction between mind, body and world. This is an argument that I will debate through the latter parts of the thesis as I defend CTM's claim of representational content. My main position in all these arguments will, at its core, be that Enactivism is throwing the baby out with the bathwater, for it attempts to reduce the mind, losing explanations to cognitive features for which they offer no satisfactory alternate explanation.

1.3 - Penrose and non-computable algorithms

There are two ways in which CTM can fall under attack. One is that we propose there are certain thought processes going on within our mind that we argue simply cannot be described as computational. The other is that we reject the existence of the tools that makes computation in the mind possible, mental symbols and representations that is. The latter ultimately poses a much greater threat to CTM as a whole and it is something that will come up later on once we start discussing enactivism, but in this section I will have a look at some examples of the former type of criticism. In *Shadows of the Mind* (1994), Roger Penrose sets out to explain his claim that the reason why computers seemingly never reach human-level thinking is because the proposed computationalism in human minds is either a) not the whole story or b) not entirely figured out and properly simulated. If a mind was entirely computational, it would be an abstraction of symbol manipulation, which in turn could theoretically be simulated in another system like a computer. Penrose's claim as to why we fail to simulate human thinking in computers is because they lack the element of awareness and understanding, which in turn is a product of our human consciousness. He also wants to argue for the idea that consciousness is not something that is beyond physical explanation, and that artificial intelligence on a human level (and beyond) is indeed possible in the future with a widened scientific world view. Here, 'consciousness' is introduced as a key ingredient

for unlocking capabilities beyond what we currently understand as basic processing in the brain. The key for this widened science, according to Penrose, is that it must include more exploration into quantum mechanics.

The central claims are as follows: Human intelligence is not just about computational power, i.e. the processing of algorithmic tasks. If that was the case, then computers would already be way ahead of us when it comes to 'thinking', for there simply is no contest between humans and computers when it comes to raw calculation capacities. What does make a difference between humans and computers is the intuitive understanding of mathematical principles that allows us to make judgements and draw conclusions a computer cannot. This includes judgements on questions such as whether or not there is an odd number that is the sum of two even numbers. This type of understanding is a product of us having a consciousness or an awareness that is above what computation alone could accomplish. Penrose focuses on mathematics for the purposes of illustrating examples, but the principle of awareness and understanding can be applied to a much wider range of cognitive cases such as decision-making (Gigerenzer, 2008) and perception (Noë & O'Regan, 2001).

The core of Penrose's argument is thus that there is a yet-to-be-discovered ingredient to the human mind, a source or catalyst for the emergence of understanding/awareness. This ingredient is allowing for an element of non-computability that makes us essentially different from computers in such a manner that even the best 'human simulation' or Turing test will eventually fail as long as it only relies on computational processes. Still Penrose explicitly says that he wants to stay clear of the mystical and avoid falling on a conclusion that our consciousness is something that cannot be understood by scientific exploration, and thus stays clear of any form of dualism or ghost in the machine. Instead he sets out to speculate where the scientific explanation for consciousness could be found. He makes it clear that he does not believe science in its current state can facilitate an explanation, but could in the future if our grasp of the sciences and our world-view expanded due to new discoveries. He still thinks it possible and worthwhile to find hints of where such development would be necessary, his own speculation being that this 'awareness' is born out of quantum coherence (Penrose, 1994). Penrose thus ends up with a position that the human mind is perfectly scientifically knowable, but only through a future more developed science, not our present day scientific theories. The computational powers of the mind, if they are of such a nature, are not enough to tell the story. For that we need an additional component, external to the system, that completes our intelligence.

His argument draws heavily upon the theorems of Turing (1950) and Gödel (1931/1992), exploring how Turing machines relate to the computable model of the mind and the mindset of computationally inclined AI research while also using Gödel's incompleteness theorem to show that a mind-model reliant solely on algorithmic action is insufficient to match the human mind and will eventually fail (an objection also brought up by John Randolph Lucas [1961]). In the context of Penrose, the argument goes roughly as follows: For a formal system F to be complete and consistent (or sound),[6] it needs to assert the truth that it is consistent. However, in a complete system there is a statement that cannot be proven, a self-referential paradox of 'there is no proof for P' (P being one of many axiomatic sentences in the system). Since proving this causes a contradiction, and denying it causes incompleteness, the only way to pursue completeness is to take a step outside of the system, creating a new system F'. The problem here being that this will continue ad infinitum. Suppose that the formal system that needs proving here is the mind of the person performing the proof. In doing so it becomes F', but this cannot be. If F is the whole of the mind we cannot step outside the system and create a superset. Yet, we are rarely faced by the paradoxes created by this theorem, and they don't seem to be a hindrance in our understanding of the completeness of system F. The essence of what Gödel's theorem is trying to show, as Penrose puts it, is that "the powers of human reason could not be limited to any accepted pre-assigned system of formalized rules." (Penrose, 1996, §4.2) Penrose is keeping Gödel's theorem in mind as he focuses on what he sees as an evident difference in human minds and computers; the subject's reaction to a non-stopping algorithm (in the Turing sense).

Penrose further seeks support for this notion by looking at natural selection. He makes an argument in Chapter 3 that a general ability to understand is a much more plausible advantage to be taken up by natural selection than an ability, based on advanced and sophisticated mathematics, to pick out when an algorithm is non-stopping or not. In addition, he argues that the rest of our sophisticated mathematical thinking, while useful, is not something that would have earned any evolutionary priority in the early stages of our mindedness as it is less useful to have in comparison to the ability to make quick and frugal decisions. Penrose is not alone in spotting the value of frugality of mind, and I will return to this below in Chapter III where I will talk more about Gerd Gigerenzer and his work in heuristics and psychology. For now, however, I will continue with Penrose's case.

[6] In logical terms: For a system to be **complete** means that for a certain property, every formula with that property can be derived using the system. In the case of Gödel, the system needs to derive every true formula, including the formula that proves the truth of the system itself. A **sound** system is a system that can infer logically valid formulas, i.e. formulas where it is impossible for the premises to be true, yet the conclusion based on these premises to still be false. Finally, a system is **consistent** if it does not contradict itself.

Penrose notes that while there are some computational alternatives that could be suggested as candidates for detecting the so-called 'loop-sentences' problem for computational minds, none of them really brings a satisfying result. One of the suggestions is a gauging system that sends out an alert when it has discerned that the algorithm probably does not stop. Penrose rejects this as a good solution. A computational gauging system would itself, by its very nature, be subject to possible loop errors. One could propose variants set up to avoid these loops by, for example, introducing a random or time-based element, but it would do little good. A gauging system based on a random element can never reach certainty in the way that we humans can be certain about that non-stoppable nature of certain algorithms. Such a system can only ever reach a state of probability. Removing the random element and instead having another variant of a gauging system that alerts when the computation has 'gone on for too long' (thus relying on a set measurement of time) would many times be wrong due to there being computations that take an extreme amount of time yet, obvious to us, is bound to stop at some point. Penrose here gives an example of the computation $2^{2^{\wedge}65536}$ 1's being printed at the speed of a 1 every 10^{-43} seconds. This task obviously ends, we can judge its finite nature by common sense, but the task itself would take longer than the age of humanity up to this point. After these examples, the question remains of what kind of computational system could circumvent loop sentences. As stated above, I think the answer to this lies in heuristics.

In general, Penrose's theory was faced with a great deal of criticism and in 'Beyond the Doubting of a Shadow' (1996) Penrose responds to a number of these criticisms that were raised it. He remarks that the majority of the criticisms were aimed at the implications drawn from Gödel's theorem. Still, even if one were to remove Gödel's theorem from the equation and claim that it is of no relevance to the human mind (one could for example take the approach that the formal system composing a human mind is in fact not consistent, even though we'd like to think so) Penrose would still have a point that the issue of a non-stopping algorithm and our judgement of it as such is something that genuinely needs looking at. Even after dodging the Gödel bullet it remains unexplained how we humans can understand mathematical principles in such a way but are unable to come up with a good idea of what the formal criteria for judging an algorithm as non-stopping would be. However, this argument relies on an assumption that one particular commentator did not like Penrose taking for granted. Chalmers (1995a) writes that "Penrose is not taking AI seriously" when he generalizes all AI creations to be "something akin to theorem-provers in first-order logic" (Chalmers, 1995a, p.10). If AI was not bound to algorithms to make the kinds of judgements a human mind can, the argument that humans can judge that an algorithm is a non-stopping one but AI computers cannot might still hold true *at present*, but it would not follow that an AI computer would never be able to make those judgements because of inherent limitations in its cognitive model. Chalmers points out that there are

many other computational procedures, connectionist networks being an example, “which are not decomposable into axioms and rules of inference” (Chalmers, 1995a). However, it is important to note that this is just a way of avoiding the problem of non-computability and cutting short Penrose's ideas of what directions AI *must* take. What we have by the end of this is a claim by Penrose that human minds are not wholly formal systems (especially not in the realm of explicit judgements), followed by Chalmers' claim that AI minds aren't necessarily either, at least not in the sense of classic digital computation. Chalmers’ claim opens up a new world of possibilities for Penrose where non-computability is no longer the only way to go. This takes the air out of Penrose's argument that the human mind is necessarily non-computable, but it does not deny the possibility that there is an element of the human mind that is non-computable. Personally, I think that modularity defeats Penrose’s argument pretty handily. Penrose describes the computational mind as a single system, thus the Gödelian trap becomes relevant, but a modular system does not need to self-verify in that same manner. A modular mind is constituted by multiple inter-connected systems and sub-systems, some of which our cognitive ‘awareness’ has minimal control over, and can only access when reading the output. A central processing system surrounded by modular sub-systems is able to constantly check and verify its own constituent parts, and is able to make judgements upon data processed by its various modules. One module can thus assess the validity of another, making self-referential completeness irrelevant. I will go into further depth on modularity in Chapter III, but for now I will simply state that I do not think the Gödelian argument is too great a worry for a modular CTM.

1.4 - Frugality and heuristic psychology

In his paper 'Why Heuristics Work' Gerd Gigerenzer explains his theory of mind called the adaptive toolbox. This theory describes the mind as "a modular system that is composed of heuristics, their building blocks, and evolved capacities." (Gigerenzer, 2008) What he means by this is that the cognitive capacities in human judgement consist of a set of heuristic tools that, much like the different tools in a toolbox, are selected depending on the situation. This thus creates a modular system of making judgements where each module is much more successful at solving certain problems than others. In general terms, a heuristic is a seemingly simple rule of thumb used in decision-making to come up with good-enough hunches of what to do. What he argues for is that we for the most part (but not exclusively) use these heuristics when making decisions. An example of an everyday heuristic is this: I’m trying to choose a good wine for a dinner party. Among the wide selection I notice one bottle in particular that is on sale and that I recognise from a recent advertising campaign. Since I recognise this wine or its brand, I favour it heuristically thinking that it has to be

more popular. After all, I haven't heard anything about all these other wines on display. I purchase the bottle and walk away. I have, by the use of a rule of thumb,[7] come to a quick decision that I judge will be good enough. Indeed, the wine I chose is unlikely to be the worst on the shelf (unless the advertising campaign was to save a failing brand) but it is also unlikely to have been the best. In a non-heuristic fashion, in contrast, I would have employed further reasoning and research, read more labels of bottles I didn't recognise, picking out the flavours that I know I enjoy and that would go well with whatever we're having with dinner. Additionally, I may employ critical thinking, making the judgement that maybe the most recognisable brand is not the best, as a heavy advertising campaign also points to mass production, which *could* mean lower quality on the individual bottle. Thus, heuristics are tools for fast and easy decision-making, sacrificing optimality of result in favour of saving on the effort that is usually involved in more rigorous deliberation.

Gigerenzer describes logic theory of mind (henceforth just "logic theory") as being the view that the mind is an intuitive logician. Such a system focuses on solving syllogisms to make judgements; a syllogism being the inference of a conclusion from two or more premises. Logic theory thus presents a formal truth-seeking model of mind, much like how Penrose envisions computational systems as a whole. Gigerenzer claims that the relevance of such models when it comes to human psychology has always been discussed, and many times refuted, ever since psychology started out as a subject (Gigerenzer, 2008). Probability theory of mind (henceforth just "probability theory") is different from logic theory in that it views the mind as working to achieve the goals of the mind through risk-taking. The risk-taking comes in the form of statistics where, instead of deducing truth, a calculation of probability is made through induction where the most likely answer to a problem will be the answer.

The heuristic theory of mind is different from the two previous theories in many respects, but one in particular; a heuristic model of mind is not as reliant on information as the previous two. Gigerenzer describes a heuristic model as focusing on acting fast in situations where unknown information and probabilities make finding the best solution (optimization) impossible. A heuristic model thus uses tools to quickly find a good-enough solution (or judgement) to a problem (or situation). While this type of approach lacks the truth-finding capacities of logic theory, it succeeds much better in fast processing due to not having to traverse through extensive amounts of information and evaluation. Gigerenzer compares this to the tools in a handyman's toolbox, where each heuristic strategy is best suited to handle a certain set of problems, while they would be less effective for other sets.

[7] In this case the recognition heuristic, which I've listed before and will describe below.

There are six main misconceptions that Gigerenzer wants to refute. The first being that since heuristics do not produce optimized results, they're second-rate strategies to solving problems or making judgements based on inference. Gigerenzer here points to computational intractability that causes logic and probability approaches to fail in delivering an optimized solution. The point he wants to make is that none of the three systems can be labeled the 'best' in every situation. Heuristics are thus not second-rate because there are situations where heuristics produce the best results, these being situations where we need to make quick decisions with little available information. From this, one could say that while heuristics work like a toolbox, the usage of logic, probability and heuristics can be seen as a choice of tools in its own right. In that respect, one would not call a hammer second-rate to a screwdriver just because a hammer produces less than optimal results when forcefully hammering in screws. A hammer works excellently with nails, where screwdrivers are completely useless due to lack of fastenings and mating capacities.

The second misconception is that we would rely on heuristics only because of cognitive limitations. Again, Gigerenzer refers back to computational intractability which prevents logic and probability from being utilized fully and properly. Thus it is rather a limitation in those strategies rather than in our ability to perform them.

The third misconception is that we only rely on heuristics in routine decisions of little importance. Gigerenzer argues that we also use heuristics when we, for example, decide if we want to donate organs or not. What he points at here is that in USA (where organ donation is something you opt for) there are fewer organ donors than in France (where you are a donor by default). He argues that there's a heuristic rule of thumb for sticking to the default if a choice is given, unless other factors would cause the agent to go away from the default.

The fourth misconception is that the usage of heuristics is mostly an indication of lacking cognitive abilities, and that those with higher cognitive abilities employ more of the logic and probability strategies and integrate more information into their reasoning. Gigerenzer here simply claims that experimental evidence does not support this claim. Rather the opposite: cognitive capacities seem linked to our adaptive selection of heuristics.

The fifth misconception is that affect, availability, causality and representativeness are models of heuristics. Gigerenzer replies to this that these are “mere labels, not formal models of heuristics. A model makes precise predictions and can be tested, such as in computer simulations.” (Gigerenzer, 2008, p.21)

The sixth and final misconception is that more information accompanied with a computational approach is always better than a mere heuristic approach. Aside from the examples with intractability given earlier, Gigerenzer here gives an example about research on stock market investment behaviours to prove that it is sometimes better to ignore information when it comes to handling a “partly uncertain world” (Gigerenzer, 2008, p.21). What the research boiled down to was that when comparing an intuitive and well-documented heuristic ‘‘Allocate your money equally to each of N funds", or "1/N", to other methods of optimizing ones allocation of funds, the 1/N strategy worked better in predicting the future. In contrast, the more complex methods worked better in fitting past data. Typically, Gigerenzer claims, the larger uncertainty in predictiveness, the larger number of funds and the smaller room for learning, the more successful predictions given by 1/N. Heuristics are thus very effective when we have small amounts of information to work with, which works well when (as in this case) making decisions about an uncertain future. While 1/N is used in investments here, Gigerenzer points out that 1/N is not an investment heuristic. Heuristics have much broader applications. An example here in the tools-realm would be that a screwdriver is not used for a single type of screw. When and if a heuristic is useful depends on how successful it is in a certain environmental structures. A successful heuristic pairing of mind and environment is then adapted over time. The adaptation takes place through individual reinforcement learning, social learning and evolutionary learning. The latter means that some heuristics come to us instinctively without us having to be raised or educated to use them.

According to Gigerenzer there are ten fundamental heuristics, which I will dive deeper into in Chapter III. For now, I will simply list them as follows: The recognition heuristic (give higher value to recognized alternatives), the fluency heuristic (go with the faster alternative), take the best (go through cues until one alternative seems better in this regard), tally (do not estimate weight of arguments but amount of arguments in favor/against), satisficing (pick the first alternative that is good enough for your aspiration), 1/N , the default heuristic (both explained above), tit-for-tat (cooperate first, then copy their attitude), imitate the majority and imitate the most successful (these last two are self-explanatory). These heuristics are evolved tools based on memory and imitation capacities in our minds. We are born with the capacity, and they are nurtured into capabilities. As such, they are tools for survival which our species have gained over time, both in an evolutionary sense and within the growth of a single individual.

As I mentioned above this can be likened to, and be a potential bypass for, Penrose's (1994) argument that developing a cognitive capacity in terms of a general understanding of one's environment makes more evolutionary sense than developing innate sophisticated mathematics. Going back to Penrose's argument for understanding non-stopping algorithms, however, how exactly

does Heuristics help here? One could imagine the fluency heuristic to be a candidate. This would be akin to the gauging of stopping after a certain amount of time. What we have here instead is the human recognizing that a solution is far from being found and thus takes the quicker alternative of assuming there is none (i.e. it is non-stopping). This can be supplemented by an initial check if there is a stated end. $2^{2^{\wedge}65536}$ 1's being printed at the speed of a 1 every 10^{-43} seconds is an obviously finite *task*, albeit a very long one. It is intuitively finite and it seems the only support for this argument in relation to computers and time-based gauging algorithms would be that the computer in question did not understand or perceive the finiteness of the task. It would be unlikely that the heuristic for judging an algorithm to be non-stopping (if a heuristic or a combination of heuristics are used) would be something very specific to exactly that task. A heuristic that is all about non-stopping algorithms would have quite an incredulous evolutionary tale, judging by the impracticality of it and the lack of mathematical problems relating to survival. It could be adaptive of course, something learned from mathematical understanding and elementary mathematical training. I would like to point out here that there is a possible discrepancy in the nature between '$2^{2^{\wedge}65536}$ 1's being printed at the speed of a 1 every 10^{-43} seconds' and 'find an odd natural number that is the sum of two even natural numbers'. The first is very much a mechanical task. It is an action with a beginning and an end. The latter is indeed also a task, but I do not believe that most people would approach it as such. It would be much easier to simply solve the latter through the axiom of 'two even numbers never add up to an odd number'. This axiom or rule is something that we perceive as true. A person who has never pondered upon this mathematical fact before might be told one day that 'two even numbers never sum up to an odd number'. If he is prone to critical thinking and not just someone who takes everything he is told as truth at face-value, this person is likely to test the axiom a few times in his mind. After a few steps it would not be unreasonable to imagine that he would apply a mix of the 'take the best' and tally heuristics to come to the conclusion that this is indeed true. A person who is not instructed might discover the axiom on his own (and in the same way) by testing a *notion* that this is the case, the notion or suspicion being gained from lack of experience of cases that proves it wrong. Once an axiom is 'gained' it can simply be used as information for the future. What this would mean is that once the axiom is in our knowledge bank we don't need to go back and rediscover that this is a fact over and over again. It could thus be argued that 'find an odd natural number that is the sum of two even natural numbers' is not a task in the same respect as a very long algorithm simply because we either discover the axiom heuristically or simply never attempt the task because we already have the answer to it in our knowledge base. We *understand* based on previous knowledge that it is the case.

1.5 - Quick Conclusions

So after looking at the worries currently surrounding the modern state of the Computational Theory of Mind I want to reiterate what I think is important for the future of the theory. Even though it might seem tempting to some to head straight into non-computability once certain arguments about the flaws in hard logic and pure inference systems start to (rightly so) re-emerge from the early days of computation, there are still too many good alternative explanations available that maintain the core of what CTM is all about, as well as keeping to its strengths. Gigerenzer's heuristics (1999, 2008) have empirical backing, make sense from an evolutionary perspective and also seem to meld quite well with the modular form of CTM. What really creates the problem of non-computability to start with is that the seemingly most prevalent understanding of CTM is antiquated and not fair to the research that has been done in the past decades (Chalmers, 1995a).

As for the EEEE theories, I do not think that CTM supporters should be afraid to give these theories credit where it is due. Embedded, in particular, is a very helpful way of looking at how we function cognitively with the world as part of our toolbox. It brings, in most part, sensible explanations that I would argue actually make improvements upon CTM rather than abandoning it. It is, however, the abandonment of CTM that one has to be careful with when it comes to EEEE. Embodied, Embedded, Enacted and Extended theories of cognition are very good at modelling active agents navigating a world, especially in fields such as robotics. However, if left at the stage of robotics the results we gain will most likely stay at being very impressive reconstructions of *simple* cognitive systems. By contrast, the strength and focus of CTM has always been that of *higher* cognitive systems, being weak at producing navigators but more successful in creating *thinkers*. Just as Rowlands (2010) I think that a form of amalgamation is a good possible solution of gaining the best of both worlds. However, in contrast to Rowlands I think the key in this endeavour would be to bring heuristics and embedded cognition into classic CTM. Why not place the thinker on the shoulders of the navigator?

The real threat against CTM comes from the attack on representational content and representation-based symbol manipulation. In future chapters, I will highlight these attacks in the form of Hutto (2011a, b, 2013a, b, & Myin 2012) and Noë (2004, 2009), both of whom champion a radicalized enactivist approach to cognition. Exactly what this means is something I will go over in the very next chapter, where I will properly introduce the E-theories and describe how they open up new ways to think about the mind where we previously may have seen limitations. Furthermore, I will discuss the differentiation between *passive* and *active* cognition, and how these different ways of framing perception and mind-world interaction changes the cognitive landscape, proposing a mind

that breaks free from the input-output sandwich and transforms into something predictive that expects and meets the world instead of waiting for the world to come to it. Predictive processing will be an important part of this thesis and the main weapon that I will seek to use against the enactivist argument against CTM. Ultimately this thesis is a story of compatibility and modernization. In the following chapters all the way to the end I will maintain the argument that CTM is, in its foundation, the best model for the human mind, insofar as the mind is an abstraction of the processes taking place in our brain, and the interaction that these processes have with our body and the outside world. I will defend mental representations as key to computation, as well as the best way to explain how we form concepts and models of things in the world, both concrete and abstract. I will also explore further the idea that our mind is divided into two types of systems, à la Kahneman (2011), and that these systems are built on the idea of efficiency and frugality. This in particular will be fully explained in Chapter IV.

Chapter II - EEEE

In the previous chapter we introduced the Computational Theory of Mind, a theory that proposes that our mind functions in many ways like a computer, with internal symbols and a processing language to parse them. We also looked at one line of reasoning against CTM; the idea that some areas of our cognition are simply non-computable, and that the mind cannot be realised in a complete computational system. I challenged the worries of Penrose (1994) regarding how our capacity for 'understanding' was seemingly beyond computational power, primarily by pointing to how the usage of modular sub-systems or heuristic fail-safes could equally explain why we don't fall into the Gödelian trap of incompleteness or endless mathematical thought loops. In this chapter, we will explore a second line of reasoning against CTM, found at the more radical spectrum of the collection of theories abbreviated EEEE; Embodied, Embedded, Enacted and Extended. What we will see below is that these four theories bring new ways to think about cognition, moving beyond the bounds of classical CTM. What all of these theories have in common is that they all seek to take cognition beyond just the brain in some form or another, some even taking the mind itself *Out of Our Heads* (Noë, 2009). This line of criticism that we're now turning to, taking the form of radical enactivism, is the one that will stick throughout the rest of the thesis. It is the one that I deem most important to save CTM from, as it involves an attack on the core of what makes CTM tick: the idea of mental representations and information processing.

2.0 –EEEE and Cartesianism, a short introduction.

What makes the EEEE family of theories unique is that they all in some ways seek to incorporate more than just our brain in our cognitive processes. This can take the form of off-loading certain cognitive responsibilities onto the world, to thinking with our bodies, to at its most extreme actually extending what we consider part of our mind to incorporate even tools external to our brain and body. In this chapter, I will seek to properly introduce each of these four E-theories, putting the most emphasis on Extended and Enacted cognition. The reason for this is that Extended cognition is based on a thought experiment, Otto's notebook (Clark & Chalmers, 1998), that will be very important for discussion later on in the thesis. Extended cognition carries with it a very unique discussion of where we are supposed to draw the borders of the mind, and once we reach the later parts of the thesis, these boundaries will be an important way to differentiate between stances.

As for Enacted cognition, it will prove to be the most important aspect of all the E-theories. More so than the other E-theories, enactivism is supported by a strong anti-Cartesian sentiment. Cartesianism is a term sometimes used to describe theories like CTM, and what is really being targeted by the phrase is a kind of brain-centric internalism. For example, Noë writes that "seventeenth-century philosopher René Descartes … held that each of us is identical to an interior something whose essence is consciousness; each of us, really, is an internal *res cogitans,* or thinking thing." (Noë, 2009, p.7) Furthermore he states that while Descartes believed the thinking thing to be immaterial, and in contrast most 'traditional' modern philosophers and scientists believe cognition to be materially situated in the brain, "[i]t is precisely on this point, and this point only, really, that today's neuroscientist breaks with tradition." (Noë, 2009, p.6) Thus, the term Cartesianism comes from the argument that traditional theories of cognition still hold to Descartes' ideas that we are an internal, secluded thinking thing, the only true difference being that we have moved away from substance dualism. This comparison to Cartesian ideas is further reinforced by the accusation that CTM and similar views propose a form of Cartesian Theatre when it comes to vision (Dennett, 1991); the idea that vision is an event that takes place in some specific locale of the brain, with a tiny (metaphorical) person sitting inside of our mind as the audience. When I mentioned in the previous chapter philosophers making an active move away from CTM, enactivism is a prime example of such a motivated move, due to their disapproval of these perceived Cartesian elements of the theory. I disagree with this assessment made by Noë, and I will hope to prove so by showing that computationalism is in fact capable of compatibility with many E-theory elements. In that way, while I will still champion a brain-centric view of the mind, perhaps even 'secluded' in the sense that we are our brains, it is still capable of reaching out beyond these boundaries and into our bodies and the world.

2.1 - Embedded & Embodied cognition.

Embedded and Embodied cognition are perhaps the easiest of the E-theories to digest, in that they are mostly non-controversial ways of looking at cognition as something more than just internalized brain-processes. They key feature for both of these types of cognition is off-loading. What I mean by off-loading is that certain amounts of cognitive responsibility or effort are taken off the brain's shoulders and instead placed somewhere else, effectively reducing the amount of effort needed to work out certain cognitive tasks. Additionally, off-loading can lead to scaffolding, enabling us to reach a higher level of cognitive competence than what would have been possible with 'brain power' alone. Even though I refer to Embedded, Embodied, Extended and Enacted cognition as a collective

set of 'theories', the former two have a special relationship to the latter two in that they work as foundational frameworks upon which extended and enacted cognition can be built. Extended cognition is very reliant on features found in embedded cognition, just as enactivism relies on Embodied processes.

Embodied cognition presents the idea that we use our bodies in ways to aid cognition. For example, when we speak we often use gestures to help express ourselves. However, even when on the phone, these bodily gestures continue, despite there being no direct communicative benefit to them. For example, David McNeill (1992) suggests that bodily gestures have a great impact not only on our capability to communicate through the use of gestured symbols, but also actually facilitates verbal expression as well as our thoughts. In this manner, gesturing or pacing about in a room can aid us in our thinking. It is often depicted in cartoons that when someone is thinking very hard they scratch their head or pace about in a circle. This depiction may not be just a way to convey to an audience the existence of certain internal states, but in fact is also a behavioural tool to facilitate these very same internal states. Aside from this, embodied cognition is a crucial part in certain new theories of perception, including enactivism and prediction error minimization (a part of predictive processing which we will get to in Chapter V). By moving our bodies we change the perspective through which we perceive the world. This may seem like an obvious statement but through this simplicity can also be found quite remarkable ways in which we off-load cognition into bodily motion. A baseball player positioning himself to catch a falling ball will keep his eye on said ball, moving along with its trajectory and increase and decrease his speed depending on the visual feedback relating to the relative change in distance between his position and the ball's. In this way, the player can, through means of bodily action, calculate and predict where the ball will land, while at the same time making sure he is there to catch it. There is a distinction to be made here though between what could be called weak and strong embodied cognition. In the weak sense, cognition can be embodied in the sense that it is sometimes facilitated and aided by bodily action. Here, bodily movement aids cognition, but is not a crucial necessity: If I wanted to know what an object looked like upside down I could either turn a mental image of the object in my mind, or I could simply use my hands to turn it physically. Both of these actions have the same end results cognitively speaking, but one is less cognitively taxing. In the case of strong embodied cognition, these bodily actions are *necessary* for the cognitive function, and would otherwise not result in anything. This stronger claim is embraced by enactivism, where the central claim is that we *enact* the world around us, and thus create perception through action. We will see more of this later on.

Embedded cognition, equally, is about off-loading our cognition onto the world. However, for embedded cognition this is primarily about how we make use of external tools to facilitate thought.

These kinds of tools involve things such as keeping notebooks and memos in smartphones to facilitate memory, or to make use of pen and paper to write down mathematical work as we calculate, to avoid having to keep it all 'in the head' and risk losing track. Calculators are an even starker example of how we can achieve tasks such as calculating very large numbers without actually taxing our own brains all that much.

2.2 – Extended: Is supersizing an option?

In *Supersizing the Mind* (2008), Andy Clark defends his theory of extended cognition against criticisms directed towards him as a result of his initial works published on this theory, most notably *Being There: Putting Brain, Body and World Together Again* (1997) and *Natural-Born Cyborgs* (2003). In doing this, he also attempts to show why this view holds up better than its opponent, or counterpart; brain-bound cognition.

The main difference between brain-bound cognition and extended cognition is that the former holds the view that, paraphrasing from Clark, cognition ends at the skull and skin of the cranium. That is to say, it is something exclusive to the brain. Extended cognition wants to say that cognition is not bound to the brain in this way, and that our minds can bleed out into our bodies and, even further, the world we interact with. Clark describes brainbound cognition as an input-output sandwich, where cognition is the filling in the middle, never going out of bounds. Even though both views would allow for tools to be used in order to aid cognition, it is only in Extended that the tools can become an integral part of the cognition itself.

Clark (2008) compares this to how babies act clumsily and experiment with the functions of their bodies as they grow and learn how to manoeuvre in the world, such is also the case for extended uses of prosthetics, extra limbs, or simply brooms/rakes. Evidence suggest extensive use of a tool that extends the body in any way (say a broom) eventually causes the nervous system to act as if the arm of the person holding the broom had actually been extended by the length of the broom itself.[8] I.e. the body is not acting as if the arm is manipulating the broom manipulating the world, but rather it skips the middle man and treats the broom as an extended limb manipulating the world.

Moving on to language, Clark presents a case in Boysen et al. (1996) of a chimpanzee who is presented with two bowls of treats. The chimp Sheba has been given training in numeric symbols and can understand the meaning of numbers. Her choice between the bowls is set up in such a way that

[8] For example, when poking something with a tool we feel the sensation as if it was at the tip of the tool and not in the hand holding it (Yamamoto & Kitazawa, 2001).

when she points at one bowl, the chimp Sarah next to her gets it. The two bowls have obviously different amounts of treats in them, and even though Sheba knows after repeated choices that the picked bowl will go to Sarah and visibly hates getting the smaller bowl, she keeps picking the big one as if her instincts were in the way. This is where the scaffolding comes in. By using covers over the bowls, and numerals to indicate the number of treats inside, Sheba can suddenly pick the smaller bowl and thus successfully receive the big bowl as she wanted. The point here being that by interacting with symbols representing the objects (numerals) instead of the objects themselves (the treats), the instinctive prompts never came into play, but the cognition involved in solving the rather simple puzzle could carry through. As such, language was used as scaffolding to overcome cognitive limitations.

The most important argument from Clark involves the offloading of cognitive pressure onto the world. In their paper 'The Extended Mind' (1998) Andy Clark and David Chalmers present a thought experiment called "Otto's notebook" and Chalmers makes a reference to a similar thought experiment involving his iPhone in the foreword to *Supersizing the Mind* (2008). Chalmers' case is a lot milder than the imaginary case of Otto, but they work on the same premise of taking notes to store information and memory in other places than the brain itself. Chalmers here mentions things like writing down his favourite dishes at a restaurant for later reference. Chalmers here would have no problem remembering these dishes, but since he can quickly note them down on his iPhone that he's always carrying with him, he can choose not to bother spending time or effort remembering if he found something he liked, and instead simply store that information on the phone. In the case of Otto, however, the thought experiment is built up such that Otto often forgets or has difficulties remembering things due to Alzheimer's. He thus has a notebook where he notes down everything noteworthy, such as addresses. When hearing about an exhibition at a museum that he'd like to go to, he reflexively reaches for the notebook to find out where the museum is. In these cases, the notebook and iPhone serve as means for offloading the cognitive strain on the brain by letting the information be stored elsewhere. However, Clark also wants to claim that this is where mind can bleed into the world outside of the skull. He claims that the notebook is part of Otto's mind (just as Chalmers, perhaps jokingly, claims in the foreword that his iPhone has become part of his). Clark & Chalmers (1998) have four criteria for deciding if an object can be part of a person's mind in terms of memory/knowledge:

1. It is frequently invoked and its availability is reliable (Otto always carries the notebook and often consults it).

2. It is not subject to scrutiny, at least not most of the time. There thus has to be a form of trust and endorsement about the contents of the book. Googling something on an iPhone does thus not constitute as a source of personal knowledge.

3. The information should be easily accessible.

4. The information noted down must have been consciously endorsed sometime in the past. It is in this way connected to what the agent has thought before (like Chalmers noting down the dish he liked, or Otto noting down something before his brain forgets).

It is also worth noting that the argument for Otto's notebook to be part of his mind comes from a functionalist[9] mindset and is based in what they call the 'parity principle'. The exact nature of the parity principle is clearly stated as: "If, as we confront some task, a part of the world functions as a process which, *were it done in the head,* we would have no hesitation in recognizing as part of the cognitive process, then that part of the world *is* (so we claim) part of the cognitive process." (Clark & Chalmers, 1998, p.8) It is not the location or substance of the process that matters (we could have a brain made of cheese[10] or, more realistically, a digital brain implant memory stick for extra storage), as long as the functional states remain the same across comparative cases, so should the identity of the mental states.

So far, it would seem that the core of Extended lies in increasing, or 'supersizing', the mind's capacities by moving cognition away from the brain. It thus leads to a type of frugality of brain power. Frugality and tractability are topics that often come up when it comes to brain and mind in contemporary discussions. Gigerenzer wants to explain frugality by relying on heuristics. Penrose explains frugality by pointing to direct conscious understanding. Clark, in contrast, wants frugality explained by relieving the brain of some of its responsibility, outsourcing cognition so to say. It is a clear connector between these two views that, on the surface, otherwise do not seem to have that much common ground.

Frugal theories of mind are very attractive from a naturalist standpoint. As stressed by Penrose (1996) and implied in Gigerenzer (1999) they make sense in a biological and evolutionary fashion, avoiding the needless complexity or massive tractability that is usually hard to explain in said fashions. Why rely on hard rationality when the work load can be lessened (and sometimes made more achievable) through rules of thumb (Gigerenzer, 1999), exploitative representation or epistemic action? (Clark, 2008, 2016) Exploitative representation is a term describing system off-load

[9] Functionalism is the idea that mental states are identified primarily by their functional role.
[10] A popular functionalist quote being "We could be made of Swiss cheese and it wouldn't matter" (Putnam, 1975, p.291)

by tracking information in an alternative way to storing it, in this case writing things down in a notebook or doing mathematical work on a sheet of paper.

Adams and Aizawa have delivered criticisms against Clark's theory, one involving a rather crude why-did-the-chicken-cross-the-road type joke: “Why did the pencil think that 2 + 2 = 4? Clark’s Answer: Because it was coupled to the mathematician.” (Adams and Aizawa, 2001, p.1) They use this joke to highlight their claim that Clark has misunderstood what it means to be a cognitive agent. Clark answers to this that one would not expect a neuron to have any sort of cognitive agency by virtue of being part of a cognitive system, so why would the pen be? I am inclined to agree with Clark in this case. A large group of sailors is not necessarily a group of large sailors. There is no reason to believe that a part of a cognitive system, especially not a temporarily coupled one, is expected to carry the same qualities (like cognitive agency) as the system it’s a building block of. Clark is willing to extend this to groups of neurons and even whole parietal lobes. He also confirms in later sections that that this applies in the same fashion to limbs, artificial or not, and notebooks that qualify through Clark’s aforementioned criteria.

Agency here (of the cognitive kind), from Clark’s point of view, seems like a fleeting thing that attaches itself to whatever is included in its cognitive system at the time that thing is coupled to the agent. However, he stresses the role of the central organism itself (the brain) as an anchor of sorts, a core part of the mind that never ceases to be part of the cognitive agency. Indeed, Clark has thus far not argued for anything suggesting that the mind would not have a fundamental and necessary core. In fact, a lot of Extended relies on the old mind, the input-output sandwich, to be there, only the filling can spill out so to say. The brain can thus be assumed to still be a necessary part of the cognitive agency. Clark is even happy to see this as a computational approach to the mind.

But what about the agent, as in the person? This worry was the reason that I added “of the cognitive kind” earlier. Clark is happy to call a pen or a notebook part of a person’s mind, if it fulfils the criteria for being part of a cognitive system. Would we, however, agree that the notebook is then part of the person in the same fashion? If yes, then the problem seems bigger. It seems against our intuitions about personal identity to claim Otto’s notebook is part of Otto, and far easier to conceive of the notebook as a tool used to improve Otto’s cognition (more than just offloading, the notebook makes Otto achieve to him otherwise unachievable capacities). But if we say no, then, this could have the implication of a split between identity of person and identity of mind – that is, there would be parts of Otto’s mind that wasn’t part of Otto the person. Perhaps the source of the worry of this split is that memories are closer to our idea of personhood, thus giving greater cause for criticism when they seem not to be connected. Memories are generally private, while text in a notebook is not. As I

mentioned, I'm undecided about whether or not Otto's notebook could be part of his mind. Clark (2008) seems in his final step to convince the reader to rely on the reader's intuition that this is the case (whether or not these intuitions can be used as proof in either direction is another hornet's nest I'm not going to disturb here). As such, a reader can become comfortable with the thought that Otto's mind extends into his notebook, but is then faced with the issue of whether the notebook is part of Otto. When the answer is a no, we are suddenly faced with the problem of a divide between identity and mind. The notebook is part of Otto's mind, but not part of Otto, so part of Otto's mind is not Otto. This, I think, is a hard position to defend and it keeps me from wanting to embrace Clark's conclusion.

2.3 – Objections to extending personal identity.

Another criticism is that of memory retrieval in the case of Otto. It is possible, as noted by Martin Davies in a personal comment to Clark, that Otto can misread his notebook. The criticism here is that the notebook would "seem more like a perceived part of the external world than an aspect of the agent." (Clark, 2008, p.100) Clark counters by saying that a person who retrieves memory from the brain could also end up in error, not because of the memory being wrong, but by something disturbing the retrieval process thus causing a 'misread' of the brain-bound memory. There is also the issue of the notebook being tampered with but, equally, a brain could possibly be tampered with by a surgeon who has greater knowledge about brains than we currently have at present and who can efficiently rewrite memories. For more contemporary considerations, imagine brainwashing or hypnotism. Could those be ways to tamper with memories in the brain? In the end, Clark holds firmly that the parity principle holds for misreading memory, and I agree that it does. However, I do think the focus on memory causes problems (as mentioned above) that wouldn't necessarily be there in more temporary couplings. Doing mathematical work on a sheet of paper or utilizing other epistemic tools to aid fleeting cognitive tasks has a greater pull to it than the cases involving memory retrieval. Additionally, when it comes to the idea of memory retrieval in a case where the person has forgotten the information (as Otto has due to Alzheimer's) it is not clear that it is in fact memory access that is taking place, rather than Otto finding out what he believed before he forgot it thus, in a way, merely accessing the past in the way one can retrace one's footsteps or look at pictures of one's younger self. In the following section I will present a strong criticism coming from Eric T. Olson that is directed at the implications of identity in the Extended theory of mind.

Olson (2011) voices similar worries to those above, making the claim that the extended theory of mind put forth by Clark & Chalmers (1998) either results in us having beliefs that ultimately are not

part of us, or a dualism where organisms are incapable of having any mental states at all. In this paper I will argue that Olson's claim relies on assumptions about the nature of personal identity and our relation to our bodies that extended mind theorists, along with many other philosophers of mind, never endorsed in the first place.

Olson's primary question is whether extended cognition implies an extended self as well. He notes that Clark & Chalmers briefly mention this notion at the end of their paper 'The Extended Mind', and that their response to this is that "It seems so." (Clark & Chalmers, 1998 p.18) Clark & Chalmers' argument is that some of Otto's beliefs are thus located in the notebook, and that his cognition extends into it. As such, there is a location outside of Otto's skull and skin where part of Otto's cognition is taking place. This leads to Clark & Chalmers' conclusion that Otto's *mind* can extend out of the body and into the world. Olson presents three questions that he'd like to explore about the relation between the extended mind and the extended self:

1) What does the thesis say other than that Otto extends beyond his skin? What is the emergent principle?

2) Would the extended self thesis actually follow from the extended mind thesis?

3) If the extended self thesis indeed followed, what would the result be if it was true?

To answer these questions, Olson (2011) addresses just what it is that is meant with 'external' in the case of Otto. Otto's notebook is 'external' in that it is outside of his skin. This is a simple interpretation that also makes it obvious that the cognition taking place is different in its phenomenology to 'internal' thinking. It requires reading and is more akin to information searching that memory triggering. A comparison is made between this and the case of a man Dave who has a USB port implanted in his brain for readily available information on an external memory stick. In this case the phenomenology would be the same as if we retrieved memory in the head, but the information external. *If* proponents of extended cognition would not qualify that as extended cognition, Olson argues, then extending beyond the body (like the memory stored by Dave in the stick) is not sufficient for being called 'external'. It would seem that the phenomenology of the cognition, in that case, must also be that of retrieving from the outside. I would argue that this is not a worry for extended cognition, because the memory stick fulfils all four criteria as long as it is (1) frequently used, (2) trusted, (3) accessible and (4) containing information (beliefs, memories) that was previously endorsed. The only reason for not accepting a memory stick in this way would be if it contained someone else's memories, and not Dave's, since that would go against (4). In reverse, taking Dave's memory stick away from him and placing it in another person's brain would break (1)

and (3), thus Dave's mind cannot qualify as ending up in someone else's brain either. Outside of that, a memory stick is an external object that is brought into the cognitive fold, so to speak, and in fact does a much clearer job of qualifying for Clark & Chalmers' (1998) criteria.

Olson also comes up with a second case, Jenny. Jenny has the same forgetfulness as Otto and utilises a notebook. However, she uses this notebook to get the information into her photographic memory. Once that has taken place, she is done with the notebook and can simply read the pages in her mind. Almost like in the previous case, where Olson made an argument for the consequences that would follow if the thought experiment was rejected, he here makes a claim for the consequence of accepting Jenny's memory as extended. If Jenny's memory is external, then it would seem that extended cognition doesn't have being external as a necessary condition. The idea Olson is getting at here is that Jenny doesn't know where, for example, the museum is located, but reads her internal notes to retrieve this information every time she needs to get there. In this way, it is similar to Otto's notebook, but internal. I fail, however, to see why we would claim that Jenny's mind is extended in this way. If she had a photographic memory that she could retrieve at will, why would we claim that she is forgetful, and why would we claim it to be any different from regular memory? I do not think this is a very strong argument.

However, even though I'd argue the Jenny approach fails, it ultimately leads into a stronger argument. What Olson wishes to do is to present the idea that external is not about location, but about being external to the subject. An external cognitive state is 'external' in that it is external to the subject of which it is a part. This leads to the argument that will follow as the basis for his rejection of the extended case: If extending into the notebook requires the 'external' cognition to be external to Otto, then it cannot follow that Otto extends into the notebook, for then the cognition is not external anymore. As such, part of Otto's cognition is necessarily not part of Otto. To Olson, this creates a problem that the extended mind theorists cannot get rid of, that extending one's cognition necessarily requires the cognition to be external to the subject, thus the subject either cannot extend its cognition, or the cognition is disconnected from the subject. For the information to remain external, we cannot extend ourselves, so we are forced into the split where cognition extends but not our identity. The beliefs stored in the notebook are thus either not beliefs at all, or not Otto's beliefs. Olson presents a variety of ways to save the extended mind thesis, only to debunk them. Thus, he in the end concludes that while it may be possible for beliefs and other cognition to briefly extend into the world, there is no case for thinking that we ourselves are extended. I would like to put forth the argument that Olson is overplaying his use of 'extended' and 'external' in order to make this problem appear.

Otto's notebook aside, the extended cognition presented by Clark & Chalmers (1998) seems like a case more akin to active 'extension' rather than things being 'external'. In fact, Clark (2008) puts no emphasis on the property of being external when it comes to extended cognition. Take the imaginary case of Emma who is playing Tetris. She uses the epistemic action in the form of zoid measuring techniques (Clark, 2008). Clark sets this example in reference to the work of Kirsh & Maglio (1994). By doing this, Clark argues, her cognition extends out into the game field where the measuring is taking place. In the perspective of her state before this feat, the game field is external. However, it is not obvious to me that you should treat the field as external once the cognition has extended into it. The question then becomes "is it not strange to say that Emma herself extends into the game field?" My answer to that would be that it is not much stranger than to say that any other tool can become part of us. In an experiment reviewed by Maravita & Iriki (2004), Japanese macaques using rakes as tools to gather food would have modal neurons in their intraparietal cortex respond to expanded somatosensory information. Before using the tools, these bimodal neurons would only pick up visual stimuli from around the arm of the macaque, but after just a few minutes of use this stimuli pickup would expand to include the rake as well. It would thus seem that the tool was integrated into the so-called 'body schema', a term in neurology meaning a subject's unconscious classification of its bodily boundaries. Back to the case of the Tetris field, even though it would be harder to get positive somatosensory results for zoids one could still imagine a similar notion of 'self' integration going on in a temporary manner, effectively extending Emma for a short period of time. There is no problem here of Y not being part of X in virtue of Y necessarily being external to X. Either we have embedded cognition, in which case Emma off-loads into the Tetris field in an empirical way, or we have extended cognition where Emma extends mind and self into the field.

I would like to make clear that I think Clark's theorem has much more appeal in its fleeting active form, the one akin to a stronger embedded cognition, than in the memory based case of Otto. It is strange that the Otto case is where the most speculation and refutation is taking place, as its treatment of memory makes it the weakest perspective. The strength in the Clark & Chalmers theorem lies in the functionalistic approach to cognitive enhancers. Otto, however, faces the difficulty of memory loss and retrieval. As Bernard Williams (1973) states, removing information from a brain and later reinserting it does not sound like a case of a person forgetting and later remembering (William's case handles a thought experiment about removing memories from a brain and storing them externally for brain repair. In terms of Otto, he'd be placing the information as a backup for when he inevitably loses it). While Williams' case assumes for the moment that our memory works like the kind of thing you can store and move around, Otto's case makes few assumptions about the nature of your inner memory workings. This means that we aren't even

certain that the way Otto retrieves his memories has sufficient functional similarity to how a less forgetful person does it internally. Thus my conclusion that the Otto case is the most questionable point of Clark & Chalmers' theorem.

So then, if we could leave notebooks behind for the moment and retrace our steps to the point I was making concerning embedded-type extended cognition, the more fluid way to treat our barriers of active cognition (i.e. processes taking place, not stored information), we will be running straight into another part of Olson's refutation. Olson wants to equate Otto with the organism Otto. If Otto extends his cognition, so does the organism Otto extend accordingly, for Otto cannot be anywhere where his organism is not without invoking a form of substance dualism. The substance dualism he has in mind is a kind of Lockean dualism where Otto is divided into Otto's extendable mind and Otto's *life* ('life' being the self-maintaining organism, within which there's his immune system and other crucial bodily functions). As Sydney Shoemaker (1999) points out in his review of Olson's *The Human Animal* (1997), Olson rejects the psychological approach to personal identity and instead embraces an animalistic[11] approach that focuses on a person as being his/her body. The argument from Olson here is that either Otto's self would be his physical body or Otto is this dualistic entity where the body is, at most, just a part (although Olson wants to resist this notion as well). In the case of the former, nothing of Otto can extend for whatever extends cannot be Otto, and an organism cannot 'get bigger' by using a notebook or a Tetris game field. For the latter, Olson's main reason for refusing a dualistic approach is the following implication that organisms, in general, would have no mental states. If there is a divide, it will be of mind and body. Since Olson is firm in that mind = all mental states, Otto will have a cognitive part separated from an organism part. The organism is as such only connected to a cognitive entity, and they both constitute Otto. Organisms that lack the mental capacities that make them persons with minds thus have no cognitive states (we can look at lesser creatures such as rodents and insects for possible examples that Olson would point to). Olson then ponders how anyone would say that an organism, despite its brain, would somehow end up cognitively diminished without an extendable and immaterial 'self' or soul. The strike of this is supposed to be that the brain is such an obvious instrument of thought that the notion of calling it insignificant to cognition would be absurd.

Something that immediately springs to mind here is the issue of physical augmentation and prosthetics. Olson, being an animalist, makes the claim that the core of a personal identity lies in the core organism, including its brain stem, cognitive centre and motor functions as core physical features, as well as the 'life' built around these, keeping them functioning and maintained. Is it then

[11] Animalism is the view that for all questions as to the nature and metaphysical state of human beings, the answer is that we are animals. See Carter (1990) and Olson (1997).

Olson's claim that an organism is only its 'original' parts, without any hope of replacement? He certainly believes that a person must be able to survive natural cell replacement (Olson, 1997). But what then of transplanted parts? Arguably organic transplants would fit better than mechanical parts in Olson's 'life' view of the organism. But even if that would be granted, what then of possibly future prosthetics that include more organic functionalities? I would imagine it not impossible that in some distant future we can have new manufactured arms with functioning blood circulation and muscle growth. If self-sufficiency and organic sustenance are essential parts of an organisms 'life' features, then even non-natural parts are only excluded by limits in technological advancements as far as we know today. If the exact same functionality can be reproduced, even Lockean 'life' functionality, is the self still limited to the parts that it was born with?

On a final note I would like to raise a concern of my own against the extended theory of mind, this one aimed at what I mentioned above to be the stronger part of the theorem. If we look at the case of Emma's extended cognition in Tetris, it is easy to draw similarities between this and the solving of a Rubik's cube. Most people who solve Rubik's cubes do not use raw brainpower and painstaking puzzle solving abilities to slowly make the puzzle come together. Rather, most people use algorithms that can be much more easily learned and reliably used. These algorithms are the cognitive work of people in the past, packed up into a neat sequence of moves and passed onto future generations of solvers as tools to be used in appropriate situations to solve a cube without risk of failure. An algorithm is a set of movements, done in a specific order, in order to move one or more squares of position A into desired position B while at the same time not disturbing squares that have already been solved. This is an obvious case of embedded cognition. The brain off-loads a massive amount of work into these algorithms, solving a task otherwise requiring a great amount of work, while basically holding a cheat sheet in their mind. It still requires mental work, no doubt about that. Using algorithms effectively to produce fast solutions is still a matter of skill. However, the level of work is marginal to the amount required to solve a cube from scratch. So while a Rubik's cube case holds for *embedded* cognition, it would be strange to say that it holds for *extended* cognition. It would feel odd to say that one's mind extended into the cube while employing the algorithm, that this manipulation of sequences made the cube become part of the subject for a brief moment of time. Is there an important difference between the Tetris and Rubik cases that makes extended cognition seem plausible for one but not the other? One could argue that the two are indeed similar and should be classified similarly, and that it is the conclusion that the Rubik case doesn't apply that is the mistake. Solving a Rubik's cube with algorithms is not only a mental task. You require knowledge of the algorithms and their application in order to use them, just as you need the knowledge of the epistemic zoid measuring strategies in Tetris in order to apply them properly. In addition, just as zoid

measuring knowledge is separate from the zoid measuring *action*, so are the Rubik algorithms separate from the sequencing *action*. The action becomes the important part, and it is correct that once the algorithms become natural to a solver of the Rubik's cube, the movements come without thinking. The solution becomes an entirely empirical task. The subject will look at the cube, seeing the positions of squares and applying the correct sequences through muscle-memory. The cognition can at that stage be said to take place in the cube as much as it does in the subject's brain. The brain would call the shots while the hands carry out the solving. For each step, this (possibly) extended cognitive system would initiate with the brain gaining visual input and deciding upon an algorithm. Muscle memory in the hands would carry out the sequences and on the cube the output would display, the output being the new position of squares. The output would then become the new input until a solution has been reached. Much in the same way, the Tetris case starts with visual input of a zoid dropping. The brain judges the field and calls for measurement. The hands on the controller carry out the commands necessary of epistemic measurement on the game field. This measuring is then visual output leading to a second input, which is then quickly judged and the zoid is placed accordingly. Each step in the Tetris case takes two input-output circuits, while the Rubik case only has one circuit per step. This, however, ends up being the only difference between the two. With this example, I've attempted to make a clearer and stronger case as to why both cases off-load as well as have the potential to lead into cognitive extension with repeated use.

This can be likened to Rowlands' (2010) criterion for his theory of the *amalgamated* mind. Rowland's theory sets out to fuse the Embodied mind (i.e. that our mind is heavily dependent and shaped by our physical setup outside of the brain) and the Extended mind (as presented above). In contrast to Clark, Rowlands' version of extending the mind does not put mental states outside of the body. Otto's notes in the notebook are not Otto's mental states. Instead, they fulfil the functional role of making that information *available* rather than merely *present* (the information of Otto's past mental states, the 'remembering'). What is meant by this is that information that is *present* is information that is out there in the world, information that we can pick up and make use of in our thought processes. This information can be acted upon to be made *available* to us (the information on a map can be made available by picking up and turning the map, interpreting the information by reading it etc.). Both making information available and sequentially making use of that information is part of the cognitive process. What Rowlands wants to argue is that the coupling of external information and mind is done through an extended cognitive process that involves the environment (external information), the enactment upon the environment (the manipulation of the information from *present* to *available*) and finally the thought processes taking place because of this. Arguably this final part takes place in the form of mental states that are *embodied* and not *extended*.

While Clark is happy to maintain brain centrality as something important to the way our mind works, Rowlands wants to do away with this 'Cartesianism' in cognitive science (Rowlands, 2010). Maybe because of this, I am more positively inclined towards Clark's extended mind than Rowlands'. Even so, however, I still argue that Clark wouldn't have to go any further than Embedded (that is, that we use the external world to offload in increase frugality of thought) to prove many of his best points presented in *Supersizing the Mind*. Not only does Extended bring on a whole new set of problems about personal identity, it also causes internal disagreements about what mental qualities actually do extend. Rowlands, for one, does not want mental states to extend but is quite happy for mental *processes* to do so. One might sidestep Otto's objections about personal identity this way but what kind of extended theory does one really have left at that point? That position seems to be more of a strong Embedded claim.

2.4 – Enactivism and experiential blindness.

Taking an even more extreme view than Clark, Alva Noë proposes that the mind is an enacted entity. That is, we interact with the world by acting upon it. This alone does not represent the extreme spectrum of the view, but what does is the follow-up claim that these actions constitute more of the mind than any brain functions do, as stated in his book *Out of Our Heads* (2009). Enactivism is a philosophical school of thought pertaining to cognitive science and perception that, as put forth by Alva Noë and J. Kevin O'Regan (2001), stresses the importance of sensorimotor knowledge and action in animal perception. The theory is a direct attack on David Marr's computational theory of vision (1982), claiming that the classic input-output view of vision is flawed and is not a plausible view of how perception works. In Noë's book *Action in Perception* (2004), the central claim is that there cannot be a passive and/or inert perceiver. In order for perception to happen, there must also be action and thought. My focus here is Noë's 'proof' of enactivism in perception through the existence of what he calls *experiential blindness*.

The classic computational view of vision (and perception as a whole) can be explained in broad terms as an input-output sandwich, just the way Andy Clark (2008) claims the mind is being viewed in general. Signals (in case of vision: retinal images) enter the brain through the sensory apparatus. Inside the brain, a representational image of the world is made and then presented as the subject of our perception. We then view the representation based on our perceived data and act upon it, thus signals go out from the brain to our body with the output; our reaction to the representation. The creation of the representation, as well as the thoughts and interpretations of said representation all goes on in the brain as a closed system independent on the world outside of it (aside from the

reliance on the input itself, otherwise the result would not be a data-based representation at all). In this way, the cognitive business of perceiving the outside world is sandwiched between the input that goes into the brain and the output that comes out of it.

In contrast to this, Noë claims that perception requires more than just input and internal representations to work. In his view, it is our acting upon the world that brings us the information we need to truly perceive the world. He is not against the idea of internal representations themselves, but rather the idea of a unified internal representation as the core of our perception. Noë likens this to "a blind person tap-tapping his or her way around a cluttered space, perceiving that space by touch, not all at once, but through time, by skillful probing and movement." (Noë, 2004, p.1) What we have here is thus a much more active view of perception, where experimenting and using one's body to prove the environment is what brings content to our perception. As mentioned before, internal representations are still formed in our mind as to what we see, smell, hear etc. but their content can only make sense as an extended perception through time and movement. Because of this, Noë also puts great weight on what he calls sensorimotor knowledge, a mixture of *understanding how* our perceptual input will be affected by the movement of our bodies (the change in visual perspective as we move our head etc.) and the *prediction and expectation of what* the incoming input will be depending on the actions we plan to perform. Noë thinks that the enactive approach is much more apparent when looking at all the other bodily senses besides vision, and that it is our fixation with vision that has lead us to sit so comfortably in the input-output model.

Noë hopes, however, that with his enactive approach he can turn the input-output model on its head by showing how the enactive approach indeed does apply for vision as well. In *Action in Perception* (2004) Noë early on makes a claim about blindness that he thinks provides good justification for believing in the enacted approach to perception. He hypothesises that there can be two kinds of blindness. One kind of blindness is the 'ordinary' one of lack of stimulus. Quite simply, this kind of blindness means that there is no information coming from the eyes to the brain, or no information/light reaching the eyes at all. In this way, the blind subject would be a person with either no eyes at all or with some sort of damage to their visual apparatus preventing them from gaining any visual stimulus. This type of blindness fits with the input-output model as well as the enacted approach, as it is simply the loss of one of the senses through deprivation of a working sensory tool. However, the second type of blindness that Noë suggests might exist is what he calls 'experiential blindness'. In this kind of blindness, the subject receives visual data but cannot make sense of it. Since the subject cannot make sense or make use of the information given to them, but still gains said information through a functioning visual apparatus, they are effectively blind while still

technically able to see. This, Noë argues, is a type of blindness that only makes sense through an enacted understanding of perception.

So does such experiential blindness exist? Noë cites cases involving cataract removal causing patients to lose content in visual input (they do not recognise that what they see is a face until it speaks etc.) as well as a case of special glasses inverting the left-right eye inputs causing the subjects to lose understanding about how to navigate the world around them. The latter is perhaps the more interesting of the cases. Receiving left-eye input in the right eye and right-eye input in the left eye causes a disruption in vision where objects are recognised, but due to the unexpected and unpredicted visual changes of turning one's head to look around, subjects report that the causal effect that navigating the world has on perspective - making objects appear to grow, shrink or move around as they move around in the world - seems to go haywire and appear completely unpredictable. This is in opposition to the common sense assumption that we would just notice things panning the opposite direction to what they should when we turn our heads right or left. In this way, the claimed 'blindness' is *caused* by input rather than the lack of it, since the input clashes with our sensorimotor knowledge. Noë admits that in this latter case, what the subject experiences is not *total* blindness, for the subject is still able to visually recognise many aspects of the world (they know a truck is a truck and that a wall is a wall). In terms of qualifying for experiential blindness, the former case of cataract removal seems to have the better claim, but even in that case Noë's assumption that the very existence of experiential blindness forces us into the enactive approach feels like a false dichotomy: What we seem to be forced to choose from is on the one hand a purely input-output based view of vision that cannot explain the existence of functionally impaired vision that is independent of faulty (or no) input. On the other hand we have Noë's enactivism where perception requires action, perspective mobility or thoughtfulness (preferably all three) for it to count as perceiving.

To answer why I think this dichotomy is false: I do not see why we would deny the importance that we use our bodies to aid in our perception. Our perceptual tools are, after all, biological tools that are part of a highly evolved and historically tested system. We use our tongues to taste, we reach out with our hands and feel to make sense of a shape (just poking a sphere does not tell us about its 'sphericalness') and we use our head and neck to fine tune our visual and auditory experience to gain the most information possible. With this, of course, one can assume that we have some sort of (mostly unconscious) sensorimotor knowledge about how the sensory input relates to the actions of our body and what changes we can expect from certain movements. This is also touched upon by Clark (2013) as he explores the implications of predictive cognitive systems in vision. However, I fail to see why this would also come hand-in-hand with the stronger side of Noë's

argument; that there cannot be such a thing as a passive and inert perceiver. Noë takes this part of the claim very seriously and is the keystone in what makes the strong sense of enactivism stand out. Perception in this claim is tied to a form of active mindedness. If one does not have the sensorimotor knowledge to understand that certain changes of perspective in the world are effects of one moving one's body, if one does not interact and 'tap-tap'[12] with one's environment and if one does not possess the ability to move and feel one's body, then one does not perceive anything. For this very strong claim to be justified, one needs more than the existence of a certain type of blindness, as it does not automatically follow with the implications of what could be a softer claim for enactivism (i.e. the acknowledgement of bodily importance without the stronger claim for necessity of action).

Imagine a person that is completely paralyzed. This person cannot move his body from the neck down. Both Strong and Soft enactivism would be happy to say that this person can still perceive the world through sight (and other senses that remain of course, but sight is the best example). Now imagine another person who is even more paralyzed. This person cannot move his or her eyes either and is constantly looking in the same direction. In this case, the stronger sense of enactivism that Noë wants to argue for would conclude that this person can only perceive the world as long as they a) can be moved around in it and (most importantly) b) have an understanding of how changes in their spatial location relate to changes in their vision. The softer claim only requires (b) to acknowledge the subject as being a perceiver. Now imagine a (probably non-human) creature that has lived its whole life in the exact same position. This creature cannot move about, has no developed sense of motor skills or how they relate to sensory experience and only possesses a single sense; vision. This is, in effect, an organic stationary security camera, a creature that only receives a single image through a set retina and will only ever receive a change in sensory experience if something in its field of vision changes. Will this creature understand the nature of the changes in its visual input? Probably not. It would not experience things as coming closer or walking away, as it would have no concept of three-dimensional spatiality having never moved about. Would we deny that this creature is perceiving *something* on some level? That is a harder question to answer (the answer is at least not so obvious that I'd be willing to conclude everyone would agree on the same answer). In Noë's case, this is a definite 'no'. There would be seeing (light hits the retina and a visual scene is presented), but not perceiving (no objects, depth or causal relations are recognised). One could compare this to viewing an abstract painting: We see something but with no concrete idea of what it is.

[12] Imagine a blind person creating an understanding of their environment by tapping with their cane in an exploratory fashion. Clark (2008) refers to the process of body babbling (Meltzoff & Moore, 1997), which is when we as humans (usually as babies) explore how certain neural commands create certain bodily movements. It is thus an exploration of how our bodies work in relation to our brains. Tap-tapping, in turn, is an exploration of the world in relation to our bodies.

If there is a difference in 'seeing' and 'perceiving' in the enactive approach, the distinction itself requires justification. In Noë's defence, one could claim that there is in fact no 'seeing' happening either, but simply collection of input data going to the creature's brain. Yet again, we could ask why there needs to be a distinction between 'seeing' and 'collecting visual input data'. Noë justified the stronger for enactivism by providing this new distinction, but did not properly explain why this distinction was needed.

As a final point, I would like to look at whether the secluded nature of the brain in the input-output model truly goes away if we endorse enactivism. Noë and other enactivists want to move away from 'Cartesianism', one of the perceived problems of which is that we are trapped inside of our heads. Enactivism claims that the dependency of action and introduction of the brain-body-world trinity does away with this worry, but I'm not sure it does. In order to test this, I will look at the classical thought experiment about a brain in a vat[13] from an enactivist perspective. If we were to assume that everything enactivism claims is true, how would we then react to a brain in a vat case? The thought experiment is as follows: there is a brain in a vat connected to a super computer, experiencing a simulated reality in which it understands itself as being a person with a body capable of moving around and interacting with its environment. This brain does not have a body, but it has gained sensorimotor knowledge through its simulated experiences. In fact, the simulation affects its perception and actions just like the real world would do on a developmental level. It would seem that, even though the brain is perceptually reliant on the external simulated world and perceives things best by tap-tapping forth with all of its senses, it is still the brain doing all the work in a simple input-output system. If this kind of thought experiment would have credibility, what difference would enactivism have made in terms of refuting the classic model? At best what enactivism does (and it does do it well, I will not deny that) is highlight that cognitive science and the philosophy of mind need to look more into how our brain uses external sources and bodily system to aid in its cognition.

How does enactivism relate to the other E's in the EEEE group of theories? It would certainly seem as if enactivism is closely related to (if not implicit in) both the embodied and embedded theories of cognition. That said, while Embodied seems to require a great deal of enactivism, enactivism in turn can do without the stronger cases of embodied theory (stronger case meaning that the brain is wired to one type of body and only one). One could, for example, imagine a case where a brain is put into a device that produces a novel range of sensorimotor experiences and relations (say a spider-shaped

[13] Definitely popularized by Putnam's (1981) refutation that this could be the case, and it is his version that I'm using here, though the origin of the thought experiment itself reaches further back with a similar scenario introduced in Gilbert (1973).

thing with both alien and familiar sensory organs). As long as the brain can 'baby babble' its way through and adapt to perceiving through this new body, there is still a case for enactivism but not one for strong embodied theory.

2.5 – Autopoiesis.

The three key features of enactivism are action, *inter*action and world-building. This much is apparent in the work of Francisco Varela, one of the people who - along with Thompson and Rosch - introduced the term "enaction" in *The Embodied Mind* (1991) as an evolution and solidification of ideas related to embodied cognition (Gibson, 1979), that grew through the 1980s. Varela (1987) points to how our framework of how we interpret and interact with the world is heavily influenced by our own bodily structure, which in turn is a product of our coupled history with the world, a causal evolutionary chain of action and interaction. He takes, as an example, the famous cantina scene from Star Wars. This cantina is full of, to us, imaginary creatures of all kinds. However, Varela points out, all of these creatures are, in their wild appearances, as limited in that regard as our imagination is. The shape and location of their eyes are different, but they do have eyes. They are vertebrate, stand up straight and interact with their environments much in the way us humans do. There is a point to be made for this. A lot of even the most outlandish creatures in fiction end up with senses and perceptions similar to ours. Even aliens from other galaxies "see" and "listen", and communicate through sounds and gestures.[14] Varela is claiming that this is due to exposure to evolutionary thinking, and that this kind of limitation was not present in, for example, the seventeenth century where (as he claims) we were more inclined to believe the plausibility of a human body with the head of a bird (Varela, 1987, p.54). The real force of this example is, especially for enactivism, the fact that the tools through which we perceive the world and the form we possess to interact with it is so intricately woven into our perception of what intelligence is, and for that matter what the world is like, that we often, perhaps without noticing, project these notions onto other forms of intelligent life. This goes hand in hand with other earlier sentiments such as Wittgenstein's famous "if a lion could speak, we could not understand him" (Wittgenstein, 1953/2001, p.223). The idea being here that the kind of body we possess and the kind of world we inhabit dictates the conceptual content of our language, to a point where two dissimilar life forms simply cannot understand each other. One

[14] Except the ones that don't. In plenty of instances across the history of science fiction aliens have been depicted with telepathic powers (including several species in Star Trek and a whole planet in Hitchhiker's Guide to the Galaxy) and, in rare cases, other exotic communications such as pheromone-based languages (like Weequay in Star Wars).

could argue that the early philosophical foundations for enactivist thought can be found in these sentiments.

To Varela, the enacted mind is autopoietic. That is, our cognitive system invents the world around us through coupled interaction with its surrounding medium and self-updating our structure in order to adjust to local conditions. In other words, an autopoietic system is evolutionarily adaptive in the way that it takes on a form *capable* of interacting with its surrounding medium, this interaction in turn leads to the interpretation and formation of a *world*. This world brings with it expectations that we act upon. For example, our human form allows us to register light waves through our eyes. We can interact with this impression through movement and manipulation in such a way that we can form an understanding of a world with depth, colours and objects. This in turn allows us to form empirical knowledge with expectations of how further interaction should turn out. We can as such gather knowledge about apples that we then use to further act upon apples in ways appropriate to our form (we can lift them up with our hands and eat them to gain sustenance). This interaction-to-world-building-to-action-to-interaction cycle is what the enactivist defines as cognition and applies to all living systems.

2.6 – Criticising Noë and the sub-personal.

Noë's enactivism makes the claim that perception is an inherently thoughtful activity. Yet, he also makes the claim that “[e]xperience is belief-independent in precisely the sense that it can look to you, say, as if the two lines of the Müller-Lyer illusion differ in length, even when you have drawn them yourself and know them to be of the same length.” (Noë, 2004, p.188) This kind of combination, combined with his views on what having a mind entails, comes with a certain baggage regarding concepts. Noë relies on the Müller-Lyer illusion to prove that what we perceive cannot be altered by our beliefs about the world, but he also wants to move away from Marr's model of vision (which he sees as the most prevalent view and intimately tied to the "standard view" brought up above) in order to present perception as something active that happens primarily in the interaction with and manipulation of the world. This enactivist view of perception still requires perception to be affected by concepts, however. As Noë puts it: "judgements and experiences can diverge and even contradict one another" but one needs "understanding of the ways experience presents the world as possibly being" (Noë, 2004, p.189) for a contradiction to be able to happen. That is why the experiential part of perception is intrinsically thoughtful; it is filtered through sensorimotor concepts that predicts for us how our interaction with the world will affect it, and thus creates an expectation of our experiences that helps us understand what it is we perceive. Noë draws upon Peacocke (and

indirectly upon Dennett) when he describes these concepts as being sub-personal primitive concepts resting on a non-conceptual basis. That is, they occur on a sub-personal level and are not constituted by lesser concepts, they are as basic as it gets and no further understanding is needed in order to understand what these concepts are.

This is the big problem for Noë though. What we are talking about when looking at enactivist perceptual concepts is three main qualities: sub-personal, conceptual and conscious. Consciousness/awareness is something Noë pushes for a lot when it comes to this theory. The type of consciousness that Noë refers to is quite broad and focused on experience. After all, enactivism relies on making the world 'come to life' through action. As such, it is a phenomenal consciousness that is being referred to, rather than the more limited or subdued *access consciousness* (Block, 1995) which instead refers to the property of a mental state or process being available for the organism as a whole, or for other mental states, to interact with. CTM can speak of consciousness this way without inviting discussion over phenomenology. Noë doesn't have this luxury.

To help with the discussion below, I'll set up a diagram comparing the three qualities of mental content, these being personal, conscious and conceptual, along with their opposites:

Personal	Conscious	Conceptual
Sub-personal	Unconscious	Non-conceptual

Some of these qualities overlap with one another, but it is not quite clear that they have any form of dependency on each other and it is possible to mix and match in many different ways. The kind of thoughts that are usually discussed in philosophy (judgements, "Bladerunner is a great movie" etc.) would go straight across the top row. A judgement such as "Bladerunner is a great movie" is personal-level (a person is thinking this), conscious (the person is actively aware of this thought) and conceptual (it makes use of concepts like "Bladerunner", "great" and "movie"). So what does it mean for this sensorimotor understanding of the world, as presented by Noë, to be sub-personal, conceptual and with an unclear statement of whether it is conscious or unconscious (although I would lean towards saying Noë would want it to be conscious). Well, this personal/sub-personal division was introduced by Dennett (1969). Back then, the distinction was working on an explanatory level, where things described on the personal level related to agents interacting with an

environment. For the subject Charlie, the personal level would thus be the things that one could say Charlie the person did and thought. The sub-personal by contrast, was all about the automated systems that one could not properly attach to persons but rather nervous systems and brains. Peacocke used this distinction in a looser sense (Hornsby, 2000; Peacocke, 1992) but it is still mystifying why Noë would pick the sub-personal level as the base for his highly active and awareness-dependent model of perception.

Judging from his claim about the cognitive status of brains as mediators of thought and not the home of the situated thinker, it is clear that what Noë means by sub-personal cannot possibly be what Dennett and Peacocke mean by it. If they were the same, then Noë would be pushing mental content and understanding into a fundamentally brain-centric view, which would go against his argument. As such, we have to postulate that Noë wants to introduce the idea of three levels of mind and brain. First, we have the physicalist level that Noë wants to refute as being a constitutive component of mind at all (except for being an organ that aids the mechanical interaction between body and environment). Second, we have the personal level. This we can assume to be more or less the same as it was in the Dennett model, an explanatory level that relates to persons and agents. Third, we have the sub-personal, which classically is exactly the same as the physicalist level. Now, however, Noë has removed its physicalist identity and instead introduced some sort of automated yet thoughtful content that serves below the level of what one could claim the "person" is and yet above that of brains and bodies. If one were to keep with Dennett's model, this third level could instead be explained as unconscious mental content (conscious but non-conceptual).

Saying that the sensorimotor part of perception takes place on a sub-personal level is quite alright if sub-personal means the mechanical handling of information going from the retina to the brain. It is even alright to make the sub-personal/physical split and claim that when this information is talked about on the sub-personal level it is talked about in an abstract way distinct from that of neural signals yet below that which one can claim to be fully personal, as that level usually requires a form of agency that simple information processing does not fulfil. However, it is still unclear how one would reconcile this with the view that perception is a) thoughtful and b) not situated in a brain-centric system. Noë wants the brain-body-world trinity to be the physical makeup from which minds arise. More so, mind is not situated in this trinity but in the interaction between the three parts. If concepts are things exclusive to minds and concepts appear on the sub-personal level, then both the sub-personal and personal must relate to the mind. As such, there is an abstract level of explanation that adheres to perception that has little or nothing to do with brains (since Noë's personal level resides in the brain-body-world trinity of action, and not in explicit brain states) that utilises concepts on a primitive level (yet are completely unaffected by our beliefs and learned knowledge, thus only

some innate basic concepts apply) and is situated in the interaction between a brain, the body in which the brain is situated, and the environment with which the body interacts, yet does not directly relate to the person that arises from this interaction. These goings-on in this in-between vague level of explanation allow the person to perceive the world. It is an explanation that is very much on the fence, very elusive, and overall hard to swallow. We may not be our brains, but Noë's fear of this possibility has caused him to throw out the baby with the bathwater.

2.7 - Hutto's Radical enactivism.

Daniel D. Hutto finds Noë's enactivism too conservative (Hutto, 2011a). Hutto rejects the idea, which he still finds implicit in sensorimotor accounts, that experience "is grounded in, implicit, practical knowledge" (Hutto, 2011a, p.28) as well as the follow-up to this: that experience is contentful. His argument is that these sentiments are unhelpful and rely on old-school cognitivism (i.e. computationalism, or Cartesianism) which in turn inhibits the theory from reaching its full potential. The key to be found in this disapproval of the "old cognitivism" is the presence of representational mental content. It is this key feature that Hutto rejects.

Hutto makes the claim that at its core, Noë's enactivism contains a contradiction. This contradiction being that, on the one hand, perception relies on intrinsic practical knowledge in order to interpret what we see while, on the other hand, perceiving involves no thought or intellectual skill. Noë walks a thin line in this regard with the help of his personal/sub-personal level distinction where the sensorimotor knowledge necessary for the complete visual experience of objects as spherical etc. is located on a sub-personal level (Noë, 2004). It is this distinction that Noë uses to dodge the contradiction and Hutto remarks that "it is not clear how [Noë] can have his cake and eat it too on this issue" (2011a, p.31). When Noë applies sub-personal knowledge to perception, he uses this as a form of passive filter through which our visual experiences are enhanced and translated. The process takes place in the sensorimotor knowledge, not the person, and as such it is not thinking that is taking place, since this knowledge is neither personal-level nor active in this regard. However, Hutto argues that experience is an occurring event that cannot be based in something passive but must be exercised. Of course, this is to be expected. A passive sub-personal cognitive filter like Noë's is highly representational in that it relies upon the content of the knowledge itself and not the exercise of knowledge. For sub-personal concepts to affect our visual experience *directly* but *passively* it must possess features that allow it to retain meaning independent of any active states. What I mean with this is that any concept that passively shapes mental content such as incoming visual experiences, and transforms them or enriches them, will have to be some sort of information-bearing structure.

What Noë has done is to sneak in computation into his enactivism, while sweeping it under the rug with the excuse that it is sub-personal and thus not part of the agent's active cognition. Hutto, being a more radical enactivist - as he himself proclaims in both *Enactivism: Why be Radical?* (2011a) and *Psychology unified: From folk psychology to radical enactivism* (2013a) - feels obligated to hold the view that *all* experience *must* be exercised. That is, our experiences must be of the enacted kind, actions and interactions creating a conscious thoughtful experience. To Hutto, this is a feature that distinguishes enactivism from computationalism in that for the latter, knowledge cannot be active, adopting Chomsky's line of "the idea that knowledge is ability is...entirely untenable" (1988, p.9). In contrast, Hutto thinks that the primary feature of enactivism is the idea that knowledge *is* ability, that is *has* to be, for action is the only thing that brings about conscious experience, which in turn is the defining feature of the enactivist's definition of cognition and mind. Hutto does admit that Noë claims that concepts are skills, but is unimpressed by this. Instead, he wants to push even further away from the classical view of cognition.

Hutto's view of the mind can be likened to a juggler who constantly needs to keep objects in the air. By doing this the juggler has created a circle, a motion of objects caught in a cycle. If the juggler were to stop juggling, the objects would fall down and the 'living' spectacle would be gone. Similarly, it is the continuous action that breathes life into the thinking creature, our thoughts only existing through action. This is further enforced by the contrast antithesis of an enactivist lifeform: the completely stationary creature I brought up earlier to show the difference between a perceiver and non-perceiver.

Hutto's radical enactivism states that "although the world is experienced a certain way, such experiencing is not intrinsically contentful" (2011a, p.34). What he really means with "contentful" is that the judgements and beliefs in a radical enactivist mind are not true or false in virtue of carrying representations or mental content that accurately or inaccurately describes the actual world (i.e. that these representations are ‘true’ or ‘false’). This is because in Hutto’s view, enactivist beliefs and judgements cannot be reduced to internal states but must be realised through the dynamic bonds the agent has with the environment

2.8 - The threat from enactivism

Radical enactivism’s objection to mental content is the largest threat to CTM. Revised or classical, computational theory would simply not have any real force behind its argument without positing the existence of mental symbols and representations. However, radical enactivism does not want any

symbols, as its supporters are only concerned with enacting the world, not representing it. Representing the world requires some amount of internalism, something radical enactivism wants to move away from. CTM needs symbols though, because at its core it postulates a computational language of thought. If there are no symbols, then we cannot form sentences, thus the language of thought would have no syntax. If we posit symbols, but have them empty of content, then we lose semantics. The result would be like a language where words and sentences lack meaning and thus refer to nothing. This latter move is suggested by Hutto (2013b) as one of CTM's only options, stating that mental representations cannot have content the way words in a language do. The particular example Hutto uses is that English 'snow' and French 'neige' are separate symbols with the same content. By contrast, Hutto argues that there is no such distinction to be made between mental symbols and their content, thus CTM should be content with symbols alone. This is not a good solution, and I do not endorse the rejection of content. When I draw up a hypothetical example for philosophical purposes I may denote two subjects as person A and person B. Here, "A" and "B" have linguistic functional reference to two (imagined) entities, but the semantics of "A" referring to person A is only temporary, held up by the current situational context, and will not refer to the same thing when some other philosopher draws up their own example or even when I do it again with a different context some other time. As symbols, A and B are short-lived, living and dying through environmental context. Even though Hutto wants to do away with representations altogether, these empty representations given meaning by functional context would be a step further in his direction towards a radical enactivism. CTM, however, requires a proper language for internal world representation to function, and we need both syntax and semantics for that to function. As such, it is this radical enactivism that I deem the primary threat that I will attempt to defend CTM from. In the coming chapters, I will introduce the means through which I will seek to create a compatibilist computational theory, and introduce the alternatives that I hope will show that the move to radical enactivism is unnecessary and in fact a disadvantageous act for creating a comprehensive theory of mind.

Enactivism attacks classical CTM with two main arguments: The first is that the world has to be enacted in order to be perceived. The idea of the passive observer which they seek to reject, and why this would also be an attack against CTM, is tied to the assumption of a passive system of vision present in Cartesianism/CTM. This assumption is born out of David Marr's theory of vision, a theory that I will explore in depth in Chapter V. I will not seek to deny the fact that CTM, and the input-output sandwich model, does indeed foster a very passive approach to perceptual cognition. However, I would also not agree to the idea that CTM is somehow bound to this passivity. In Chapter V, I will point toward the alternative of Predictive Processing, a primarily visually focused theory of

prediction-based cognition, styled in Bayesian fashion, that I will argue is by all means fully compatible, indeed constructed upon, the idea that the mind is computational and contentful.

The second argument is that mental content does not exist. This is not just born out of an argument that one can create a theory without mental content, but the criticism that content itself lacks a true function (Hutto, 2013b). This is a much harder form of criticism as it questions the very being of content itself, claiming that it is superfluous and may as well be cut away by Occam's razor. My strategy against this criticism is to prove that there are indeed areas of cognition where positing representational content *does* bring something to the table that couldn't be explained otherwise (at least not as *well*). Addressing this worry will primarily fall upon the last two chapters, but will incorporate many of the things that we will encounter throughout the coming chapters to form a bigger picture of what the mind is and why it serves best to be computational.

Chapter III - Heuristics and Modularity

Modularity and computation have historically gone very well together. Just as different parts of the brain do different things and communicate with each other, any complex computational structure will contain several processing units or programs carrying out different tasks. My view of the computational theory of mind certainly involves more components than simply a single processing unit lining up our surface thoughts to be processed computationally one at a time. Going further: I want to argue for the case that different faculties of our mind, just as there are parts of the brain, do different things, all computationally. I think that has been made clear from earlier chapters. This chapter has the purpose of showing how I see modularity being strongly connected with heuristic-based cognitive models, as well as presenting and arguing for the idea that these cognitive models can be computational, something that isn't a commonly held view among the proponents for heuristics in psychology. Thus this chapter is a triangle drama between modularity, computability and heuristics where modularity likes computability and heuristics likes modularity but is not too fond of computability. My goal is for them all to get along fine and as such it is the strained relationship between fast and frugal heuristics and quantifiable algorithmic computation that needs looking at. An important piece to this puzzle, however, is modularity, and that is what I will start out with.

3.0 – What is Modularity?

The idea of modularity in cognition started with Fodor in his book *The Modularity of Mind* 1983. There Fodor lays out a number of criteria that a cognitive system needs to fulfil in order to be considered modular:

Domain specificity: A modular cognitive system needs to be domain specific in that it only handles a certain limited range of inputs, thus only computes to solve a restricted amount of problems. Examples of such specific domains would be colour perception, voice recognition and visual shape analysis (Fodor, 1983). These are specific domains of input gathered from a much more general range of inputs in the form of visual and auditory sensory information. A modular cognitive system that handles colour vision would not bother with the visual input data that is not about what colours

things are. It would only take in the colour data and process it to give an answer to the problem of which colours are currently being registered in which areas of the eye, thus giving colour to our visual experience.

Mandatory operation: A modular cognitive system is not under conscious control. It works automatically and thus its operation is mandatory, i.e. there is no option whether or not to initiate the processes within the module, they simply trigger given the right context or input. If the right kind of stimulus is given for the system to start computing, it will start computing. This way the modular cognitive system is like its own little separate system working apart from the whole, taking care of the functions it's supposed to take care of without any need for supervision or command from any other areas of the grander system. The two following criteria are closely linked to this one.

Informational encapsulation: For a system to have informational encapsulation is for it to be limited, to some extent, in the range of information it has access to from the system as a whole, despite it being part of that grander system. This makes the resulting output from modular processes independent of a lot of information that the grander system might have. A module may for example only have access to information relevant to the particular task the module is designed for. This means that a module can come up with an answer that is refuted or seen as incorrect by the rest of the system because of additional information that the module had no access to. This way one could spot errors in one's colour vision for example. If a colour suddenly changes (say red apples turn blue), either due to a fault in the visual system or an actual change in the world, the colour vision system would still compute this input without suspicion. It cannot be suspicious because it has no information on the notion that apples are not supposed to be blue. The grander system however, or the agent, will be suspicious of this once the output is given, for such information is available to the whole. As Fodor puts it, the data that a modular cognitive system has to work with in order to fulfil its function "includes, in the general case, considerably less than the organism may know" (Fodor, 1083, p. 69).

Limited central accessibility: Limited central accessibility is just like informational encapsulation, only in reverse. Here, the central monitoring system (the set of processes that we as cognitive agents are directly aware of) cannot peer into the workings of the modular cognitive system. Input goes in, something happens inside and output comes out. To Fodor, the central processing unit is closer to the *you* that you perceive in your mind, in that it is more closely related to what we call consciousness. Limited central accessibility is there to explain why there are things going on in our

minds that we are not fully aware of. And it is true; there are plenty of thoughts and cognitive processes - some personal, some sub-personal - that we have each day that we are largely unaware of. Take for example movements based on muscle memory, where the brain sends commands to different parts of our body, yet we are not necessarily aware of every nuance to these commands other than the action we are performing. If I am running, I am not aware of the biomechanics involved in creating a successful running gait, I'm merely aware that I'm commanding my body to run.[15] One might argue that these are not real thoughts and while applicable to the brain are not similarly applicable to the mind. That is fine, but what about other more minded things we don't notice, such as being struck by a sudden epiphany on a problem one had been mulling over earlier? In such cases, it is as if some process has kept working on the problem while the conscious agent has been focusing on other things. Additionally, beliefs and moods can subconsciously affect our intuition of things around us, much to our own surprise when pointed out. For example, I may have had a bad experience with dogs in my childhood that I've since forgotten, and this has led to me having an aversion or negative attitude toward them. I may state that "dogs are the worst" and my mother would respond telling me that "you only say that because of that one time a dog bit you." These are examples of cognitive and minded activities that we as conscious beings have limited access to.

Fast processing: Processes in modular systems are supposed to be fast. However, how fast is unclear. Fodor mentions how people have been recorded to have 96% accuracy in rapid serial visual presentation tasks going at 167ms exposure per picture (Fodor, 1983, p. 63). This shows that even at such high speeds the processes of an allegedly modular system (or system*s* working in conjunction) can, for the most part, successfully keep up. One could say that generally it needs to be fast enough to go by without noticeable delay, like our language recognition and understanding for example. Both colour vision mentioned earlier and language recognition seem instantaneous to us. One could also reinterpret this criterion as regarding frugality, which links in well with the next criterion.

Shallow output: The way Fodor works with the 'cheapness' of modular output is underlined with the idea that output (or a concept created by output) with high specificity requires more from a system to produce compared to a rather general output (Fodor, 1983). The output of a modular cognitive system should be shallow. This can be interpreted in a number of ways. The general essence of it is that it is relatively easy to compute and the data retrieved from it is not very complex but rather straightforward. An argument for Fodor would be, for example, that only very basic

[15] In fact, if I were to have to think about every step I take, I may end up walking like a baby. Imagine the difficulty we experience when learning a whole new bodily task, anything from martial arts and dancing to swimming motions can feel counter-intuitive and non-fluid until we train them into memory.

concepts can be involved in the output of a modular system. By basic concepts he means things like DOG, which is far more simple and basic than the concept CORPORATION. DOG is a very simple concept in the sense that it involves physical descriptions held common among creatures that we refer to as dogs, its function being that we can accurately recognise and meaningfully think about dogs out there in the world. CORPORATION, on the other hand, is built up and supported by several other abstract concepts that require much more complex understanding of not only the physical world but of human society.

Fixed neural architecture: This means that the modular cognitive system can be realised in a specific neural architecture that can be, more or less, mapped. More generally the module can be traced to a specific area of the brain, or as a neural system that spans over small areas of several regions. What also seems important about this is that this fixed neural architecture can be traced through patterns in several subjects. If a certain module is in a certain place in one person's brain, then it should be in roughly the same area of another person's brain (Fodor, 1983). Just as we have areas in the brain associated with different functions like sight, association, motor function, higher mental functions etc., so does Fodor argue that modules should have similarly defined ranges of activity.

Characteristic developmental pace and patterns: There should be patterns in the development of these modular systems. I.e. when a group of infants grow up, there should be a recognisable pattern of their cognitive growth. If a certain part of the brain containing a certain modular system is supposed to be developing at a specific age, then infants at that age who indeed do have standard development of said area of the brain should also show a similar development in the relevant cognitive abilities (Fodor, 1983). This means that there should be a pattern in the biological development of modular systems just like the patterns in development of everything else in the body and brain.

It is important to note in regards to the encapsulation and (in)accessibility criteria that there are base-level and higher level processes going on in our minds (and especially our brains) that we neither control nor are directly aware of, even some of grander scale and importance to our everyday lives. The main cognitive systems that Fodor defines as modular are the language, language acquisition and perceptual systems. There is a very strong case for the notion that the visual system is largely encapsulated, but *does* have access to knowledge about shadows and contours, in addition to raw visual data. Knowledge about things can affect our perception of what we see and hear, as

shown in Edward H. Adelson's (1995) checker shadow illusion. In this illusion, a cylinder is standing on a checker board casting a shadow. In objective reality, the colour of the dark squares in the light is exactly the same as the light squares in the shadow. However, our minds, with knowledge of the shadow being a shadow, thus everything looking darker inside than they "should", see the light squares in the shadow as being of a lighter gray than the dark squares in the light. People often need to be convinced of the objective fact that the two are of the same shade by drawing a line between them, connecting the colour field. Even then the picture can look suspicious or distorted, as if the field changes colour halfway through. However, this is also where encapsulation has a case: Our knowledge of the connected colours being the same (that we gained through the proof of connecting them) does not at all affect our ability to see them as the same shade or colour. As such, the encapsulated visual system makes a set range of knowledge available to refine its input, while information outside of that range cannot affect our visual experience, no matter how hard we try to 'see it correctly'.[16] It seems that the visual system, in a frugal way, gets aid from a specific set of knowledge tools. These tools may be basic aids (akin to Noë's somatosensory information [Noë and O'Regan, 2001a]). This makes sense, since there would be no advantage for the visual system to take our wider scope of knowledge into account to better understand the world around us if it involved accessing the entire knowledge base for snippets of information that would possibly change the way we experience the current situation. There is simply too much information to process than what is biologically economic. Cognitive systems, like much else of an evolved organism's body, are generally efficient in a way that allows for survival in an environment that requires quick reactions. It is useful to have knowledge of how shadows and light interact with objects and geometry linked and readily accessible to a hypothetical visual module, because it helps us quickly grasp what is going on around us and takes the minimum amount of information gathering to process the visual data into something useful. Specific modular capacities that Fodor proposes include perception, language (Fodor, 1983) and cheater detection (Fodor, 2000a). Beyond Fodor, over the years many more modular systems have been proposed. For example, in developmental psychology there's suggestion of a modular capacity for reorientation following disorientation (Lee & Spelke, 2010) and Jonathan Haidt (2013) argues that human moral judgement behaviour is innate and modular. Based on this, the idea of modularity continues throughout the cognitive sciences even in recent years.

[16] There was a case of a blue-and-black dress that went viral on social media. Many saw the dress as white and gold due to a yellow backlight in the image. Instead of correcting for the light by reducing the amount of yellow experienced, their brains instead went the extra step of overcorrecting for the blue shadow on the dress itself, creating a visual experience of white-and-gold instead of the actual blue-and-black. This visual experience *could* be corrected by the use of knowledge and rethinking the perspective. This seems to clash somewhat with the checker shadow illusion case, but I would like to argue that perhaps the dress was a 'correctable' experience because of the ambiguity (there were two ways for the brain to correct; remove shadow or adjust light), while the checker shadow illusion is the mind fighting against a singular, unambiguous correction for shadow alone. See Lafer-Sousa et al. (2015).

You might notice that I've used words like "frugal" and "tool" while also addressing the benefits of quick processes where accuracy is not of utmost importance. You might also have noticed that the title of this chapter is *Heuristics and Modulatiry*. Well, what this chapter will go on to show is that heuristics-based psychology and mind modularity have a lot of things in common and can greatly benefit from one another. But first, let's get something else out of the way. When Fodor talks about modularity it is as much about brains as it is about minds, as can be recognised in the last two criteria listed. Now, I've made clear earlier that I think that when we speak of mind and brain, we are basically talking about the same thing but from two different angles, and that what relates to the mind is the abstract level of explanation (thoughts, beliefs) to what can be attributed to the brain, which in turn is a more physical level of explanation (neurons firing). In many cases one can draw direct parallels, especially when talking about systems, processes and modules. However, the issues I deal with are best treated on the level of explanation of the mind. My general rule is that while I will try to address issues purely in "mind-terms" (and certainly when talking about extended or enacted minds, 'mind' takes on more than just 'brain') to the best of my ability, sometimes making reference to brain processes brings more clarity to what is being discussed.

3.1 - Massive Modularity

While Fodor's classical account of modularity focuses mainly on a rather small set of specific select systems that fit the modular criteria, leaving the rest in a state more open for interpretation, massive modularity attempts to introduce most of the workings of the mind to modularity. Most importantly, it attempts to bring in characteristics of the central processing system into modularity. Fodor's theory is centred around the idea that modules *surround* or connect to the central processing system, which in turn, contrary to modules, takes care of things such as reasoning and decision-making in a more minute and complex manner (central system reasoning is not bound by fast processing nor shallow output). Massive modularity goes against this notion of a central system, and throws out quite a few of the criteria as irrelevant, by stating that even these faculties can be modular (which, as I like to link modularity with heuristics, I want to support). One of the greatest proponents for massive modularity is Peter Carruthers, who in his (2006) book *The Architecture of the Mind* gave the first comprehensive account of the theory. The idea that the mind was (mostly) made up of several specialised sub-systems was already popular in the late nineties and early 2000s. Perhaps the most popularised account comes from Steven Pinker's *How the Mind Works* (1997). Pinker has argued for a very evolution-based account of the mind's functions, setting it in an adaptive as well as

computational framework. Carruthers built upon this theme of specialisation and proposed a massively modular system that went well beyond Fodor's vision. To make this function, Carruthers noted that several criteria "will very likely have to be struck out" (Carruthers, 2006, p.12). More specifically, these criteria were the shallow, fast processing and encapsulated criteria. I've already touched upon the problem with encapsulation, but I have also pointed at examples of how one could argue for a middle ground of semi-encapsulation or selective encapsulation. Carruthers' claim that encapsulation would no longer be a criterion is based in his desire to include all systems under the modular definition, and certain systems (such as decision-making) have no business being described as encapsulated. However, Carruthers also does not remove the (in)accessibility criteria, but modifies it to ascribe to modules the feature of possible inaccessibility to other modular systems (it is thus a network of modules, some only having limited access to others). As such, there are a variety of restrictions between different systems and, since no system is strictly *the* central system but rather a grouping of systems, some of them being more central to cognition than others (vision is more peripheral than general reasoning, which in turn is more central), the (in)accessibility criterion would suffice to cover for both itself and encapsulation in this looser sense of modularity. Since, as described above, encapsulation is accessibility in reverse (and vice versa) having both only makes sense in a modular system where there is a distinct crossing of information between modular and non-modular systems. But with several modules working in integrated ways with each other, only one criterion is necessary. Encapsulation goes since it is the most problematic and strict by Fodor's definition. However, Fodor claims that, while many of his other criteria could be called into debate (and have been weakened by arguments from proponents of massive modularity) the heart of his notion of module lays in that it must be informationally encapsulated from information that is outside of its domain. The database could then in turn be expanded beyond the module itself (Fodor, 2000b, p. 62-63). Even so, limited accessibility still achieves exactly this, without leading into needless clashes with observable evidence (like Adelson's illusions).

Shallow content being removed as a criterion is obvious, since now all of the mind functions in a modular way, thus complex content *has* to be included (unless we want to argue against the existence of complex mental content altogether). What massive modularity aims to achieve is in the end a weaker (or looser) form of modularity where module simply means a function-specific system that can be isolated and identified within specific, although dispersed, neural structures that are domain-specific in the sense that only relevant inputs are being distributed and handled, and that a lot of them are not subject to the direct control or awareness of the cognitive agent. This last point can be a bit problematic when pushing massive modularity to its most extreme, as it seems to imply

at first glance that we are not in control (or even aware) of anything going on within the modular system, which just so happens to be the entire system. If modular systems are not subject to the will of the agent, and all systems of the mind are in some sense modular, then there's no system in the mind that is subject to the will of the agent. This interpretation, however, I think takes modularity and its criteria a bit too far. In Carruthers' version of modularity, not all of the criteria apply all of the time (Carruthers, 2006). It is obvious that some systems are more open to our conscious awareness than others, most of these are those that are still considered more "central" than others.

3.2 - Modularity into Heuristics

So whether massive modularity can be taken in its strongest sense, in a loose sense, or neglected in favour of Fodor's sense of modularity, there's nonetheless a system here that shares a lot of similarities with what Gigerenzer was trying to accomplish with his adaptive toolbox (Gigerenzer, 1999 & 2008). In his paper 'Why Heuristics Work' (2008) Gerd Gigerenzer identifies ten heuristics that he views as fundamental. These heuristics are problem-solving tools that we apply on a day-to-day basis in order to avoid having to stop and deliberate over every decision that we have to make throughout the day. These are:

The recognition heuristic: Give higher value to recognized alternatives. This could mean for example giving higher priority to brands you're used to or focusing on the alternative in a pop quiz that you recognise from somewhere else.

The fluency heuristic: Go with the faster alternative. Often to avoid inconvenience, choosing this heuristic involves going with the alternative that has the least stipulations attached to it, to maintain flow of actions. This could occur in situations where the flow is important or more comfortable, like when choosing a new phone plan and you are generally uninterested in all the features and options outside being able to call.

Take the best: Go through cues until one alternative seems better in this regard. As such, working through a long list of possibilities is much more easily handled when you can stop halfway through. Example: I try to choose between which to purchase out of two books for a long trip. I decide that what I value in this situation out of a book is 1) price, 2) length and 3) excitement. The books are both

in the same sale pile and thus of equal price, so my first cue is neutral. One book is 100 pages longer than the other, and thus wins out (it's a long trip). Thus even though the shorter book may be more exciting, the length was valued higher and was the first step where a discrimination could be made.

Tally: Do not estimate weight of arguments but *amount* of arguments in favour/against. This can be likened to a "diplomatic" approach where each argument has a value of exactly "one". I may for example tally what the positives between going to see a movie or spending more time reading philosophy. I can tally between these and count the number of positives on each side, making the larger tally the victor. This is likely to be cancelled if one argument is simply too great to be treated as a value of "one". For example, when deciding whether to go see a movie with friends or spend time at home with my partner: 1) If I go to the movie I get to see my friends, 2) I've really wanted to see this movie for a while, 3) I just remembered it's my partner's birthday and I really need to be there. Here, it's two against one, but the one is simply too great for the Tally heuristic to remain a desired method, and thus cancels.

Satisficing: Pick the first alternative that is good enough for your aspiration. Much like *take the best*, this heuristic differs in that you know exactly what you need. As such, you're not looking for something that sticks out but rather what first comes up as satisficing your current desires. This could be equated to accepting the first candidate that qualifies for a position and offering them the job without looking at the remaining applicants. This heuristic, much like *take the best*, very rarely receives optimal results, but that's not what heuristics are meant for.

1/N: The default heuristic. Often compared to allocating your funds equally into *N* amount of funds. Betting on all horses assures a win, and is thus attractive when outweighing losses are not apparent. If I can't decide between ketchup or mustard I may go with a little bit of both, and when I choose how to spend my time I may allocate it equally between tasks without any specific priority.

Tit-for-tat: Cooperate first, then act toward the other as they acted toward you (thus if both cooperated, cooperation continues). Also known as the reverse golden rule.[17] This builds on trust, where cooperation is assumed at the outset (giving the other the benefit of the doubt) and then respond to friendliness or hostility in kind.

[17] As an interesting point of trivia, this is also the same golden rule that appears in LaVeyan Satanism.

Imitate the majority and *Imitate the most successful*: These last two are self-explanatory. The psychology behind the former is also connected to the informal fallacy *argumentum ad populum*, showing more clearly than other examples that heuristics are not perfect and can sometimes lead to problems in modern society, yet is also widely used. "Everyone is doing it" is a common thing to hear to justify certain actions, especially about things like going a little over the speed limit and other minor but commonplace crimes and risks.

All heuristics exploit evolved capacities in that they are reliant on recognition memory, recollection memory, imitation etc. These are capabilities we are born with, and that we turn into capacities through practice. The 'evolved' capacities that Gigerenzer is speaking of here are thus not entirely shaped by nature nor nurture alone, but rather by both on a step-by-step basis. This seems to imply that an evolved capability can thus become dormant unless turned into a capacity through practice.

Gigerenzer claims that the reason why a psychology based on heuristics works is because heuristics are tools that have been customized over time, both in terms of evolution and in terms of individual training and discovery, to create something ecological (as opposed to logical) that works much better in a Darwinian sense, where the problems that need solving are less of the formal nature that logic and probability theories apply to, like mathematics or scientific research. Heuristics are fast and frugal, thus better suited for survival (Gigerenzer, 1999). This makes more sense than having a psychology entirely based on more complex and information-hungry logic or probability. To Gigerenzer, these methods are not apparent to us (we are not aware that we employ most of them).

Heuristics are methods for making decisions and judgments, which are in turn considered to be two very central modular faculties in Carruthers' massive modularity system. Massively modular or not, I want to argue for the adaptive toolbox to be modular and in turn, in opposition to Gigerenzer's claims, computational. The Darwinian aspects of both modularity and heuristics-based psychology are highly apparent. They both seek to explain the workings of our mind from an evolutionary developed standpoint, probably as a result of the contemporary developments in cognitive science and evolutionary psychology. However, their goals differ. Gigerenzer contrasts his heuristics with computational alternatives, implying that the heuristics themselves are not to be considered as computations, but as non-computational processes. Carruthers, on the other hand, extends computational theory of mind further than Fodor did, as Fodor points out in a response to Carruthers and Pinker's new computational trends: "There is, in short, every reason to suppose that the

Computational Theory is part of the truth about cognition. But it hadn't occurred to me that anyone could suppose that it's a very large part of the truth; still less that it's within miles of being the whole story about how the mind works." (Fodor, 2000b, p.2)

And yet, I would argue (in a very non-heuristic way) that the similarities between modularity and heuristics *outweigh* their differences, and that those in turn could easily be revised and reconciled. If successful, this would mean that the heuristics of the mind actually refer to a cluster of frugal decision-making modules used as a developing toolbox to help make quick decisions where extensive rational thinking is, due to restrictions such as effort or time, not desirable or, due to lack of information, simply not possible.

3.2.1 - Similarities

Both heuristics and cognitive modules are viewed as tools, in the former case more explicitly. Modules are applied as situations that require them arise. In perception this is perhaps not as obvious, as we spend most of our conscious (and arguably unconscious) time perceiving things. As such a vision module would almost always be "on". What about a module to catch cheaters? Fodor proposes the existence of such a module, arguing that we are much better at catching an underage drinker than we are finding an odd number, as we will see below, even when both tasks are the exact same Wason selection task (Fodor, 2000a). The Wason selection task is a card-based puzzle, meant to test capacities for logical thinking. It was devised by Peter Cathcart Wason (1968) and goes as follows: There are four cards on a table, one displaying a vowel, the second displaying a consonant, the third displaying an even number and a fourth displaying an odd number. The subject taking the test would be told a rule: “if there is a vowel on one side of the card, then there is an even number on the other side” (Wason, 1968, p.273). The goal of the test is to select the cards needed (and only those needed) to be flipped in order to determine the truth of the rule. The correct solution is to turn over the first and fourth cards, because you want to make sure that there is not an odd number on the other side of the vowel card, and that there is not a vowel on the other side of the odd numbered card. However, when testing people with this task, Wason noticed a very low rate of success, as well as a trend in errors. While nearly all subjects correctly picked the first card, the fourth card was almost never selected. More favoured was the third card. Years later, Tooby & Cosmides (1992) discovered that the success of this test was highly context-based, and if one moved away from numbers and letters, and instead framed the test in terms of social relations, success

rates shot up. A particularly successful example of this was putting the subject in a social situation as being a cop checking a bar for underage drinkers[18]. The success rate of the task seemed to be context-sensitive: when we are faced with the task in an abstract form (the arbitrary correlation between types of numbers and letters) we struggle, but as soon as it is put in a context that we recognise socially, our reasoning improves and we excel. This caused the suggestion of a cheater-catching module (Fodor, 2000a, Tooby & Cosmides, 1992). In this modular sense, when placed in the correct situations, the performance of a certain task (like getting the correct answer) becomes more intuitive (is fast and automatic) and requires less cognitive 'work' (is frugal).

Likewise, Gigerenzer's adaptive toolbox has many heuristics that are context sensitive. *Tit-for-tat* is obviously tailored around social interactions, and the *satisficing* heuristic does not apply if we are unaware of what exactly would satisfy, in which case *take the best* would be more applicable (and in turn the latter can be suppressed by the former). Note that the heuristics I mention here are specifically those that Gigerenzer has identified as fundamental. They are thus not all the heuristics they are (the toolbox is adaptive and ever-expanding) nor are they necessarily the definitive list of fundamental heuristics in human cognition. Gigerenzer could be wrong. There could be more, there could be less, or there could be a different set where some remain and some are switched out. Thus, when I'm selecting specific examples I do this to illustrate the point but do not necessarily subscribe to the idea that "yes, these are our specific heuristics".

Both modules and heuristics are designed to be frugal and efficient, but also serve the purpose of bypassing computational intractability. That is, they are imperfect or limited in just how far they can take their processes in order to be able to arrive at quick and sufficiently accurate output to domain-specific input. In terms of heuristics, instead of context-based input triggering a module, we have predicaments of choice calling for certain heuristic tools.

3.2.2 - Differences

I'm not going to argue that all modules are heuristics-based, nor that the varied behaviour of modules applies to the adaptive toolbox. If a heuristics-based psychology exists within our minds, it is

[18] Here the four cards would be underage, adult, has-a-soda and has-a-beer.

a subset of all the modules present there, but modular nonetheless. It is not even constituted by the whole set of our decision-making and judgement faculties, but a subset of those as well.

The main difference between Gigerenzer's heuristics and Carruthers' modules is that the modules are described as not having the ability to be consciously activated. Gigerenzer is not only able to describe his heuristics, but also deploy them. Once you know how a heuristic operates (since it is a rather simple rule of thumb) you can easily employ it in higher reasoning. If I wanted to, I could employ *1/N* as a *strategy* for monetary investment. Likewise, instead of simply going for what I recognise as described by the *recognition heuristic*, I could employ it in a (relatively) more lengthy reasoning process where I could rationalise the decision. For example, when asked which city has the highest population, Stockholm or Kaohsiung City, I could go for my (heuristics-regulated) gut feeling that Stockholm is the answer, or I could use the same strategy to think to myself that "Stockholm is known to me while Kaohsiung is not, so maybe it is known to me precisely because it is a more populated city. I haven't heard of Kaohsiung so it can't be that big or important." In both cases I would be wrong, but that is not the worry. The worry is that I could consciously employ a heuristic, which violates it being modular in the strong sense.

However, there is an important difference here. When I respond to my gut feeling through the *recognition heuristic*, I am handling output that has been given to me (not even necessarily reactively or reflexively) through a process I was not consciously aware took place (thus a gut-feeling). This process might be similar to what I arrive at using my (somewhat) flawed reasoning employing the same strategy, but it involves far more justifications and variables ("maybe it is because", knowledge of the possible ways population size can affect the fame of a city etc.). As such, it is more akin to a post hoc reasoning of the gut feeling already present. In addition to this, nowhere does it state that a heuristic or module does things that cannot be done otherwise through general reasoning or by employing similar strategies. Only a select few modular tools posses capabilities that are beyond the rest of our mental faculties (the visual system, for example, is very specific and not easy to reconcile or translate into a trail of thought). There are cases where modules would do better, as in the case of the Wason selection task, but even then subjects did not have an impossible time working out the problem in the "non-modular" case (the results for the correct answer were less than 10%, but that's still a sizeable percentage). Thus conscious employment of strategies *similar* to the heuristics in our (hypothetically) modular adaptive toolbox does not violate the criteria of mandatory operation and limited accessibility, it simply means that we're doing it the hard way (like backtracking to make sure that we locked the door, we are double-checking). This difference is, as such, not a violation of

criteria. The difference is that heuristics are simpler, which makes them easier to understand and re-employ as tools of general reasoning.

Looking at compatibility with massive modularity in particular, we could argue the possibility that limited accessibility has a much wider spectrum in just how limited it is depending on what the different modules are and how close they are to the "central cluster" so to speak. Indeed, if we truly want to take massive modularity to its furthest extent of dividing *all* of the mind into modules and sub-modules, then we would need some way of taking into account our conscious awareness of many of these processes, as well as our general-reasoning capabilities with more robust and much less frugal cognitive tools such as hard logic and mathematics. To Gigerenzer the toolbox is but a toolbox, and our general reasoning is just that. With a central system, there is no problem with this. The clash occurs when you want to get rid of the central processing system.

As I've mentioned before, my argument does not need massive modularity to be completely successful. In fact, I would rather prefer if I could defend the existence of a central processing system. What I need is to prove that heuristics-based psychology is compatible (and favourably so) with modularity of mind, and for this I need to defend a modularity of mind that is somewhat less restrictive than Fodor's classical account, while still retaining the essence of the theory. With these factors in place, I can achieve the purpose of this chapter: to prove that heuristics can be (and are) computational. The next section will deal with a not-so-different but definitely separate account of heuristics-based psychology by Amos Tversky and Daniel Kahneman (1974). This account for heuristic rules and how they interact without thought processes will underline just how computational heuristics actually are, without moving far away from Gigerenzer's account. In fact, Tversky & Kahneman help greatly in illuminating the working parts of mental heuristics, on both good and bad, which in turn starts reeling in Gigerenzer's model to join the party. The computational party that is.

3.3 - Kahneman's Heuristics, Extended and Encapsulated.

No, this is not about heuristic rules written down for reference in a notebook or downloaded from a USB stick. Rather, this section is about the extended *use* and *presence* of heuristics *within* the mind as proposed by psychologist Daniel Kahneman. To Kahneman, heuristics are used in all manner of processes involving data of limited validity, including many visual processes such as distance judgement (Tversky & Kahneman, 1974). While the heuristics presented by Gigerenzer are all fairly

central modularly speaking, Kahneman's account reaches much further. Although, one could argue, they still remain firmly within the realm of judgement-based modules and it is not entirely clear whether or not the core "visual module", if such a singular module exists, (more about that in a later chapter) is one of these or if distance judgement would make its home as a connected module.

While Gigerenzer states that a heuristics-based psychology is less than optimal, Kahneman goes further and claims that it is in certain cases even the cause for severe errors in judgement. Even when given additional helpful tools and information, such information is often disregarded in favour of the standard operation of the relevant heuristic. For example, when a group of test subjects were presented with the ratio in the amount of lawyers and engineers present in a group, and asked to attribute described characteristics of members in that group to either lawyer or engineer, the probabilistic advantage of knowing how many engineers as opposed to lawyers were present (in this case 70 to 30) was disregarded in favour of simple application of the *representativeness heuristic* (Tversky & Kahneman, 1974*,* p.1125). Knowing that there are more than twice as many engineers than lawyers adds an important piece of information that seemingly *should* affect the outcome of a person's reasoning when it comes to selecting whom a certain characteristic belongs to. If there are 100 descriptions and any one of them is chosen at random, it's more likely that the person described, no matter what the description, is an engineer. If this information is considered, that should lead to a pattern during a series of choices pointing to more people being pointed out as engineers than lawyers. However, according Tversky & Kahneman the findings tell a different story. There was no discernible pattern pointing toward the probabilistic majority. Instead, almost each and every description was allotted into what best fit the stereotype of engineer or lawyer. Kahneman claims that the representativeness heuristic is a psychological rule that says the probability of A belonging to B is higher if A is representative of the qualities of B. In this case, characteristic A is more likely to be describing a lawyer if that characteristic is better representative of a stereotypical lawyer than a stereotypical engineer. The subjects in the experiment knew of the probabilistic advantage of one group over the other (there were two types of tests, flipsides of each other when it came to the ratio information). Even when given completely worthless descriptions pointing neither way, the probability distribution from subjects' responses was even between engineer and lawyer, instead of the expected slant in the posited majority's favour.

This seemingly subconscious habit of ignoring even helpful information in regards to following a frugal rule feels very similar to how Fodor describes the encapsulation of modules. Even though I mentioned earlier that encapsulation was one of the original criteria that was changed in massive

modularity, note that I also pointed out how limited accessibility fills its place sufficiently anyway. Encapsulated or limited accessibility, the results of the experiment presented in Tversky & Kahneman (1974) point to heuristic processes behaving in a similar way to a module that is limited in available data. The fact that it is ignored even if the cognitive agent has been informed and knows better suggests that once heuristics are applied, they operate independently of contradicting or interfering information in the agent's possession. It doesn't seem to be by choice but simply by habit of applying that specific cognitive tool in that specific way. One could wonder if the subjects of the experiment, if having the relevance of the given ratio explained to them, would have been more likely to use the hunch they get from their use of the *representativeness heuristic* and then modify that by further rationalisation and analysis of wider knowledge. This would go together with Gigerenzer's line of thought that we are not slaves to our heuristics; it is simply the instinctive and evolved behaviour (Gigerenzer, 1999). If we have no need to override it, or realise that maybe we should apply extra thought, then the gut feeling is the response that we'll give despite its flaws. According to Tversky & Kahneman, the subjects would retain their bias toward the heuristic rule even if encouraged to be thorough and rewarded for a correct answer (Tversky & Kahneman, 1974). However, it remains unclear just to what extent the usefulness of the ratio information was pointed out to the subjects, but that will remain beside the point in the long run. In any case, the fact that heuristics cause us to get things so wrong gives further support to the idea that they are module-like, as the limitation of the knowledge base taken into account for the process is yet another similarity that the two share.

3.4 - Heuristics as Computation

Heuristics is in many cases seen as a non-computational alternative to solving problems, or a gut-reaction applied to a problem that doesn't include any rational thought in the way of hard logic variety algorithmic thinking. In the majority of philosophical instances these problems take the form of the making of a decision, picking between options or making a call with insufficient or overwhelming amounts of available data. Outside the ramifications of philosophy and psychology, and moving into the realm of mathematics and computer science, heuristics are problem solving strategies for, once again, finding solutions to problems. In these cases, the problems have the nature of being one or several of the following, as listed by Michalewicz & Fogel (2000):

Too many possible solutions in the search space: This is a very straightforward issue relating directly to frugality and computational intractability. In fact, this is perhaps the most commonly cited

reason for the necessity of heuristics in almost all accounts thereof. The premise itself is simple and I've gone through it before: considering all options is in certain scenarios simply not practically possible due to the size of the search space. Even with great structure, there is simply not enough time to consider all the variables. This could be anything from a slightly too time-consuming computation in a pressed situation, or an actual practically endless computation that simply cannot ever be completed.[19]

The complexity of the problem necessitates a simplification which in turn makes the results useless: This characteristic hangs on the idea that when we attempt to solve a problem, what we are working with is in fact a *model* of the problem, a simplification of the real world, isolated and strapped in an environment where only the relevant components are present. This is especially true in mathematical or theoretical problems. The issue appears when the model gains low fidelity due to oversimplification or too many rough approximations. For example, when calculating for efficiency in shipping routes by truck, the model would have to account for traffic density. However, we cannot predict exact future numbers, but we *can* determine daily averages. It is thus these averages that we have to rely on when creating the model. If these averages are not reliable (say traffic density swings heavily and randomly on a certain road, making the average of this wide range seldom a realistic depiction of actual traffic) then neither is the solution. This leads into the next problem:

The evaluation function requires a series of solutions instead of just one due to changes over time: Taking the same example as above of wanting the most efficient shipping route by truck or car, there are all manner of singular random events that would require their own calculations in order to appropriately anticipate. A car crash could severely change what the optimal route will be. Similarly, roadwork can change the variables of certain routes, but vary greatly in impact due to the element of luck.

The options are so constrained that finding even a feasible answer (let alone optimal) is difficult: This means that even though there are hard constraints to a problem, i.e. constraints that *cannot* be violated (to the problem of helping an old person across the street, shooting her is not a viable option), there are also soft constraints. These soft constraints are things that we'd rather avoid but might not be able to. In an ideal world, no soft or hard constraints would be violated by the solution. However, the soft constraints can be set up in such a way that some combination of them simply

[19] The Boolean satisfiability problem is an example of this, as checking all the solutions to this problem would take even a constantly running computer billions of years to perform (Michalewicz & Fogel, 2000).

have to be violated in every single scenario. It is thus a task of weighing these soft constraints against each other to find a solution that has the least level of constraint-violation. The problem occurs when quantifying these violations, as each soft constraint can have its own unique weight, and even several subjective weights to different people. If helping an old lady cross the road somehow (in some unfortunate universe) boils down to a choice of her hip breaking or the picture of her recently deceased husband being lost, the helper might prefer to save the hip while the old lady might prefer to save the photograph. Either solution gets the lady across but neither is optimal, nor easily quantified.

The solver has set up a psychological barrier for themselves, preventing them from finding the answer: The solver or decision-maker can sometimes set up a barrier that prevents them from approaching the problem appropriately. In mathematics, Michalewicz & Fogel (2000) takes the example of asking students to (a) 'find a solution for x' and how students are more capable at this than when they are asked to (b) 'prove something about the solution for x'. Despite these two tasks being the same, just worded differently, the students put themselves in a different mindset when tasked with (b), making them unable to solve it if they are insecure about proving something. In more general situations, this could involve the self-imposed inability to even approach the question of "which is the most populated city out of *x* and *y*" because they "don't know anything about geography" even if the person in question was equipped with the means to arrive at a correct answer.

Note that in these cases, heuristics are used to find non-optimal solutions to problems that have a defined answer, i.e. "which path is the shortest?" or "what is the cheapest way of managing the factory?" The big difference with these problems and the decision-making situations handled by Gigerenzer and Kahneman is that the latter are often more open-ended tasks. Still, in both areas heuristics remain essentially the same: they are tools for solving problems that methods via straightforward reasoning or hard logic are not equipped to handle.

The heuristics found in computer science and mathematics are created by design. They are tools devised specifically for a certain problem, with specific trade-offs in mind in terms of optimality, time or probabilistic accuracy. This would seemingly put them at odds with the heuristics of, say, an evolved toolbox à la Gigerenzer. However, this is not necessarily the case. Just like evolved heuristics, designed heuristics improve or change over time. The difference here is that changes in designed heuristics are due to academic advancement while evolved heuristics change due to, well, evolution

(and personal-level learning and development). In one way or another, they both develop. Similarly, the evolved heuristics of today have evolved in such a way that they practically are designed to handle specific problems, just like designed heuristics. Evolved heuristics work through the evolutionary development of the brain and subsequently the mind. It is a process both of survival of the fittest as well as individual learning and fine-tuning of what works and what doesn't (sufficiently). Designed heuristics also undergo a level of natural selection, only this time in terms of failed and successful theories and models. To make a distinction between evolved and designed heuristics does not achieve much, as both share the same development and end result, the difference being the paths the development takes in terms of time and causal chain. In fact, it makes more sense to distinguish the two as survival heuristics and academic heuristics.

But do they function in the same way? Well, they both work by limiting or minimising workload in order to balance effort against results. Heuristics in any form will always be about getting as good a result as possible while saving as much time and effort as one can. In short, all heuristics are frugal, this is their essence. Both also handle and find workarounds for flaws in rigid mathematical or logical systems. Computational intractability is a major motivation for both the evolved toolbox and a mathematical heuristic algorithm. And that is a keyword here: algorithm. As mentioned before, Gigerenzer does not see the evolved toolbox as a set of computational tools in the classical sense. That is, it is an alternative to logic and probability models of mind (Gigerenzer, 2008), both pointing far back to the Turing days of computability of mind. Nonetheless, that does not explicitly make them non-computational. That would require a very conservative view of how we define computational and non-computational, such as that held by Penrose (1994) as he assessed that the only alternatives to solving problems outside of hard logic involved strategies beyond quantifiability and symbol manipulation. Seeing as how all the heuristics of computer science are expressed in terms of algorithms manipulating symbols through computational procedure, and while being used this way arrive at similar solutions to heuristics in psychology, does the groundwork for the possibility that Gigerenzer or Tversky & Kahneman's heuristics could be translated into computer heuristics. Or better yet, that they already are algorithms within the mind.

3.5 - A Difference of Perspective

Could it be that modularity of mind and heuristics-based psychology are two theories about largely the same ideas, just from the differing perspectives of separate fields of study? It would not be

impossible. In addition, even in the strictest of senses, the fields of psychology, cognitive science and philosophy of mind have exchanged ideas and influenced each other many times, especially in recent decades. In a less restrictive interpretation of the subject matter one could claim that the fields of study have actually mostly been sharing the same field, and crossed each other's paths extensively. It is only natural that similar ideas would appear in fields sharing the same subject matter, which in this case is the mind. I think I have in this chapter shown that there are more than enough obvious links between the idea of our mind being made out of modules and the idea that we utilise quick mental tools specialised for different types of decisions and judgements for us to dare make the jump that heuristics, as presented in both Gigerenzer and Kahneman, can be considered modular (as processes and strategies of modular systems). And not to stop there, there are even things that these heuristics can show empirically about modules that had previously been discarded or not thought of. For example, Tversky & Kahneman's representativeness experiment shows that encapsulation might be much more relevant a criterion than Carruthers' thought when revising modularity of mind theory for its upgrade to massive modularity. Kahneman's theories also create links between peripheral modular systems and the application of heuristics, suggesting that heuristics might not be as centralised as first thought when only taking Gigerenzer's account in mind.

Can heuristics be considered to be computational processes? Yes, I do believe so. In this chapter, I have been looking at similarities between heuristics and modularity, as well as comparing heuristic models from different fields. In both instances there has been no reason to doubt the computability of either, and enough similarities to assume that something akin to computation could be going on even as we use fast and frugal rules of thumb in our minds. Additionally, looking at the phenomenon of heuristics itself there's little to be found that is so unique and so obscure that it would be considered impossible to somehow fit into an algorithm, or turn into syntax-based rules like a computational language. Cognitive computability is, after all, the manipulation of mental symbols in an algorithmic, syntactic manner, through the process generating output from input. In fact, this kind of description lends itself really well to the type of toolbox Gigerenzer is describing. The adaptive toolbox is a set of adaptive tools employed for the purpose of solving problems in the form of judgements or decisions when data is lacking (but not absent). As such we have insufficient input to employ our higher-level thinking properly, or we lack the time for such a process to take place (or it simply isn't feasible for an exhaustive process to be computed in the first place, like the endless or life-span exceeding computations presented by Penrose among others). In such cases, these tools take what data there is and process it, using whatever problem-solving technique that is appropriate. Michalewicz & Fogel describe this as "turn[ing] the crank" (Michalewicz & Fogel, 2000, p.139), we

engage in a process and like a jack in the box, the answer pops out.[20] Many of Gigerenzer's heuristics are of this nature: We simply apply a set of rules to a situation and work through the motions. *Take the best* for example is a very mechanical method, going down a list of features one-by-one and stopping when one side wins out. What is going on between start and finish, or during our choice of method? Could it be manipulation of mental symbols the way mathematical heuristics is a manipulation of mathematical symbols? Yes, it would in fact be consistent with how they are written down: Gigerenzer's heuristics in particular are exceedingly formulaic in the way they are written. Looking over the core heuristics listed in his work, there are a lot of if-thens,[21] inherent weights (to be found in the *recognition heuristic* and *take the best* where one is "searching through cues in order of validity" [Gigerenzer, 2008, p.24]) and tallying (obviously represented in the *tallying heuristic*), all of which could be produced in symbolic form and run through an algorithmic table if put to paper.

So what do the results in this chapter mean for the argument as a whole? Well, it shows for one that there is good empirical evidence that something like heuristics happen on a psychological level in humans. We do not use all of the data available but have developed frugal methods to deal with quick decision-making. These methods are sub-optimal by nature (Gigerenzer), yielding sufficiently accurate results fit for basic decisions. However, when it comes to problems requiring higher-level thinking, our innate heuristics (as opposed to those we design in mathematics and computer science specifically *for* complex higher-level problems) or heuristic-like thought processes yield not only sub-optimal results but can also lead to conclusions full of errors in judgement (Tversky & Kahneman, 1974). This is due to the fact that these processes, be they heuristics or not, do not consider all of the information given. In fact they don't even consider all of the information the agent has readily available. The processes seem at times purposefully ignorant of statistical aids. We seem programmed to follow these cruder methods. Now, it just so happens that this type of information blindness is something that comes up in other parts of cognition as well, like vision, for example: Shadows on a checker pattern board will trick our mind into perceiving shades of colours differently (Adelson, 1995). Even if we physically prove to ourselves that a shade within the shadow and a shade outside of the shadow are indeed the same by drawing a similarly coloured, and shadow independent, line through the two fields, the presence of the shadow still causes our minds to trick ourselves into perceiving a shift in shade of the monochromatic line we just drew. This kind of information blindness also just so happens to be present in the modular theory of mind, explaining

[20] This in reference to heuristics in programming. Generally speaking, researching and using heuristics as methods of logic to guide problem solving in programming, and indeed be programmed tools themselves, is an integral part of computer science.

[21] A few listed heuristics follow this structure. For example, the recognition heuristic: "If one of two alternatives is recognized, [then] infer that it has the higher value on the criterion." (Gigerenzer, 2008, p.24)

how and why these kinds of things happen. I think modularity of mind is at the very least a good model, a good abstract simplification, of how the mind works. Modularity also shares a lot in common with these heuristic like processes in our psychology, as well as the heuristic theories developed to explain them. I would say it's not too extreme a conclusion to draw that heuristic and modular models have a lot more in common than we might think, while also accepting that according to the evidence, they might just be a very good explanation for how a lot of the human mind works.

3.6 – Wrapping up modularity.

In this chapter already I have thus delved deeper into what modularity means and how it is structured. I have also gone into the subject of heuristics, showing how these tools are able to solve problems and make decisions in a fast and frugal, if also suboptimal way. I have also argued for why I think that these modules could be described as computational capacities, and in virtue of this making an example of how a computational system does not need to be information-hungry or greedy, but may develop tools to reach cheaper and sufficing answers, while only applying more rigorous faculties when needed. I have also touched upon the subject of mathematics and computation. In the next chapter I will be exploring the difference in mathematical abilities between humans and computers. After all, humans are really bad at mathematics when it all comes down to it. There's no human brain, no offloading strategy that can compete with a simple calculator when it comes to simple arithmetic. Is it that our brains lack the horsepower or is it that we do, in fact, approach mathematics and handle the cognitive processes differently, and what does that mean for the computational account? I will try to convince you that while it does have some consequences, it's not all that bad. In fact, you'll see that it's partially beneficial.

This is not to say that we *do* mathematics that much different from computers. After all, we taught/programmed our own mathematical methodology into machines. However, as is only logical, mathematics is not necessarily of a nature that our human mind is optimised for. As Penrose puts it, being really good at maths is not really a great survivalist trait in the way that it would've evolved as any type of primary cognitive faculty (Penrose, 1994). The kind of mathematics that we do on paper in academia and classrooms is an artefact that we are able to wield, much like how Gigerenzer points out that we are able to apply logical or probabilistic thinking at will - and often reach better results in environments suited for that kind of thinking (Gigerenzer, 2008) - but it is not part of our naturally

developed primary cognitive tools. There is, however, an invention of ours that *is* optimised for that kind of thinking by design: the computer.

Chapter IV – Mathematics and Vision

In this chapter, I will discuss how mathematics relates to human thinking. It may be tempting to point out that if we are computational beings, why are we so much slower at simple arithmetic than actual computers? After all, mathematics is simply the algorithmic manipulation of symbols, much like how I've described the claim of CTM. How do we solve this? In the pages below I will attempt to show how we as humans may be more horizontally capable when it comes to computational cognition, compared to the limited-purpose computers that we design, but also that for various reasons we are more vertically limited in our purely mathematical faculties. What I mean with horizontal and vertical is that we have a broad spectrum of tasks that we do well, but that artificial computational machines have better vertical capacities: their design grants them more powerful processing in a narrow selection of tasks.[22] After having a general look at mathematics, and revisiting the arguments of Roger Penrose, I will move on to introducing Kahneman's two-system model of the mind, where we differentiate between fast and slow thinking, and how this could potentially relate to our conundrum. Additionally, I will explore a theory of vision that is often seen as very closely connected to the core tenets of CTM; that is Marr's theory of vision. This theory is a computational *and* mathematical description of how visual data is processed by the mind and turned into images inside the brain. Marr's theory is rigorous and describes several steps through which we transform two-dimensional sensory data into three-dimensional representative objects within our mind. That is, the theory describes how we come to understand these images that display upon our retina as a world of depth, full of autonomous objects detached from each other. From a philosophical standpoint, this perfectly describes what CTM at its core is trying to say; that we computationally process content in order to create a representational version of the world within our minds, a representation that is not only a diorama-like picture of what we see, but also an understanding of the shapes and forms of the objects within, and how these objects are made up of separately identifiable parts.

4.0 – Human vs. Computer Mathematics

A computer and a human are both playing chess. The computer has been programmed with bottom-up computation, meaning that besides the basic knowledge of the workings of chess that it was programmed with, it has a cognitive architecture designed for increasing system complexity as it

[22] This would not necessarily be the case in the hypothetical existence of a general-purpose AI.

goes along, learning and refining its strategy through gathering data after each game and learning from mistakes. In addition, the computer parses information much faster than the human, making decisions within seconds or even less. In the end, the human somehow corners the computer into a stalemate where the human has a clear advantage. Should the computer break said stalemate by advancing and taking a piece, it would lead to a definite victory for the human player. There are two options for the computer; keep stalemate and make a draw or break the stalemate and lose. The computer advances and is after that defeated. Why? A case just like this was presented by Roger Penrose (1994) and was an actual problem for even the best chess computers around at that time due to their programming including the drive to always improve upon their material situation. This concept is widely known in chess as 'eating'. It is assumed you should eat when you can because improving upon one's material situation is always a good thing. However, there are exceptions to this norm, and unless specifically programmed into the chess computer, it doesn't seem to learn this on its own.

Chess computers are one of the products of the Artificial Intelligence endeavour. It is a scientific - and in foundational aspects, philosophical - endeavour to, quite simply, artificially create an intelligence, a thinking thing. The focus of creating a thinking machine often results in attempts at mimicking the qualities and psychology of the human mind. This is where philosophy truly enters the picture. The history of AI is intertwined with the history of the Computational Theory of Mind (CTM), a philosophical theory of mind that stretches back into the mid 20th century with the Turing machine (Turing, 1950), and its building blocks even further back to the late 19th century.

Roger Penrose, supporting his argument against computer-simulated human intelligence, points to the nature of our mathematical understanding as something that cannot be explained in computational terms. For example, he references Goldbach's conjecture; a to this day unsolved mathematical problem stating that every even integer that is greater than 2 can be expressed as a sum of two prime numbers. Anyone testing a few examples in their head or on paper can verify that each tested instance is true: 2+2 makes 4, 3+3 makes 6, 5+3 makes 8... This can be put under continued verification, and has been continuously verified by mathematicians such as Tomás Oliveira e Silva up to exceedingly high numbers, "yet it is such an awkward case that all attempts to establish it have so far failed" (Penrose 1994, p.193). What that means is that while this conjecture can be verified by examples - and while most people after several of these examples would be happy to say "yes, this is true" - it cannot be established as a truth by a mathematical proof through rigorous deduction. Despite all verifications, the truth-value of the conjecture remains inconclusive. Penrose makes references to other similar problems such as non-stopping algorithms (in the Turing sense). He argues that in order to judge that an algorithm does not stop (i.e. that there is no solution for it) one

cannot rely on a computable algorithm to reliably make that judgement for the system. On the other hand, many cases of non-stopping algorithms can be judged as true through common-sense notions of human mathematicians (and in many cases non-mathematicians as well). This would suggest that there is something non-computable going on, something that is not purely algorithmic that is helping us understand these things. What Penrose argues with these examples is that our human understanding of mathematics and our handling of truth-values on algorithms seems to involve something other than your standard mathematical methods, since these are insufficient to conclusively give answers to said problems. If we look at the human mind as something working with rational deduction and mathematical thinking, and since we are failing to put down answers on paper, we seemingly rely on some tool or strategy unknown to us in a direct sense, something that we are not aware of yet utilize to arrive at answers. This leads Penrose on a hunt for a plausible solution, arriving at the conclusion that humans have a unique sense of understanding mathematics that cannot be expressed computationally, i.e. it cannot be expressed in the same mathematical language that the problem is stated in.

On the other hand, humans are also often demonstrably bad at mathematics. We are outclassed at processing speed compared to most computers when it comes to calculating, and we often make mistakes or lose track when handling any and all kinds of mathematical problems. Kahneman (2011) makes a stark contrast between our ability to evaluate the expression of a woman's face and being faced with a simple act of calculating a multiplication. As humans we directly and intuitively recognise the mood of a person by looking at their facial expression. An angry person is immediately recognized as angry to us, there is no process of evaluation or measuring of the angle of the brows, no weighing of different traits counting for or against the conclusion that it is an angry face we see. It simply comes to us directly. We can also immediately draw further conclusions as to what we could expect to happen should the picture start to move (we can see if she's about to say or shout something, and we may predict the nature of what she's about to say on account of her current facial expression). In other words, not only do we immediately recognise what kind of facial expression this is, we also immediately mind-read and predict possible future events. On the other hand, when seeing 17 x 24 presented to us, things are different. We immediately recognize this as a multiplication problem, but we do not recognise the answer in the same immediate epiphany as that of a facial expression. At best, we can make a guess as to within what numerical range a probable answer would fit.[23] In order to arrive at proper judgement of the correct answer, we would need to start thinking. A noticeable kind of thinking, a slow kind of thinking, as opposed to the fast kind of thinking going on with facial recognition that we can only assume happens so quickly that we are

[23] That is, unless you have trained your memory to associate that particular multiplication with its product.

never consciously aware of it. The differentiation between these 'modes' of thinking is what Kahneman calls *fast* and *slow* thinking. When looking at these two cases from a computational perspective, it would seem that both should not be too different from each other. Just as the combined facial features of a picture of an angry woman are all component parts of ANGRY WOMAN in our minds (and in addition all the further information we can immediately and effortlessly extract from the concept in relation to context), 17 x 24 is but a representative formula adding up to 408, a formula possessing a syntax that we can understand.

When calculating 17 x 24, or any other mathematical problem, there is an event going on in our mind. We are thinking, we are calculating. Some people might experience a sensation of their brain straining, and some might not be able to solve it in their head but have to rely on pen and paper. Furthermore, even people who can solve it in their head need to resort to multi-step tactics of dividing up the numbers further and process the calculation little by little. In these calculations, several things can go wrong. We can forget to carry a number, we can draw an incorrect conclusion (such as 17 x 20 resulting in 360 instead of 340) and even lose track and "drop" the numbers we were "holding" in our head while tending to secondary and tertiary steps, thus forcing us to restart or give up. Trying to keep it all in the head becomes analogous to juggling too many balls.

Now imagine a computer being put up to the same tasks. The computer might need to possess some form of special facial recognition programming of course, but will still probably end up short in comparison to what any human can do. It will likely take longer and the results will sometimes fail (the computer might not recognise a face at all or see a face where there is none). In fact, Whitworth & Ryu say of state-of-the-art automated airport check-ins that "minor variations, like lightning, facial angle or expression, accessories like glasses or hat, upset them." (Whitworth & Ryu, 2007, p.230) And then, on the other hand, the computer will immediately and accurately calculate 17 x 24, without any risk of being mistaken or losing track. If the human mind is computational, why is it that humans and computers are different in this way, and does this pose a problem for CTM? One might think that the answer to the latter is yes, for computational theory of mind suggests that human thinking is processed through, of course, computations. Computation is simply processing of representational symbols through language-like algorithms, and simple maths problems should be fairly straightforward in such a system. Someone proposing that human thinking is computational all the way might be tempted to describe human beings as rational and logical beings, and yet we find numerical processing much harder to the much more subtle and complex ways a hypothetical visual module would be processing facial recognition. This becomes a problem, because mathematics is nothing but numerically quantified puzzle-solving and logic. Mathematical thinking involves deductive thinking, hard logic and probability, and is much more rigorous than, say, Gigerenzer's take

on heuristics which we looked at in the previous chapter. The latter is explicitly an alternative to logic and probability, based in fuzzy logic[24] and uncertain probability. If the human mind cannot excel at thinking within its own domain, then how plausible is it that rational and computational thinking really is our domain? One could answer that it might simply be a matter of computers having more processing power. This would explain why even simple single-purpose computers like calculators are much faster than humans at plain old mathematical problems, but pale in comparison to the many other things humans are better at, as "seeing" the validity of unsolved conjecture or accurately recognise facial patterns, both pointed out above. These cases of human superiority could then be ascribed to lack of proper programming on the computers' part. We simply do not understand enough of our own processes to artificially manufacture them in our computer devices, thus ending up with inferior thinkers on superior hardware. However, this sentiment is simply not true. In fact, the human brain has far more processing power than a calculator due to the difference in *sequential* and *parallel* processing. The neurons in the human brain are much slower than the circuitry in a digital computer. It is so much slower, in fact, that a million computer events can take place in the time span of a single neuron event (Whitworth & Ryu, 2007). However, computers are sequential. Supercomputers utilise *some* parallel processing, but not nearly to the extent of the human brain.[25] As such, computers bound by sequential processing need to take care of events one-by-one (although really quickly) while the brain processes several tasks at once, for a *much* greater processing potential. One can compare this to the difference between one quick person checking 40 jars for a hiding Ali Baba, and 40 average people lining up to check a single jar each (Whitworth & Ryu, 2007). The answer to the difference in human and computer performance in mathematics must thus lie somewhere else.

4.1 - Two Systems and Some Modules

As previously noted, Kahneman divides human thinking into two categories: fast and slow. This is, in fact, something that Kahneman takes quite far as a distinction, and is laid out in his book *Thinking Fast and Slow* (2011). To him, fast and slow thinking is not just a case of categories, but a case of systems in such a way that all fast thinking belongs in one system, and all the slow in another. These are of course relative terms, and to further clarify what “fast” and “slow” mean in this context he

[24] As opposed to binary logic (Boolean logic), where statements are either TRUE or FALSE (1 or 0), fuzzy logic deals in degrees of truth where the truth value may be anywhere in between 0 and 1. See Zadeh (1965).
[25] For a comparison of raw processing power, it took Diesmann & Morrison 82,944 processors and 40 minnutes to simulate 1 second of the human brain. See Wilkinson (2013).

further details the unique characteristics of each. In this theory human cognition is, as such, divided into a System 1 and System 2.

System 1 operates outside of our accessible conscious[26] will, automatically and quickly. Processes handled by System 1 will appear to us as immediate and intuitive conclusions and judgements. Output from this system is best checked or questioned as it is prone to biases and systematic errors.

System 2 requires cognitive "effort" to operate, handling complex computations that are apparent to our consciousness. These kinds of operations are the ones that we may relate to as 'thinking' or active deliberation. It is where we like to attribute cognitive agency. It usually prefers to operate in a comfortable low-effort mode (it's rather lazy) and receives impressions from System 1.

There is something important to be highlighted here of how these types of thinking relate to consciousness, or awareness. While it would be easy to say that System 1, being fast and automatic, are largely unconscious processes, while System 2 processes, being deliberate, are largely conscious, this is not entirely the case. The relationship between the System 1 and 2 divide, and the states of consciousness and unconsciousness, is more of a correlation, and examples can be drawn in both cases where this relationship is not true. The essence of System 2 is deliberate conscious initiation of the thought process. We do not start thinking in a System 2 fashion without consciously initiating it. This is why it is a choice if I want to go with the gut-feeling presented by System 1, or to apply further considerations using System 2. Another hallmark of System 2 is that we follow, or are aware of, all steps in the cognitive sequence, but this is where we are making a general statement that may not always be the case. For example, we often experience starting to think about a problem, not finding the solution, then waking up the next day to a sudden revelation, as if the thinking process that we thought we abandoned kept going in the background, and found an answer without our direct awareness. Similarly, we may experience gut-reactions that are fast, effortless yet we are very much aware of what reasoning lies behind them. Someone may suggest that we should run across the train tracks as a shortcut, and my initial reaction would be that it is a stupid idea. When asked why I would be able to defend my gut reaction, even though the thought process leading up to it was automatic. It would not fit to say that I was entirely unconscious of my System 1 process and only aware of the result.[27]

[26] Again, this is 'conscious' in the sense of access consciousness (Block, 1995). That is, we do not have direct interactive access to the initiation of these processes.

[27] Haidt (2001) would offer a counter-position to this. He writes of how our moral judgements are largely post-hoc justifications to emotionally based gut-feelings. In this way, we make up a story to explain our intuitions. There is an argument to be made that System 1 intuitions *could* be similar.

Does this division of fast and automatic versus slow and deliberate sound familiar? I think it does. It is very similar to modularity theory's peripheral vs. central processing systems as I have described them in the previous and introductory chapters. Just as in Fodor's classical view, System 1 takes up a similar role to the modular systems outside of central processing. They are encapsulated, to an extent, and largely inaccessible to our conscious will. That is, we cannot will a System 1 process to initiate; the processes are fast and operate automatically. System 2 is comparative to the central processing system, where the cognitive "agency" resides and thinking is slower and more obvious. According to Kahneman, System 1 generates the impressions that System 2 uses to form beliefs and make decisions, much like how a module would send data of relatively simple complexity to the central processing system to further utilise in more general and versatile ways. These systems are not completely dependent on each other. As mentioned before, one can go with the intuitions gained from System 1 without checking them in System 2, but System 2 can also operate while ignoring System 1. Going back to the first example: In this system divide, recognising an angry facial expression is something attributed to System 1, while attempting to solve a mathematical problem would fall almost solely under the responsibilities of System 2 (Kahneman, 2011, p.21). Only the initial ballpark figure or gut feeling about the answer (if any at all) would apply to System 1.

Evolutionarily speaking, System 1 is the system that we share with the animals. This system accounts for our perceptual faculties, instinctive habits and emotive responses to situations. In contrast, System 2 is a relatively recently evolved part of our cognitive system, and something argued to be largely exclusive to humans (Evans, 2003). In this sense, System 2 is what allows us to perform abstract hypothetical thinking, helping us to predict future events and adjust our actions accordingly on a much grander scale than what instinctive behaviour allows for. Take learning for example; an animal without reflective or abstract thinking is able to learn things about the world and adapt and improve its behaviour over time in order to better accommodate for future situations, but it can only do so by learning through experience. Such an animal cannot prepare itself for future events that it does not have the opportunity to first learn about first-hand. For example, an animal of that kind cannot make a plan for how it will react during the event of a natural disaster, barring that this has somehow happened to the animal before (and it survived) or if such disasters are common enough that the species as a whole has learned how to react to it through instinct and parenting. But then take something more rare and complicated, like a nuclear war. Such an event is likely to only happen once, if ever, and would bring catastrophic consequences that in no way can be learned about through first-hand experience. In fact, a nuclear war has thankfully never happened. Yet, we possess the forethought to take measures to avoid nuclear war. How do we avoid nuclear war? We hypothesise steps that would lead to nuclear war, and we avoid those, or we work on counteracting

events that we predict would lead to such a situation. For that, we need to be able to apply abstract hypothetical thinking. We need to learn about these situations *before* we actually experience them (Evans, 2003). As such, being able to simulate these hypothetical situations in one's mind is very beneficial - even if it's just something as benign as a teenager rehearsing lines for a date or mentally running imaginary scenarios to prepare for the worst. This kind of learning capacity is beyond the scope of System 1, and is thus a System 2 level of learning. This is not to say that System 1 cannot learn things, like pointed out before; it merely learns at a more basic and direct level.

While System 1 is fast, System 2 is slow. Since System 2 is the domain-general system for our deliberate thinking processes. The thinking that takes place is very much how we may experience a train of thought; it is a sequential, step-by-step process that requires effort and operates from our working memory (to use a computational phrase). In contrast, the automatic processes of System 1 are not limited in the same way and can run parallel to each other. Like I mentioned before, System 1 is comparable to modular systems operating outside of central processing. As such, it would be wrong to call System 1 a singular system in the true sense, for it is in fact a collection of several System 1 modules all operating independently of each other, each for their own domain-specific tasks.

Thinking about System 1 and 2 in this way, could we explain a computer's shortcomings in facial recognition in modular terms? Yes, it can easily be described as humans having a much more sophisticated visual module than the computer, which again goes back to the programming argument earlier. It would then not be too far-fetched to say that the reason we have a harder time calculating maths is simply because we lack an appropriate System 1 kind of faculty for that kind of problem, i.e. we lack a mathematics module. After all, modules are context sensitive and it could be the case that mathematical problems presented as numbers cause most of the workload to be handed to the central processing system. Even with massive modularity, one could still argue that the lack of a specialised mathematics module results in the cognition being handled in a less-than-frugal non-modular way.

I am willing to admit here that I might be too optimistic in my comparison of these different mind theories. Kahneman's work focuses on the effect of biases on our ability to judge and choose, as well as exploring our intuitive abilities (Kahneman, 2011). Just as Gigerenzer, Kahneman deals with heuristics. However, the two differ in that Gigerenzer is presenting an account for why heuristics are useful to our everyday decision-making, while Kahneman is more concerned in pointing out how our biases in System 1 operations cause irrationality and bad decisions. Kahneman's view of our

intuitions is thus more negative than Gigerenzer's. Where one sees a sufficiently useful toolbox, the other sees problematic systemic errors in judgement.

Furthermore, Gigerenzer sees the distinction to be made as being between rational thinking and an offloading toolbox used for quick decisions. What bridges these theories with modular theory is the common theme of decision-making, context sensitivity and limited and frugal use of information. It is the important core that remains and tempts comparison as they arrive at roughly the same answers for the same kind of problems. If they can somehow be used to support one another, that would be a great discovery in the endeavour to solidify at least common aspects of these theories as plausibly true.

Even if we take modules as being the answer as to why humans are better than computers at recognising faces, there is still the problem of the computers' superiority in mathematics specifically, and the mystery of why we humans are so inferior at it despite greater processing powers, as stated previously in the sequential/parallel processing comparison (Whitworth & Ryu, 2007). There is also the issue of the cognitive struggle required to keep numbers and equations *in the head* and our general practice of offloading temporary data storage onto paper notes and similar things. In the following two sections I will tackle these two issues and possibly point out some counterexamples to some of these norms.

4.2 - Size and Storage

Digital information storage allows computers to store information and utilise it in later processes. This is often referred to as "memory". Humans can also remember things, and what we can't remember we can offload onto the world in the form of personal notes. If computational theory of mind is true, external and internal storage of information for humans is quite similar: both involve the manipulation of symbols to retrieve information. In the mind these are mental symbols, and in a notebook these are words in human languages, which in turn require a "decryption key" of sorts, that is an understanding of the relevant language. Unlike the case for computers, humans require a lot of effort with their internal memory at times, especially when it is involving a lot of information partaking in one process. A human can store memories gathered over an entire lifetime, not all of them accessible at all times. We forget things and remember them later. Sometimes an external cue can trigger a memory, such as an old photo or a note on a calendar. When we retrieve something from memory unexpectedly in this fashion, it feels as if we had forgotten it and just then regained access to that particular memory.

When calculating a mathematical problem, even quite simple equations can prove difficult to keep all in our heads: In modular terms, in central processing. In computer terms, in cache. It is as if, comparative to a computer, humans have a limited working memory for just how many mental symbols we can juggle at any one time. If we keep too many numbers in the air, we are likely to forget some or lose track of our line of thought. If I were to tell you, the reader, to solve the following, you would probably have no problems solving it all in your head: 2+3*5. Assuming one knows the order of operations (BODMAS: **B**rackets, **O**rders, **D**ivision + **M**ultiplication, **A**ddition + **S**ubtraction), i.e. you know the language and how to read it, this is easy and you'll arrive at 17. Now try the following:

3*5/2 + 3(2^4 + 7) - 2(5 - 3)(3*4)/5 =

Solving that in the head will probably be quite difficult for most, at the very least it will require memorisation of individual values as operation moves onward (the answer was 66,9). As more and more individual problems that require separate attention appears, it will be harder and harder to contain it all in working memory without some sort of external help. Now compare that to this:

2 + 3 + 7 - 4 + 5 - 6 + 2 - 1 - 4 + 6 + 7 + 8 - 9 + 4 + 6 + 3 + 7 - 8 + 5 + 3 + 6 + 14 + 30 - 12 + 5 + 11 + 7 - 2 + 14 =

Arriving at 109 here is a lot easier due to the fact that only a single number has to be remembered for each step as we move through the operation. However, a single error will cause further errors down the line. And humans do make those errors all the time. Perhaps it's due to hasty misreading or some kind of cognitive interference, but sometimes 35 - 7 becomes 29. So could it then be assumed that humans have a very limited working memory storage?

According to psychologist George Miller, who invented what would later be called Miller's Law, experiments show that on average a person can hold about 7 objects in short-term memory (Miller, 1956). This means a person can on average solve an equation with 7 active numbers held in short-term memory. Beyond that, we are prone to errors or simply overload and mind-blank (to follow the juggling analogy, we drop the balls). This does indeed seem like quite the limited memory when compared to most modern day computers. On the other hand, human long-term memory is in contrast both flexible and vast. With enough rehearsal and reinforcement, short-term memories become long-term memories. As such, given enough time and dedication human memory can achieve quite remarkable things (anything from ancient Greeks memorising epic poems to savants memorising digits of *pi*). However, feats of memorisation in this way do in no way help humans in the race against computers. If one subscribes to a loose version of extended minds and Otto's notebook

(Clark & Chalmers, 1998), one might be inclined to allow humans some leeway in judging our mathematical skills, as the use of external objects greatly improves upon this ability, but the fact remains that compared to computers, human working memory in brain alone is suboptimal when handling big problems as we are prone to losing track when keeping too much in working memory at the same time.

Going beyond working memory, computers have recently outclassed humans in the long-term as well. Mathematicians Boris Konev and Alexei Lisitsa (2014) have recently discovered a proof for the Erdős discrepancy problem using computers. The problem with this proof, and why it has raised some controversy, is because "[i]ts gigantic size is comparable, for example, with the size of the whole Wikipedia, so one may have doubts about to which degree this can be accepted as a proof of a mathematical statement." (2014, p.6) This mathematical proof is so big that no human can confirm it to be true, long-term memory usage or not. This has caused a stir in the mathematics field as this is a step into non-human mathematics where we need computers to check.

4.3 - Fast and Slow

The second issue is the slow processing speed of human arithmetic compared to even something as computationally simple as a standard calculator. This slowness is despite the availability of parallel processing. What makes computers calculate fast and humans calculate slowly? Perhaps the most obvious suggestion is that there is a distinct lack of a specialised arithmetic module in the human mind. If there is no such thing as an arithmetic module, or if any kind of arithmetic system is at best rudimentary, then that would make a strong case for why humans are lacking in raw calculation power.

Imagine all the small processes going on in your mind and body. There are so many processes going on that we hardly notice on a daily basis. Take breathing for example. Have you ever gotten the feeling that you've just realised that you need to actively breathe? Once one starts thinking about the process of breathing one often finds oneself having to put effort into the breathing process, taking the reins and controlling the flow and rhythm of air going in and out. The same thing sometimes happens with walking. Such a simple thing as walking becomes a much more tedious task when one is concerned about keeping a specific pace that is outside of what comes naturally. The point I'm trying to make is that something we are directly aware of is much more straining than something that we do by habit. Jogging is easy as long as one doesn't pay attention to the remaining

distance or the impact of each step. I find from personal experience that "zoning out" helps keeping the pace and often results in a recovery from perceived fatigue.

Kahneman's System 2 points to an aspect of human cognition that is often downplayed in modular theories, massive modularity most of all, and that is the active taxing cognition. It is an attractive view that most of our cognitive processes are relegated into appropriate modules that work away on their own, but there is a form of agency that is often forgotten. This agent is a form of meta-cognition. It is the central processing system, the conscious thinker, the System 2. As Kahneman describes it, System 2 is for tasks that require attention (2011). When these tasks are approached and System 2 is properly utilised, it is in a manner that is deliberate and (in relation to System 1) careful. This deliberate nature requires more cognitive effort on the part of the thinker. For a person with dyslexia, reading is a struggle and requires focus and effort. In contrast, for a person without dyslexia, reading comes with a flow where information is gathered from the text at a non-taxing pace. Going a bit further, for a person not used to academic writing, reading an academic text can also be taxing. The same can be said of reading any text in a subject with which one is not familiar. It is not only that these texts might be more thought-provoking, but the language itself requires more effort in the form of decoding into "mentalese". In modular theory, the language module is one of the most explored concepts, thanks to Fodor (1975) and Chomsky (1968). If modules are a true way of describing the mind, one could suggest that maybe a person suffering from dyslexia has trouble with their language module. One could say that in the case of a person that possesses a module for a certain task and one that is lacking that same specific module, the difference between these two in their abilities regarding that kind of cognition would be the kind of difference in reading ability that we see between a dyslexic and a non-dyslexic person. What I am trying to say is that even though humans study math as a subject, we might all be a bit dyslexic when it comes to arithmetic.

So is it likely that we lack a module for mathematical calculations? As Penrose states in *Shadows of the Mind*, it makes much more evolutionary sense to have a general understanding of maths than it is to have cognitive faculties designed to handle complicated mathematical problems. As he puts it: "For our remote ancestors, a specific ability to do sophisticated mathematics can hardly have been a selective advantage, but a general ability to *understand* could well have." (p.148) Even though mathematics has been with us since ancient times, one could argue that it could hardly have been that strong a trait to give selective evolutionary advantage, and then have time to evolve into a fully-fledged module. However, there are those who are more positively inclined to the existence of a mathematics module, especially for simpler forms of arithmetic. Take for example Dehaene (2011) who suggests that we possess a rudimentary ‘number sense’ for simple maths, something even

present in us as children, as well as in certain animals.[28] If that is true, then we are, at least on some level, capable of fast mathematical calculations. We can also take a look at perception and see how computations take on an almost mathematical shape. We as humans constantly involve ourselves, unknowingly, in distance measurement, measuring angles, counting objects,[29] determining symmetry, predicting trajectories of flying/falling objects and all kinds of additional things involved in the task navigating and interacting with a three-dimensional world. These tasks come to us naturally and feel automatic, leading us to believe that a great deal of these capacities lie in peripheral systems.

Additionally, not all mathematical capacities need to be arithmetic, and not all mathematical operations taking place in the mind need to be expressed in numbers. Numbers in measurements are abstractions, part of a language we created in order to calculate things beyond what our minds alone could do, beyond the capacities of our modules. While a group of three bottles on a table may hold a mental representation of THREE, it would be less intuitive to say that the baseball player measuring the trajectory of a falling ball would have mental calculations going on in his head that are similar to the diagrams we would draw up on paper. Yet, this ability we have of determining vectors in 3D space does seem to take on a mathematical nature if we are to express such abilities computationally. Humans may be slow at advanced arithmetic calculations, but we are quite apt at intuitively determining the path that moving objects will take. For example, when jogging down a road I may see someone else who is about to cross my path. My brain is able to tell me, accounting for speed, distance and angles, if I need to slow down/speed up to avoid bumping into the person at the point our paths cross. In this sense, we may have modular capacities of a mathematical nature that goes beyond basic number senses. So why are we still slower than computers at arithmetic? Because the more advanced aspects of our maths-based modular capacities are not of that domain, but rather tied to vectors, geometry and (as I will explain below) vision.

4.3.1 – Mathematics in vision

Above I made examples of how a rudimentary number sense may be involved in vision, in the sense of quickly counting groupings of items (like bottles) to be represented by a single numerical in our

[28] For other sources with similar angles, see Devlin (2001) and Schliemann & Carraher (2002).
[29] If one imagines two tables with bottles stacked on each, at lower number ranges we can quite handily determine which table holds the greater number of bottles, without necessarily counting the bottles in a system 2 fashion. With larger numbers, we may rely on density, but when it comes to smaller numbers we clearly 'see' five bottles as a group of '5', and that group looks different from a group of three bottles. The types of objects don't matter for this recognition, only the number. Five bottles and five pencils are both equally groups represented numerically as '5' (or FIVE). Thus we know, intuitively and universally, what a group of 5 objects looks like.

mind. However, I also argued that we have other capabilities tied to visual measurements that can be expressed as mathematical in nature. David Marr (1982) argued for a model of vision that carried out such vector and geometry-based mathematical computations in order to produce visual input from the eyes into the brain. Marr's theory of vision involved computational processes, akin to the type of computation envisioned in classical CTM and he argued that these processes taking place in vision could be abstracted as mathematical formulas. Take for example his statement of the function of the retina:

"Take the retina. I have argued that from a computational point of view, it signals $\nabla^2 G*I$ (the X channels) and its time derivative $\partial/\partial t(\nabla^2 G*I)$ (the Y channels). From a computational point of view, this is a precise specification of what the retina does." (Marr, 1982, p.337)

Marr's theory of vision introduces a thorough computational (and mathematical) proposal for how our vision works, from light hitting the eyes to a fully represented image in the mind. Specifically, Marr's theory tries to explain how we translate two-dimensional input into three-dimensional representations in our mind. I will dedicate the rest of this chapter to exploring this theory and its history, showing why it was (and still is) a respected theory for visual computation. However, Marr's theory is not without problems. Unfortunately, the failings and issues with Marr's theory have, due to its connection to CTM, caused similar criticisms to be levied against CTM as a whole. As we will see below it is a very rigorous model of vision, requiring a great deal of processing to keep the flow of input going. As such, it lacks frugality in comparison to more recent alternatives like Predictive Processing (Hohwy, 2013; Clark, 2013, 2016), another theory of vision that I will look at in the next chapter. My aim in the coming sections is to explore the idea of (modular) computation in vision, but also go into further detail as to why Marr's theory has become a convenient specific target for general criticism against the computational theory of mind. In the next chapter, I will argue why these criticisms miss the mark when it comes to more contemporary computational theory, by showing that we have other, equally computational, options in the aforementioned Predictive Processing.

4.4 – Marr's Theory of Vision

Marr's computational theory of vision was developed during the 1970s and was posthumously brought forth as a coherent whole in the publication *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (Marr, 1982). By then, the theory was already known, and had been presented to contemporaries for analysis and review through its

appearances in publications such as *A computational theory of human stereo vision* (Marr & Poggio, 1979), *Visual information processing: the structure of creation of visual representations* (Marr, 1980) and an attempted summary in *Marr's computational approach to vision* (Poggio, 1981).[30]

It is important to note that this theory was developed from a neuroscience perspective for how the visual system recognises and evaluates geometric shapes in the visual field based on raw data signals. However, the elements that the theory touches upon made it naturally interesting for philosophers of mind and researchers of artificial intelligence.[31] The aspect of the theory that has been of the most interest to philosophers is perhaps its most basic claim and premise: the transformation of a two-dimensional image on the retina into a three-dimensional object in the human mind. Not surprisingly however, this highly computational approach to the human visual system has also fallen under the same canopy of criticism from the EEEE theorists as all other aspects of computationality. In fact, Marr's computational theory of vision is a convenient example to draw upon, since the very discoveries that inspired theories such as enactivism also proved that Marr's theory was dated (Clark ,2013). In addition, the claims of Marr and the claims of enactivism and its peers are directly opposed to one another. As put in *Radicalizing Enactivism* "REC [Radical Embodied Cognition] rejects the idea that the exercise of *conceptual* or *representational* capacities is needed and instead looks to the history of organismic activity as providing the something extra." (Hutto & Myin, 2012 p.17; my emphasis) The basic premise of Marr's computational theory of vision is the transformation of 2D input (in this case two two-dimensional images entering the cognitive-visual system as data via the retina) into 3D representations. As such, the existence of internal representations is an integral part of Marr's theory and as I have argued earlier; representations and internal mental content are an important part of the computational theory of mind. Enactivism often finds itself to be in direct opposition to this type of internal content (see Noë, 2009; Chemero, 2009; Hutto, 2011a) much in the line of the type of dynamical systems theory found in *Dynamical approaches to cognitive science* (Beer, 2000). The relation between dynamical systems theory and enactivism, as well as the state of internal representations in enactivism is something I will go further into later on in this chapter.

[30] Tomaso A. Poggio worked with Marr at several instances of the development of the vision theory and has many insightful things to tell about that period in the afterword of *Vision.*

[31] I'd argue that when it comes to matters of mind, brain and computational theory and the relations between them, separating these fields, especially in the 70s and 80s, is largely pointless as they were - and still are - all players in the more specialised yet widely multi-disciplinary field of cognitive science.

4.5 – Marr's four levels of description

Marr argues that vision involves the processing of information with the purpose of capturing data about the world and transforming it into a representation that is useful to us. This representation is what our mind uses to conceptualise what we perceive, such as animals, trees, geometric shapes and the increasingly specific subsets of these. The information processed in this way is not tied to the physical constraints of its hardware (or if it is, only loosely). What goes on at a purely physical level does not suffice to explain the depth of what takes place on a much more abstract or "mind"-level. As such there are several levels of description through which we can describe, study and explain the processes of a computational system. In such a computational system that Marr would describe human visual system as, the levels of description would be three-fold; computational theory, algorithm and implementation. The latter would in this case be the more physical level, which in turn can be divided into mechanisms and components, for an actual number of four levels of description (Poggio, 1981). Computational theory is by Marr considered as the top level of description. It is the level through which we explain and describe the abstract processes and properties of the visual information. I am going to go through each of the four levels of description here in order to accurately outline what each level encompasses. It is important to note that these levels can be studied or used as frameworks for discussion independently of each other, though each higher level in the hierarchy will by its nature involve elements of the lower levels beneath it. By going through these levels of description I hope to also be able to highlight just how essential representations are to Marr's theory, which will be of importance later on as I will continue my defence of internal representations in cognition as a whole.

The lowest level of description for Marr's computational system is the analysis of the components of said system. In a system like mathematics, or more specifically in a system designed to calculate and derive value from a certain type of mathematical problem (this could be a computer program or more generally a certain area of mathematics), these components would be the fundamental symbols and numbers representing values. In a system like a computer, these would be the parts of the circuitry. Finally, in a brain the bottom level would constitute the neurons and synapses. The bottom level constitutes the analysis of these components, the understanding of how they work. Whenever cognitive scientists consider the workings of the brain by looking at exactly how neurons, synapses or more broadly how the physical aspects constituting the brain work, it is this lowest level of description that is being talked about. Such analysis often flows over to the next level going up: the study of mechanisms. To once again draw an analogy to mathematics, this is the level of description that involves the study of adders and multipliers. These are the mechanisms through which the numbers on the lowest level are manipulated. But it is not the manipulation itself that is of

import here, but the analysis of the mechanism. Here Marr's example for a mechanism of the brain is memory (Marr, 1980, p.199), though I think the study of neurotransmissions seems a more apt example, as analysis of memory may be getting closer to the third level which I will explain now. To discuss the nature of neurotransmission itself, as a mechanism, is second-level. To describe how neurotransmissions function in a set subsystem like the visual system, and how they transfer information from the retina to the visual centre to the brain, that is to step into the third level of description: algorithms. This is where we start to leave behind the purely physical aspects of description present in the lower two levels to instead focus on not the components or their manipulation but the application of both of these into algorithmic information processing.[32] While discussion about the information gathering of the visual system does indeed involve mention of synapses, transmissions and neural pathways, it is the step-by-step *processing* that is of relevance at this level of description. In mathematical terms the third level is simply the level involving specific algorithms made up of numbers and the adders and multipliers to transform these values into the information that is desired, structured into steps so that the input is processed and transformed into relevant output. It is the method through which we manipulate one set of values in such a way that the end result gives us information in the form of another desired value. When discussing an algorithm for solving a certain type of mathematical problem, or when drawing up an equation to calculate a specific problem, the numerical values, while relevant, can take the back seat in the analysis without much trouble. In fact, for any algorithm the individual numbers themselves are merely applied values, as its structure of mechanisms has created slots for values that can vary greatly between individual instances of calculation. An algorithm can easily be discussed without any meaningful values present. In fact, it is often easier this way. It is easier when explaining an algorithm to a class of students to substitute specific values with generalised substitutes and placeholders, replacing "1","2" and "5" with *x, y* and *z*. *x, y* and *z* would not produce any meaningful output if put into an algorithm requiring numerical values, for it is the numbers that carry meaningful content to be processed. Yet, an algorithm can be perfectly understood with placeholders like *x, y* and *z*. This is because the third level of description is about *describing* the algorithms, and while understanding an algorithm would call for some competent level of knowledge about the two lower levels of components and mechanisms, I need to understand what numbers are and how to do addition and multiplication, these do not themselves require description for the algorithm to make sense.

[32] This is my reason for setting memory aside as an example of second-level description. I think the discussion of memory is much more apt as analogous to algorithms than as a mechanism in and of itself. That said, one can argue that memory is a mechanism, as a driving force that introduces knowledge, in an algorithmic system like visual processing. In the case of "visual data of X + prior knowledge of X = recognition of visual of X", memory could be denoted as the adding mechanism, 'fetching' the prior knowledge so to say.

Thus far we have three levels of description; two lower levels falling under a common label of implementation, involving more physical elements in relevant systems, and one higher up, involving the more abstract study of structuring the latter two to produce output from meaningful input. One could think that this is sufficient. The algorithm level successfully describes a visual system by including the neurons, synapses, impulses, transmissions and the shape of the structure which they take in order to process input from the retina into output in the visual centre. But Marr was not happy with this. Instead he suggested a fourth level: computational theory. "Neuropsychology and psychophysics … *describe* the behaviour of cells and subjects, but do not *explain* it." (Marr, 1980, p.203) The top level of description is not so much a level of description as it is set out to be a level of explanation. This is the level at which we explain what is going on, on a more abstract level set apart from the basic components. It is this level that imposes meaning upon the lower three. While a computational cognitive system may involve the description of perceptual systems, memories, concept formation etc., it is the top level that describes the cohesive theory behind these, addressing what memories, thoughts and vision are and what they mean to us. This level is also the most abstract because of this. In cognitive science, the expertise and input of us philosophers is very often confined to this top level. It is important to note that "top" level does not imply any form of superiority either. The study and understanding of each level is equally important in forming the whole, but no matter what area one is interested in (components or algorithms) in Marr's model, any such research would no doubt, though to varying degrees, incorporate the description of theory. In any computational system, no matter how mechanical or anchored in the physical, this top level is always an abstract one, for it is a description of theory to be applied to the lower three. This is where Marr's theory of vision will take form. The reason for why he presented this four level model was precisely to describe what he thought was a problem with the study of vision at the time: vision could be described neuroscientifically but lacked explanation. Marr set out to make a computational theory *explaining* vision, what it does and what it meaningfully involves.

So what exactly *is* Marr's theory of vision? I've thus far explained *why* Marr wanted to make a theory of vision but not touched much upon the theory itself. Well, to put it in simple terms; at its core, Marr's theory of vision is a step-by-step explanation of how representational cognitive (visual) content is transformed from raw input into a primal sketch, and then further into a representative 3D model in our mind. It is a very relevant theory for this thesis, as it incorporates the fundamentals of the early computational systems, representations and algorithms, input and output, while also proving a great example of where computational models can go wrong and how they need to adapt in order to fit into the modern landscape of cognitive science.

4.6 – Three Steps to Representation

There are three main "steps" of representation in Marr's theory: the primal sketch, the 2,5D sketch and the 3D model. Experientially speaking, it is the 3D models that feed into our conscious visual experiences of an object or scene, for they are the final output in the unconscious (accessibility-wise [Block, 1995]) mental process of transforming visual input into mental representations. Even when we imagine objects in our mind, it's these final 3D models that are being used (or you could say "displayed in the theatre of the mind"), for they are the representations of the various objects in the world as we understand them. I would like to address here that this interpretation of where conscious experience enters the picture is not entirely uncontested. Jackendoff (2002) explores the perceptual architecture of Marr's theory when exploring the taxonomic structure of lexical items. In his interpretation, anything beyond the primal sketch (and even parts of that) includes elements that we are certainly aware of, these elements being called "percepts" (p.328). These percepts are a step ahead of Conceptual structures and Spatial structures that he mentions later (p.346), and as these are described as basically equivalent to Marr's 3D-model representations (more on that below), it is only logical that the percepts have more in common with the 2,5D images or some half-step in-between. While I try to stay away from the subject of consciousness and phenomenal experiences in this thesis, we could describe these percepts in computational terms as being cognitive elements that are sub-personal, inaccessible or that we are otherwise unaware of, just as I am unaware of the electric signals in my brain.

Mental representations were very important to Marr when he constructed his theory, and he agreed with the notion that when we think about objects, they are represented in our mind as mental objects that we manipulate. The kinds of discoveries that led him to this notion were for example Shepard & Metzler's experiments on mental rotation (1971). In these experiments, the speed at which subjects could discover if two differently rotated 3D objects possessed the same shape depended on the differentiation of the angle between the two objects, with greater differentiation bearing negative impact on the time it took, suggesting that the subjects could mentally simulate rotating the objects in their mind in an attempt to compare the two. For this to be the case there would need to be mental objects to rotate, thus making a case for the existence of these as representations. Marr (1980) identifies the three steps of visual processing as follows:

The first step is the derivation of a raw primal sketch. A raw primal sketch is nothing but a very simple outline of important points of reference and angles in an input image. Marr very much treats vision as involving images, analogous to how a computer or a digital camera does it; information is gathered on the lens/retina and an image is formed from this. The primal sketch is the first step in

this and the derivation of this sketch involves combining zero-crossings found in the image into edge-segment descriptors. Zero-crossing is a term for the point at which a value goes from a positive into a negative (or vice versa). In this context, it is the point at which one shade of information (for these purposes black and white information "pixels") changes into its negative value. To put it simply: a raw primal sketch consists of "pixels" of black and white, where the zero crossings mark the boundaries between these. These zero crossings are then segmented in order to create reference points for the edges of objects and potential depth changes. These points are further processed and grouped together to form place-token units, the outlines of individual objects and artefacts. Virtual lines are then drawn to further define and represent the local geometry of these objects. The result is a sketch-like image showing the changes in light and vague object shapes, comparable to a picture displayed in an intensity array, where dots are arranged in varying intensity to form a grayscale-like pattern.

This sketch is then decoded by the brain. This decoding involves segmentation of the sketch into regions with meaning. That is, the mind applies knowledge of objects in the world to segment the information in the image into such objects, interpreting ball shapes or table shapes into ball regions and table regions, fencing these regions off and defining them as such objects. As such, Marr's theory of vision is quite information dependant, and a lot of knowledge, like awareness of shadows and depth, goes into interpreting the raw data. This way of imposing meaning upon an image via knowledge application was mentioned to be inspired by Tenenbaum & Barrow's experiments on the role of interpretation in segmentation (1977), and Eugene Freuder's (1976) computer system designed for recognising hammer heads in images, then searching for an appended shaft to confirm the object as being a hammer.[33] This latter example is particularly interesting as it involves applying specialised knowledge to confirm the identity of a whole object with but a single element of the object as a starting reference point. It could be comparable to recognising an eye, then searching for mouth, nose, eyebrows and other features to confirm the visual experience of a face, then further taking the face as a reference point for identifying a human being. While these kinds of programs were less than successful at the time, not able to move onto more advanced things, Marr attributes this failure to the lack of formulation in the top-layer description, further emphasizing why he believes the fourth level of theory to be the most important to focus on. This is not to say that the fourth level is more important. Rather, it is that the fourth level was previously neglected, and is now in need of more focus.

[33] Since the head of a hammer is the most unique shape, this was the segment of the object that was prioritised when searching a space for a hammer. When the head of the hammer was identified, then the system would look for an attached shaft to confirm that this was, indeed, a whole hammer.

Of course, humans view things in three dimensions, and in order to get to our 2,5D image and eventually a 3D model we also need to determine depth. In order to do so, Marr explains, we need stereopsis, which he argues is a relatively early process occurring at the tail end of the primal sketch phase, but ahead of the 2,5D image phase proper. In stereopsis, or binocular vision, the brain compares two images (at this stage: two sketches) and compares points of reference. The fact that there are actual points of reference selected and compared in the two images is a necessary thing according to Marr, since without that there is simply no comparing the images. This line of thinking, I will show further on, is one of the weaknesses in Marr's model and greatly improved upon by predictive models. In either case, Marr recognises that there is a problem in this approach: How does the brain select reliable points of reference from the sketch? These points of reference have to represent, in both retinal images, the exact same location in the world for a judgement of depth to be as advanced as it is in our vision. Since Marr's primal sketch consists of grey-level shades created through zero-crossings, virtual lines and "pixels", one could suggest that these points of reference need to be selected from areas in each image with similar shades of grey, but this according to Marr is not a workable idea for there may be several locations in each image sharing the same shade of grey. Instead, these points of reference need to be objective markings taken from intensity changes in the image. This Marr calls the stereo problem, and it has two rules for how to compare the two images in stereopsis:

The first rule is uniqueness. The two images in stereopsis are bound to be unique from one another. After all, the images are to be compared for disparity and thus no valuable information can be gained from such a comparison if the two images are identical. However, the images can only be unique in one value: position. The information in an image can only be shifted and differ from the second in virtue of being at a similar but still separate vantage point (this being the left eye sitting slightly closer to the right eye, creating a slight difference in angle). If the images were unique in further ways, this would disrupt the information gathering since there would be no way for the system to determine what differences were due to angle and which were due to other circumstances. Imagine for example if my left eye image contained objects that were simply not present in my right eye image. This would disrupt my depth perception as I have, under this model, no way of gathering objective markers, and any gathered from the unique object would not find a correlation in the image where it is absent. Take an extreme example of the two samples being unique to the point of dichoptic presentation, where two completely different inputs in left and right eye causes the brain to make an alternating approach rather than a blended one, viewing one image at a time as the viewer experiences two images switching back and forth (Wolfe, 1983). This example

is of course extreme but serves to show how uniqueness has to be limited in order for stereopsis to function.

The second rule is that of continuity. When viewing a set of frames in a movie or a series of changing images in a flipbook, the continuous similarity accompanied by the gradual changes creates the illusion of motion. Similarly, Marr argues that it is in a similar way that the human brain computes structure from motion. It is important to note that in Marr's model, we are actually dealing with mental 'frames' and individual sketches from which we derive visual information, so the analogy to flipbooks and movies is particularly apt in this case. Marr makes reference to research done by Ullman (1979a, b) in which Ullman explains the derivation of an object's shape from motion through the process of identifying continuous elements throughout the frames and via this determine shape in virtue of the change in the elements' positions. As such, we arrive at the following: In order for a primal sketch to be decoded and translated into a 2,5D sketch via introduction of depth elements, two things will have to be true. Firstly, two primal sketches from two different vantage points will have to be compared to one another, sharing values in visual data with the only discrepancy in these values being the differentiation in viewpoint positioning (i.e. two eyes next to each other). Secondly, there will have to be a continuous series of these frames from which motion can be interpreted and thus further shape can be computed. When these two are true, stereopsis is achieved.

4.7 – Attaining a 3D-model representation

According to Marr in the previous section, in attaining stereopsis we gather information from two sources: stereo disparity and motion. When introducing the 2,5D sketch however, Marr adds some further sources of information: shading, texture gradients, perspective cues, occlusion and contour. Important to note is that Marr maintains that these information channels rely exclusively on information from the image itself and not any *a priori* knowledge about the objects in the image (but may involve information about shadows and such). Any a priori knowledge used for object recognition comes later. As such, the data gathered through these sources relate to and describe the image as a whole (in the form of a "scene") rather than any specific objects within (Marr, 1980 p.208). The 2,5D image is thus a representation of all the surfaces in an image derived from a primal sketch, at this stage making all shapes and orientations *explicit* rather than inferred. Instead of seeing mere changes in light/dark intensity derived from zero-crossing data, we have now moved onto depth and even further into contours, shading and shapes. This is starting to look like an image containing objects. We cannot mentally rotate just yet, nor are we at a proper representation of objects within the world, but we have arrived at an explicit, if mostly mathematically expressed,

image of what our eyes see. It is important to note that Marr still views the information present here as largely mathematical. To imagine a 2,5D sketch to be something akin to an art student's still life drawing would be wrong, though not entirely off the mark. Marr makes it clear that he still imagines a 2,5D sketch in terms of data, visualised as dots on a field, now with additional informational cues to describe depth and shapes. Largely, this does not matter, for the 2,5D image is a representation not accessible to our visual experience anyway, and talking about how one can visualise it would be frivolous if taken too far.

The main purpose of a 2,5D image is to take the information gathered in a primal sketch and make it consistent with the nature of the world around us. We live in a world of shapes and orientations, and the job of a 2,5D image is to represent this. In this view, we are very much locked inside of our heads, and this image is the first step in letting us know how things actually are on the outside through a visual medium. In essence, this is the first half of the input-output sandwich which classic computationalism has been accused of, and while I do not necessarily agree with the negative connotations of this model, it is easy to see at this stage why Marr is a prime representative of this viewpoint.

Once the 2,5D image is created and all surfaces identified, the 3D-model(s) can be constructed. In order to construct a comprehensive 3D-model though, Marr argues that one needs a coordinate system. In a 2,5D image, the information contained in the image depends as much on the shape of the objects within as it does on the viewpoint of the observer. How the observer sees an object through angles and lighting determines what sides of an object are visible and how much is hidden from our view. Yet, when we perceive objects and recognise them, these unseen angles are not experienced as hidden or unknown in our understanding of how a recognised object is shaped. When I see an apple, I do not image just half an apple and then something unknown on its, so to say, "dark side" (the blind angle of which I cannot gain any optical data). No, instead I see an apple and recognise it as one of many other apples that I have seen before. I can rotate this apple in my mind and understand its shape without having to see what the apple looks like all over. Of course, this could result in misleading assumptions and errors in my assessment of the apple's shape (it could have a big bite in it that was obscured from my angle), but these would be particular features limited to unique situations. In the majority of cases, my assumption about an apple's shape would be correct. This would be, in a representation-based computational model, because I have a representation of a generic apple object in my mind through which all other recognised apples will primarily be understood, i.e. an "apple type" representation. Unique features of each apple (bites, deformities etc.) later shaping these into more specific representations, i.e. "apple token *n*" representations, where *n* identifies one particular apple in the world. More importantly for Marr's

theory: if we see an object at one point in time from one viewpoint, we need to be able to recognise the same object again at another time and from another angle. As such, we *need* to be able to construct a full representation of an object through only partial data.

Since we are able to understand the shapes of objects without needing to see the entire object from every angle, this would suggest that 3D-model representations are not viewpoint-based like 2,5D models. If 2,5D models are 50% defined by the viewpoint of the "scene" presented and 50% defined by the shapes of structures within, 3D model representations will have to be 100% defined by the shape of the represented structure. That is, 3D-model representations are entirely object-centric. If we are to represent the features of an object without making use of its coordinates in a context like a 2,5D image's "scene", we will need, as mentioned above, a coordinate system through which to internally map the features and shapes of these objects. Marr believes that mental 3D shape representations are described in volumetric rather than surface terms. What this means is that these shapes are described through centre of mass, size and axis of elongation. This assumes that complex objects like dogs and humans are segmented into smaller volumetric shapes which in turn are described not only in the terms above but also through their relative position to each other. This creates a relative coordinate system from which parts of an object can be identified through their location not in a viewer-centred context but an object-centred one. As such, we receive optical information about a shape that, through identification of finger shapes attached to hand shapes attached to arm shapes, we can correctly recognise and internally represent as a human arm. This hierarchical structure also allows for our visual recognition system to be modular, allowing arm-objects to be recognised on their own or collectively with human-objects or even say gorilla-objects. The surface shapes and characteristics can differ but the volumetric shapes remain within an acceptable relative variable through which we can categorise shapes into types.

4.8 – Summary of Marr's theory.

In summary, Marr's theory of vision involves the computational processing of visual data into three-dimensional representations of objects in our mind. The first stage of this process is a simple sketch of intensity changes in light and colour, creating a two-dimensional image of the perceived geometry. Next, this sketch is further enhanced with contours outlining surface shapes, building a viewer-centric 2,5D image bound by a coordinate frame. Finally, these contours are translated into volumetric information that, parsed through a modular object recognition system creates 3D-model representations of individual objects perceived in the 2,5D image. This kind of visual system would require a lot of information to be constantly gathered and processed for vision to function at the

level at which us humans experience it. This kind of system definitely works from a very input-output oriented structure where the mind gathers external input, computes it through algorithmic steps and produces as output mental content for later utilisation. Marr's theory is quite computation-heavy, as changes in the visual field means that each new image requires all steps to be repeated once again. If we interpret these images as retinal "frames" much like how a computer screen produces frames per second, this kind of system would seem quite resource intensive on the mind and brain. In the theory's favour: the last step in the computation of an image, the 3D-model representation, the volumetric approach allows for a minimal amount of elements to be computationally considered, while also utilising a modular object recognition system through which specific shapes can be grouped together or recognised individually, though stored in wider categories. Thanks to this a finger can be recognised as "a finger" and does not specifically have to be "a finger on a human hand on a human arm", since the same finger shape could just as well in other cases be recognised as "an ape's finger on an ape's hand on an ape's arm", or to step into the realm of imagination "an ape's finger on a panda's paw on a giraffe's leg". This kind of modularity means the same resources can be spent on a wider array of situations and computations, which means greater efficiency and thus frugality. Another show of frugality is that Marr specifically mentions *a priori* assumptions to be crucial for the translation of 2,5D images into 3D-model representations. These assumptions involve continuity, structural changes in visual motion, rigidity, shapes from contours etc. These assumptions are of a rather basic sort, close to what one could expect in a checker shadow illusion (Adelson, 1995, 2000) where local contrast alters the experience of lightness in shades, though not at all as specific in the knowledge dependence as seeing the yellow in grey bananas.[34] The fact that memory and knowledge can twist and shape our visual experiences is too apparent to be ignored, and we need more integration with these areas of the mind going into visual processing for a theory of vision to feel complete. In this regard, Marr's theory falls short in comparison to more recent theories that focus more on pro-active perceptual cognition, like enactivism and Predictive Processing, the latter of which we will get introduced to in the next chapter.

[34] Hansen et al. (2006) performed experiments where subjects got to interactively adjust the colours in a picture of a banana on grey background. Even when the image reached the point of an achromatic banana, subjects could still see yellow and wouldn't stop until reaching a slight blue hue, yellow's opposite. Hansen theorises that the knowledge that bananas are yellow thus caused an internal adjustment effect that needed to be cancelled out. With enough of an opposite colour added, this external adjustment cancelled out the internal one, resulting in a grey visual experience.

Chapter V – Predictive Minds

In the previous chapter, we explored Marr's theory of vision. While it is a thorough computational account for how information enters our eyes and results in a three-dimensional visual experience, it also suffers from being very input-hungry, requiring us to take in and process as much of the world as we possibly can in order to creaty a visually rich experience. Furthermore, Marr's theory also easily falls under the criticism of 'Cartesianism', with the idea that our minds are passively awaiting visual input, creating a worry that we in our visual experience are just locked in secluded movie theatre in our minds.[35] This chapter will involve diving into the potential replacement to Marr's theory; the Bayesian-based Predictive Processing, also sometimes called prediction error minimization. While Marr's theory represents a version of CTM under attack from enactivism – passive, input-hungry and detached from the world – Predictive Processing fosters the idea that we meet the world halfway, positing a frugal and active version of perception that at the same time relies on computational representation. Before we discuss the theory, I will take a moment to introduce why I think vision, out of all the senses, has become the topic of so many philosophical theories and discussions about perception.

5.0 – Vision, Old and New

Vision is often the focal point when discussing the nature of perception. Noë's enactivism favours examples relating to visually interacting with objects in the world, most prominently the idea of turning and moving about to determine the roundness of an apple (Noë, 2004). Why is vision the one of our six senses that is given the most focus in these discussions about how we come to know and interact with the world? I can see three main reasons:

The first reason is that the many nuances of vision are easier to discern, grasp and communicate linguistically. Visual perception, though it has a high level of complexity to it, is fairly accessible when it comes to description. There is a lot of language available to describe visual experiences and through this language alone in many cases we are able to recreate the same visual experiences. I may describe a red dress on a curly-haired woman, standing next to a stream during a warm summer day. If someone were to receive that description, that person can reliably picture something like that

[35] Comparable to Dennett's (1991) term 'the Cartesian theatre'.

visual scene in their mind.[36] When a person sees something, that experience is like a window out into the world where they can discern shapes, objects, movement and colours, and differentiate between tokens of all of these. By contrast, senses like hearing, smell and taste are fairly limited in what they can perceive as well as what language is available to describe these experiences. Gastronomes might find it easy to describe the taste of a dish in all its nuances, but the judgement of the average person would be far more simplified. Adding to that, there is less of the world that we can discern from taste alone. Tasting something simply discerns taste. A knowledgeable person may use taste combined with smell to discern he presence of harmful substances in food or become aware of nearby danger through the smell of a predator. However, you could not use these senses alone to navigate the world in any meaningful way, not on the level at which human taste and smell lies.[37]

Tactile sensation is a form of perception that perhaps comes close to vision in regard to linguistic availability and navigational application. There are a lot of ways in which we can describe how things feel, including heat/cold, coarseness/smoothness, size and shape. We can describe these features in comparison to other objects, and even when encountering a type of object that we haven't felt before, we can still often communicate an approximation of what it feels like touching the object. We can even use touch, and frequently do, to discern weight and relative location. Touch is perhaps one of the most vital senses for navigation and survival. A blind person can still operate and move about in the world through the use of tactile senses, in some cases making use of tools like guide sticks. On the other hand, a person who cannot feel will encounter great problems in interacting with the world. A limb that is numb can feel alien and difficult to control. In addition, numbness and unawareness of pain can lead to accidental self-harm. However vital the sense of touch is to survival though, its value to philosophers discussing perception still falls second to vision, and this is where my second reason comes in.

The second reason why I think vision is the prime source of thought experiments about the nature of perception for philosophers and cognitive scientists, is that it provides much more interesting discussion material and raises many more issues than the other senses. In vision one can discuss the experience of colours and the nature of visual illusions, for example. These kinds of discussions bleed into many big important topics of the mind, such as the nature of qualia and the ways in which, and to what extent, knowledge affects and shapes visual experiences. These issues directly focus on the relation between mind, perception and world. By finding out how the world is filtered by our mind,

[36] That is not to say my act of describing a photo of said scene would necessarily (or even likely) give rise to that very photo being pictured in the other person's mind, but their representational states would be similar and contain the same types of details.

[37] Snakes might have better luck in this regard. Of course, this is true of other animals with more refined senses than humans, while other senses may be less refined relative to humans. Since our subjective experiences are that of humans, this discussion is best left human-centric.

we learn more about the latter. To any philosopher of mind who wishes to gain insight into the nature of the mind, vision thus becomes a highly important focal point on many issues. I would have to point out that when it comes to illusions, tactile and auditory illusions are at least as interesting as visual illusions. Take for example the rubber hand illusion (Ehrsson et al., 2004) where synched tapping our (out of view) hand and a rubber hand in our visual field will cause us to feel as if the tapping sensation belonged in the false hand instead of our own, or the McGurk effect (McGurk, 1976) where visual experience of a mouth speaking affects how we perceive the sounds that we register as related to it.[38] However, also note how both of these illusions contain, in fact hinge upon, visual aspects. Additionally, the sense of smell is a very interesting topic in relation to memory, being a very powerful emotional trigger (see Herz et al., 2004). What makes vision stand out is that it features across all bases. Simply put, vision lends itself best to this task simply by providing us with the most material to dissect. However, there is one last reason that I think is the focus of discussions regarding the mind and perception, even the relations between mind and world in general. This reason applies to the academic landscape of my current writing, and strongly relates to the rise of enactivism.

New and old theories of mind alike revolve around features directly represented in visual theories. When we discuss the representations, we mention such capacities as picturing objects in our mind (like dogs or tables) or simulating action via mirror neurons by *observing* said action in other creatures. One of the focal points of this thesis is the attack on computationalism coming from enactivism. The classical form of CTM, along with the input-output sandwich, is seen by many as directly represented in David Marr's theory of vision. On the other side of the fence, recent success of theories on predictive brains (a lot of which is based on findings in predictive vision) and action-based robotics is seen as a big victory for enactivism, which core tenets revolve around enacting the world, a lot of the time described in terms of action facilitating *vision* (like saccadic eye movements). As such, both CTM and enactivism rely on or are identified with theories and statements regarding vision. This in turn naturally creates an academic landscape where vision continuously ends up the focal point, or at least the origin from which the discussion takes flight.

However, I contest the idea that Marr's theory represents all that computationalism can or even should be, while also arguing that predictive brain theory is actually more compatible with computational theory of mind than radical enactivism. The way I will go about this is arguing for the idea that a brain with a predictive framework has the capability of performing Bayesian inferences, which in turn allows me to argue that such cognition, these Bayesian processes, are - perhaps

[38] Both of these will be brought up again in chapter VI.

necessarily - computational in nature. There are a number of issues that will have to be addressed along the way. First of all, does a computational brain automatically lead to a computational mind? As stated before, while I believe the two to be very closely related, I have to treat the two as distinct entities as one incorporates much more than the other and both possess features that fail to overlap with the other. Secondly, does a computational brain harm enactivism in any way? At first glance, I highly doubt it would. My pursuit of proving the computationalism of predictive brain theory is less a stab at enactivism and more of a redemption of the computational theory of mind. Thirdly, just how far can we go with the Bayesian model and, even more importantly, the predictive one? Some philosophers put great emphasis on cognition being 'active' in the sense that we seek out input instead of waiting for input to come to us. This correlates with predictive theories, where input is first predicted, and the actual input that follows is either verifying or corrective. Andy Clark is one such philosopher, wanting to put much more emphasis on the predictive and action-oriented nature of the mind as its key feature, if not as an all-encompassing aspect of what we do.[39] I'm personally sceptical of the extent to which prediction could be said to be the key feature of cognition, more so for the mind than the brain.

5.1 – Predictive Vision

So what is the current landscape when it comes to theories of visual systems? Well, it involves a lot less input compared to that offered by Marr's theory, and the gaps in the data resulting from this are instead filled in by our own mind, via prior knowledge and, perhaps surprisingly, guesswork. It would probably be more accurate to say that according to these new predictive theories vision doesn't so much involve less input as it *requires* less. The same amount of light is still hitting the retina and the same amount of information is still being passed on from the eye to the brain - what *does* differ however is that this flow of information is quickly intercepted by an equal flow of predictions. One of the proponents for this predictive theory of mind is Andy Clark, and I have made reference earlier to his (2013) paper expressing just those kinds of sentiments, though the idea of perception as something involving knowledge-driven probabilistic processing goes all the way back to Helmholtz (1860/1962). This idea has since lingered in the study of the brain and perception but has now stepped up on the main stage of Philosophy of Mind, enticing philosophers such as Clark (2013,

[39] During a conference I attended in Edinburgh in 2014 this topic came up, and Clark seemed quite happy to concede that everything the mind does is prediction in some form or another. This includes calculation, deliberation and memory. These may seem unintuitive candidates for prediction, and later on I will indeed argue against them being predictive in nature.

2016), Hohwy (2007, 2013, 2014) and Noë (2004), as well as cognitive scientists like Hinton (2007a, b), Friston (2005, 2008), and Knill & Pouget (2004, 2011 [with Moreno-Bote]).

A very fleshed-out version of a predictive vision framework can be found in Clark (2016), where he goes through in detail just what is meant by his suggestion that the brain is foremost a prediction machine and how this links into action and ultimately how it can support the idea of the human mind as embodied. This of course links back to the second chapter where I discussed the nature of the four 'E's in the EEEE family of cognition theories, of which Embodied cognition was one. I shall now continue by presenting a general view on the predictive mind in a summary that I hope does it justice. Note that this view is not exclusively held or created by Andy Clark and thus I will henceforth refer to it under the more general term Predictive Processing. A more descriptive term would also be to call it a prediction error minimization framework. The reason for this is that, as will become apparent below, at its core Predictive Processing is all about minimizing the amount of error signals generated by the system. For now a general view will do, but it will become much more important later to distinguish between Clark's view and that of others, as there are different interpretations and variations in approaches that carry certain important distinctions when it comes to subjects like internal representation, which will become a topic in Chapter VI.

Predictive Processing's take on perception is one where we are constantly 'guessing' input or expecting the world before we actually perceive it. Constantly when walking through the world we expect things to happen, we guess what is going to happen next based on previous experiential knowledge. When I press the buttons on the vending machine, I 'expect' the chocolate that is tumbling down. First, I predict what will happen when I push the button, I expect to see a chocolate bar in wrapping tumble down before me. After that, I predict the thud of it landing. When I reach in, I predict the feel of the plastic wrapper on my fingertips and finally I predict the taste of the chocolate bar when I eat it (at varying detail depending on if I've had this particular type of chocolate bar before, though the general idea of a chocolate bar is already a sufficiently narrow selection of predictions in the wide array of possible things to taste for me to have a pretty good idea ready in my mind of how the chocolate will taste). Unless something unexpected happens, all of these predictions will be affirmed by what happens next, as I've come to trust through experience and the general reliability of things in the known universe. So in Predictive Processing, how does this sequence of predictions affect the soon oncoming series of signals impinging upon my sensory organs once the button on the vending machine is pressed? The answer is quite simple: they will measure up against each other; the incoming signals either affirming or disproving the prediction cascade. The short of it is that when we predict sensory input, these predictions intercept the actual sensory input coming in from the world outside. What happens next is that the correct predictions will simply cancel each

other out. The prediction was correct and thus there's nothing to pass on since my mind has already predicted this to be the case and, once affirmed, has no need of corrective data in order to generate data of what is going on. In contrast, it is the 'prediction errors' that cause interest from the system. These errors are sent upstairs for a closer inspection, so to speak. Imagine envisioning a famous painting in one's mind. With enough effort, we can form a picture of it in our mind's eye, an image generated by our memory mixed with assumptions we may have about lighting, angles or brush strokes. Now imagine that we open our eyes and see that very painting before us. The things that our vision will primarily pick up are the unexpected elements, those we didn't envision, didn't remember or simply were flat-out wrong about. There would of course be no difference in the light particles hitting our retina or the initial signals going from the eye through the optic nerve into the brain. Eventually however, the bottom-up information in these signals will be intercepted by top-down predictions, ensuring frugality in what is being processed (the information here simply being the sensory response to whatever physical energies that impacted upon them, which is then to be translated into, in the case of vision, visual data of colours, brightness, shapes etc.). The result is that 'error signals', i.e. signals that do not match our predictions, will be allowed to pass and reach higher levels where these same errors will serve as corrections to our initial prediction. Once our prediction incorporates the elements that it got wrong, we then have a correct version of what is being seen out there in the world, with only part of the input actually processed. We have thus used top-down connections alongside bottom-up error signals in order to generate a vivid visual experience: a representation of a painting in our mind (Hinton, 2007a). Thus we've spent less processing power (or energy currency) compared to a more classic approach like Marr's, but arriving at the same destination.

An important aspect of this predictive model is that it is hierarchical. Much like how Marr's model is made up of steps from primal sketch to 3D-model, each step leading to a more complex mental structure being presented, or how the features of 3D-model images are grouped in hierarchical structures where the parts can be recognised separately from the whole, so is Predictive Processing's model hierarchical in the structure of its predictions. This hierarchical structure forms a spectrum of prediction levels where the lower levels focus on predictions regarding the most basic spatial and temporally specific features of incoming information, while the higher levels are increasingly abstract the further up they go. For example, going back to the example of the vending machine: The expectation of a 'Mars bar that I paid for' or similar would be an example of a fairly high level expectation. It involves the concept of 'Mars bars' and the expectations of getting what one pays for, which also involves the previous knowledge that I had put coins into the machine just moments ago. An even higher-level prediction that we could use as an illustrative example would be to walk out the

door in the morning and expect ‘a lovely day’. This would involve a myriad of assumptions and abstractions. Even so, any events that would occur to contradict these expectations would stick out as errors that would have to correct my idea of what kind of day I was having. If nothing happened or was perceived in such a way as to contradict my expectations, then I would indeed, at the very least in my mind, have had a lovely day. In either example, these high level predictions are not very spatially or temporally precise. I predict a Mars bar to fall out the vending machine, but I do not know precisely when. I do have a rough idea of course: I do not expect it *tomorrow*, having the knowledge that I’m pressing the button to confirm my selection in mere moments, and that vending machines do not take hours or days to deliver the goods. Equally, the prediction of a lovely day is specific to that day, but not precise to a moment in time where the lovely day will be had. In addition to this, I do not have a prediction of *where* the lovely day will be had. Equally the expectation of Mars bars from vending machines, while I do have an idea for a place (being the vending machine) this precision of spatial prediction is reduced the further I am removed from the event itself, and is not a necessary part of my expectation. To clarify: I could have this expectation before entering the room where the vending machine is and find that I had misremembered its location or that it had been moved. I could find it out of order and simply opt for a machine in a different location. These kinds of errors have much less impact on higher-level predictions. As we go down the hierarchy, the predictions become increasingly fine-grained, thus this type of temporal and spatial precision becomes much more crucial for our predictions to match the constant flow of information.
Whenever an error is detected at the lowest level, the error signal is passed on to the layer above it, functioning as input for that layer. Thus error signals create a cascading flow of input, with an equally cascading flow of predictions coming from the top. What follows is a very frugal system that actively rather than passively perceives the world, continuously updating its understanding of the input and improving upon subsequent predictions of what is going on. The distinction between active and passive here lies in how action relates to input. For example, Marr’s system of vision *reacts* to incoming input, but the input is always first to arrive in the order of processes. By contrast, Predictive Processing prepares for the input before it arrives. As such, the relevant processes surrounding the single instance of perceiving a tree start before the actual visual information of the tree arrives on the retina. The flow of these processes is also dynamically regulated, in that more uncertain situations result in more input than predictions (where error signals would be passed on unchallenged) and on the flip-side other situations could be to varying degrees lacking in input, causing our predictions to step in and help us fill in the gaps. However, in the case of the latter I would like to highlight that unless these gaps were miniscule or simply full of perceptual ‘noise’, we would still be dependent on actual input for our experiences to appear experientially genuine. While predictions help with frugality, they require to be met with confirmations in order to actually become

experiences. Otherwise, the lack of input would be cause for error signals, or simply a lapse in perceptual activity. Even though we are able to imagine things in our mind (sometimes very vividly), these imaginations do not become illusory experiences if left unchallenged.

Imagine seeing the same video clip over and over until you can replay it perfectly in your mind, or a person who is listening to a song she's heard many times before, and singing along with the lyrics. In the latter case, that person would almost be listening to the song independently of the song itself (maybe even to the point of an involuntary earworm). Indeed, if a song had a sudden (deviating and unexpected) second of silence in a random location, a person familiar with the song who was humming along would still be able to 'hear' the missing notes in their mind in spite of the newly introduced gap. However, the person would still be aware of the gap (predictions would be corrected and an error signal passed on). Our mind would not deceive us to believe that there was no actual gap. An example of such a deception would be what is commonly known as the temporal induction of speech, an auditory illusion where "replacing a phoneme in a recorded sentence with a cough resulted in illusory perception of the missing speech sound." (Warren, 1970, p.392) In this experiment, even with the cough completely replacing the phoneme, the sentence would audibly appear uninterrupted (the cough would still be heard of course). This is a case where noise is filtered and separated from the subject of our focus, and prediction fills in the gap. It would be important to note here that replacing the cough with a silent gap instead made the absence of the phoneme much more apparent, thus breaking the illusion. In such a case from the framework of Predictive Perception, the gap was clearly obvious enough for it to appear as an error signal rather than something that should be patched. Moving back to my example of the manipulated piece of music, even though we can keep 'listening' to the tune in our mind without the need for the tune itself (even if the song is cut in half you can 'finish' the other half of the song in your mind if you are so inclined), we are not tricked by our predictions into hearing the music in the sudden gap. Our predictions have not made us blind to the gap's existence. What we 'hear' in our mind when we imagine a song is experientially different from the kinds of experiences our confirmed or corrected perceptual predictions bring us.

5.2 – Bayesian Adherence of Predictive Processing

According to people like Hohwy (2013), Friston (2008), and Knill & Pouget (2004, 2011 [with Moreno-Bote]) these predictive perceptual processes could very well adhere to Bayesian principles of probability. In fact, it turns out that Bayesian processes would help a great deal when coping with non-linear interacting causes. What this means at the outset is that neural computation

accommodates for uncertainty. Predictive Processing in perception is largely about predicting *causes* that explain input. We are in this view encapsulated in our skull, with our perceptual organs and perceptual processing being our only tools to receive, translate or understand information about the outside world. When we get information about a hat shape and need to figure out the cause of this, we employ predictions that cascade from high-level to low-level hypotheses looking for confirmation in the incoming data, with residual error signals being what comes back up for re-evaluation; a likely contender in this case is that there must be a hat in front of us. This may sound like an obvious statement, but the prediction of hats or other singular objects appearing in our field of vision is but the most simplistic example. Examples of more complex causes, and more importantly non-linear ones, can "range from simple perceptual interactions like a cat being partially occluded by the fence, to deep and complex regularities such as the impact of the global financial crisis on the effort to combat climate change" (Hohwy, 2013, p.59).[40] These kinds of complex causes are likely to have not a singular possible explanation but several. A hot-air balloon that takes up an increasingly larger area in my field of vision could either be approaching or, more unlikely, rapidly increasing in size. A cat shape interrupted by the pickets in a fence could either be a whole cat standing *behind* the fence or, morbidly and once again less likely, slices of cat placed in the gaps in the fence. Even though all of the explanations found in these examples are *possible* causes, it is clear that some are more *probable* than others. We do not usually make mistakes with these. However, there are other situations where mistakes are more likely to be made. Imagine for example that you are walking down the street and a stranger approaches, walking in the opposite direction. The person suddenly looks at you and starts to smile and wave. There may be a moment of surprise here. You do not know this person, so what is the cause of what is going on? We have knowledge that we apply in our predictions that people who smile and wave want to make contact and most often know the recipient of this treatment. Is this stranger simply a person we know but have forgotten about? Is it a case of not recognising a face? The assessments we make here shape the actions that follow. Perhaps we choose to wave back and suddenly someone else comes up from behind. The two strangers greet each other, our mistake has been revealed and perhaps a feeling of embarrassment creeps in. This is a case where, thanks to the complexity of variables and the small difference in probabilities, the context generates a lot of uncertainty. In all these cases, higher level predictions function as priors for lower level activity, that is they express the initial probability distribution for an event before incoming error signals correct this probability. Low-level predictions are more basic and concerned with fine details of incoming information while higher level predictions focus on the bigger picture and are very sensitive to

[40] One may note here an example of how Hohwy signals prediction as potentially going further than just perception.

context (something lower-level activity generally is not). As such, low-level activity can be described as myopic while high-level activity is hyperopic (Hohwy, 2013).

So we have now stated that predictions involve probability. But there are two more points to make: Firstly the internal representations generated by our predictions are not fixed values but probabilistic in and of themselves, i.e. they are stochastic. When perceiving a 'depth' feature in a natural scene or an object, we do so by representing the depth as a relative probability, not as a singular precise value (Knill and Pouget, 2004, p.712). Secondly, predictions can be laterally weighted. This is a feature of having competing hypotheses on the same hierarchical level. What it means is that over time among a group of lateral competing hypotheses, one or more particular hypotheses may have an apparently higher posterior probability than the rest. In this case, the less probable hypotheses will be reduced in priority and "silenced" while well-performing ones may be enhanced. Similarly, prediction errors themselves evolve in a hierarchical manner over time. A unit that frequently puts out prediction error signals that continually get explained away by the same higher-level hypothesis will form a relationship of reliability, giving these error signals greater weight in the "message-passing economy." (Hohwy, 2013, p.61) The system is thus not only taking in prediction errors level-by-level, but also evaluating those error signals and estimating the 'trustworthiness' of different sensory data on each level. Those deemed untrustworthy can then be judged to be mere 'noise' while the trustworthy signals will gain extra attention. This is called Precision-Weighting (Clark, 2016).

What does it mean then to claim that something is 'Bayesian' in nature? To give the answer to this, I will give a brief explanation of Bayes' rule. What it does, in essence, is it expresses a posterior probability arrived at by assessing a prior probability conditioned by a certain event or condition that changes said probability. What it looks like in equation form is displayed below:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

In this equation, the probability of A given event B is equal to the product of the probability of B given A times the probability of A divided by the probability of B. In more eloquent words, the posterior probability of A equals the prior probability of A times the likelihood of B. Prior here means a probability on its own before having been conditioned or changed by an experiment, event, newfound knowledge or evidence. It is as such untouched by new information. Posterior denotes that the probability *has* been altered by this new event or information implied in the relation

between the two terms. As you may have noticed I've mentioned posterior probability just above, where it related to the predictions altered by the new error signals flowing up from the bottom layers and thus when 'corrected' gained an altered probability value. It was discovered by Thomas Bayes in the 18th century and after spending about two centuries in a state of relative controversy, it has in modern day become a very popular theory and a tool that is currently applied to a wide variety of fields ranging from genetics to artificial intelligence to forensic science to, as we are looking at now, visual perception (Stone, 2013). To illustrate how this equation works, I will go through an example I've devised, step by step:

Suppose that I have started showing symptoms of a disease. I'm experiencing a high fever, profuse sweating and nausea. I go to the doctor and she tells me that 90% of people suffering from malaria also suffer from my particular symptoms (perhaps eager to spot an exotic disease). This piece of information can be quite worrying indeed. However, what this tells me is only really the probability of me having the symptoms that I already have, given that I have malaria. This information is not so useful given that I'm currently in the UK where malaria is hardly rampant. For all I know, I could merely have the flu. If we let A be "infected with malaria" and B be "having symptoms of fever, sweating and nausea", we can calculate the likelihood of me having malaria given the symptoms, i.e. P(A|B), by using Bayes' rule. P(B|A) or "the probability of having the symptoms 'B' given that you are infected by malaria" is 90% or 0.9. Let's assume that the risk of contracting malaria in the UK is 0.005% or 0.00005. Now, it also happens to be flu season, and currently 1 in 10 citizens are showing symptoms similar to mine. We now have all the values we need and there's but a simple task of calculating the probability of me having malaria. This calculation goes as follows: P(A|B) = (0.9 x 0.00005) / 0.1 = 0.000045 / 0.1 = 0.00045. As we can see, there is a 0,045% chance that I'm actually infected with malaria, making flu the much more likely scenario. The probability of me having malaria is reduced by the fact that there are many others around me experiencing the same symptoms, with high statistical likelihood of these symptoms having other causes thus allowing for alternate explanations to the cause of my predicament.

So what does this mean for Predictive Processing? It means that by adhering to Bayesian principles, predictive systems can 1) adjust a standing hypothesis (a prior probability) when approached by a new event or set of data (error signals) to create an updated hypothesis (a posterior probability) from which we make increasingly informed predictions that explain, or rather "explain away" said errors. It also allows such a system to 2) decide how much attention to give toward incoming error signals depending on how probable such signals are given the hypothesis (Clark, 2016). The system is thus in the business of minimizing the amount of errors in our predictions and can be referred to as a *prediction error minimization mechanism* (Hohwy, 2013), thus the name we

gave it earlier. If the hypothesis is that there is beer in the fridge, opening the fridge and receiving no beer signals is a very low probability given the hypothesis. This would generate a large error signal; as such, we may need to search the fridge with our eyes to see if we haven't missed it somehow. If in the end we confirm that there indeed is no beer in the fridge, the hypothesis that there *is* beer in the fridge needs to be heavily revised. What this does for Predictive Processing is allow a system to *learn* over time. Indeed, one of the biggest successes of the Predictive Processing theory is its capacity for learning. As our brains predict the world through guesswork, and as the world outside of us is "a world that is highly structured, displaying regularity and pattern at many spatial and temporal scales, and populated by a wide variety of interacting and complexly nested distal causes" (Clark, 2016, p.19); this in combination means that the guessing gets sharper and sharper over time. This kind of learning is multilayered (as a result of the many-layered hierarchical structure proposed by the Predictive Processing theory) and has an objective to train top-down connections to generate increasingly accurate representations of the world. To clarify, bottom-up signals are signals that rise within the hierarchy, combining at each step to form more complex signals ("red" and "cup" on one level move up to become "red cup"), while top-down signals start at the top and descend, being taken apart into lesser and separately identifiable components. In this way, bottom-up is assembly and top-down is reverse engineering. Marr's theory of vision, while relying on a priori knowledge to filter and make sense of incoming data, can be described as a wholly bottom-up system in that it assembles the pieces of information into larger and more complex structures, until finally arriving at a 3D-model. Predictive Processing, on the other hand, is a bit of a strange hybrid, where hypotheses are broken down in a top-down fashion, while incoming signals are assembled via bottom-up processes.

Predictive Processing is an improvement over earlier theorized models for artificial networks that instead were meant to "adjust the weights on the top-down connections to maximize the probability that the network would generate training data." (Hinton, 2007a, p.428) The 'training data' brought up in this citation refers to what is usually used as innate knowledge for back-propagation learning procedures, a set of data which the top-down connections were trained to 'classify' the bottom-up signals by. This procedure thus relied upon already knowing about the world so as to identify incoming data according to the framework. However, such models are supervisory in nature and require an innate set of knowledge that dictates the desired output. In that way, they didn't guess the world like the predictive model does, instead they expected it.[41] What is meant with this

[41] As PEM is a self-teaching model of mind, one could argue that at a sufficiently learned and advanced level, such a system is also "expecting" rather than "guessing". In fact, that is how a PEM mind would build up a comfortable certainty about the world. The main point of difference however, is the way in which the system got there from the outset.

distinction is that guessing is speculative and more easily convinced of error, while expectation relies on a certain degree of certainty. Compare, for example my guess that there is a cup on the desk and an expectation that this is the case. With a guess, I'm more prone to want to affirm this belief before acting upon it, and would not be remarkably surprised if it turned out my guess was wrong, I may in fact be neutral to this outcome. When expecting a cup, I may reach for it without looking, and find myself very confused when reaching for nothing but air. As such, expectations are part of our world model. What is suggested by scholars like Clark and Hinton is thus that a mind that follows a top-down predictive structure rather than a back-propagating one can do away with innateness altogether. This is a big step forward compared to more classical 'Fodorian' theories, in which Marr's theory of vision is included, that required innateness on some level or another.

5.3 – A comparison to System 1

Looking back at previous chapters I've been stressing frugality as an attractive aspect of any theory or model of cognition, as less resource-heavy systems, or systems that work toward reducing resources required for them to function, are more plausible from an evolutionary standpoint. This goes all the way back to my mention of the reasoning by Penrose that our prehistoric ancestors would have greater benefit for practical and adaptive thinking than being good at mathematical equations (Penrose, 1994, p.148). The general argument is that an advanced but 'clumsy' cognitive system would be an unlikely contender for how the mind actually works in virtue of the sheer unlikelihood of it having evolved in the first place. An emergent, evolving system that doesn't provide direct benefits for survival without spending a disadvantageous amount of resources is unlikely to have made it in the first place. A fast and frugal system that teaches itself has a greater evolutionary probability for two reasons: Firstly, it requires less resources, or in cases where great resources are required in either case; economises these resources as well as possible through adaptation. Secondly, a self-teaching system is a better model and explanation for how a basic system – more so in a historical perspective of the human species, but also from the much more focused perspective of a single human life from birth to death - survives long enough to transform into an 'advanced' system like those presented in this chapter, both Marr's and Predictive Processing. What I mean by advanced is that the system possesses features that are unlikely to have been there at the outset and most likely evolved or were acquired over time. To look back at Penrose's example of mathematics, it makes sense that this capability of mathematical thinking was not a high priority in the more dangerous times of early hunter-gatherer societies. Still, we do today have such capacities. Equally, Marr's theory of vision presents a system requiring an internal setup with the innate capacities to accurately

translate the information processed from input to output. Such a system, or any computational system for that matter, does not come about randomly through a bolt of lightning in a primordial swamp, but requires time to develop and be refined. As such, a system needs to be functional in interacting with and processing information about the world outside of our minds even in its early evolutionary stages. A self-teaching system, once the basic setup is achieved, has a much shorter road to journey before achieving present-day capabilities compared to a more top-down system relying on (extensively) innate knowledge or complex internal structure. This is not to say the latter are evolutionarily impossible, only that they are less attractive from such a perspective.

While Predictive Processing is largely of a very different structure from Marr's theory, in some ways what prediction error minimization is trying to accomplish for judgement-making about the world is very much comparable to the works of Gigerenzer and Kahneman, both of whose ideas we've been exploring in previous chapters. Are these ideas comparable to the point where one may suggest a compatibility of ideas could be attained or are the similarities merely superficial overlapping of themes and approaches? This is the question I will seek the answer to in the following section. The obvious angle I'm taking here, attempting to strengthen the position of a Compatibilistic Computational Model, is to explore the possibility of describing predictive visual processing as modular, since - as I've pointed out in previous chapters - both Kahneman's System 1 and Gigerenzer's toolbox can be described as being of such nature. In the case of Gigerenzer the key theme is frugality; just like the adaptive toolbox (Gigerenzer, 1999 & 2008), Predictive Processing seeks to trim the fat in terms of how much deliberation and resources are spent on arriving at the answer to a question or problem. In the case presented in predictive vision, the question is "what do I see?" and the answer is given by hypotheses explaining away error signals.

If one were to translate Kahneman's System 1 and System 2 model of cognition, Predictive Processing as presented here (as a model for the visual system) is very much a System 1 activity. Indeed, this is a sensible initial position to take as Kahneman himself explicitly states that perceiving the world is a System 1 capacity, along with all the other innate capabilities we share with other animals (Kahneman, 2011, p.23). I would like to note here that although we differentiate between System 1 and System 2, the former is in fact a collection of several sub-systems working autonomously; a grouping of all of our instinctive systems (Evans, 2003). By applying the model presented in Predictive Processing to our common sense judgement of our subjective experience as cognitive agents, it is fairly straightforward to assert that we are not actively or consciously aware of the back and forth of hypotheses and error signals, especially not on the lower granular levels of the hierarchy. However, does such an acknowledgement mean that we can take for granted the entirety of the predictive model of vision to be wholesale compatible with Kahneman's System 1? Hardly,

though the favourable comparison can be stretched a bit further before discrepancies are encountered. System 1 is fast, effortless, automatic and domain specific, generating gut reactions which we may accept as-is or further analyse, all of which could further be applied to predictive visual processing as described by Clark, Hinton, Hohwy et al. The processing of visual data from light rays to the experience of an 'image' is after all a largely (if not completely) automatic process that we rarely if ever experience as involving any real cognitive effort, nor do we experience any noticeable "lag" between visual input (light hitting retina) and our visual experience of the world. When I visually perceive the world, I see what is going on *in the moment* and unless I am actively searching for something, like a specific object or shape in a clutter/pattern (i.e. a visual 'chaos'), there is little conscious effort involved. So far so good, but the issue of actively searching for something is already a red flag hinting that a Predictive Processing model, or any concrete cognitive model of vision for that matter, may not fit so cleanly into a System 1 classification as one may initially intuit, despite Kahneman's initial claim. In fact, anything that requires attention from the cognitive agent is by definition a System 2 process. If one were to define the act of looking for an object within one's visual field purely in the terms of the mental cognitive components (i.e. disregarding the bodily motor actions of tilting one's head, shifting the eyes and physically move around, opening drawers etc.[42]) the question that we would have to ask ourselves is whether the active search for specific visual data (like trying to find Waldo) is something that takes place in the visual system itself (i.e. the attentive action of searching for an object in a drawer can be reduced to the cascading of hypotheses and error signals) or if this kind of cognitive process is something that merely uses the visual data produced by a *subconscious* visual system while the actual mental act of 'searching' this data for a specific shape or object takes place in a more consciously accessible central system. As mentioned before, think of access consciousness (Block, 1995). The former option would be incompatible with the idea of a predictive visual system as entirely subconscious and thus entirely System 1 compatible because of the problem of attention, while the latter would circumvent such worries as attentive cognitive actions would transition into System 2 territory without compromising the visual system's status as enclosed, specialized and independent of our conscious thought, much like a module. However, I will leave this for later as even greater obstacles very soon present themselves in this comparison.

[42] I will disregard these aspects of the action for now, but they will eventually be reintroduced into the equation as they will be highly relevant at a later stage. For now, let's just focus on the mental aspects and embrace the input-output sandwich.

5.3.1 – Experience-based Learning

Another comparative point that can be made about a System 1 visual system and Predictive Processing's visual system based on prediction error minimization is the way in which we learn from experience. As I've mentioned in the previous chapter, System 1 is not completely closed off to new knowledge and *does* learn over time. In fact, the judgements that we make through System 1 are heavily belief-based (which in turn gives rise to the belief-bias effect, more on this below) (Evans, 2003). In *Thinking Fast and Slow* (2011) Kahneman early on makes the point that System 1 deals in generating *suggestions* to System 2 in the form of impressions, intuitions and feelings about the world and the situations we find ourselves in. If System 2 *endorses* these impressions as true rather than turning them away by inhibiting them, then these impressions become what Kahneman would refer to as beliefs.[43] However, these beliefs are not ruled by System 1. After all, System 2 has to *endorse* them. Take an optical illusion, for example the Müller-Lyer illusion consisting of two lines of equal length. At the ends of these lines are diagonal lines creating arrows pointing outward at the ends of one line, and inwards at the ends of another. The line with the arrows pointing inwards will be experienced as longer, even though the two lines are actually, as mentioned before, of the exact same length. We (the System 2 "we") are bombarded by impressions from System 1, telling us to endorse these intuitions as a belief. In a situation without further context, we may well do so. However, by gaining the knowledge that these lines are in fact of equal length, by someone we trust telling us, or we ourselves measuring the lines with a ruler, we can come to the much more informed belief that our visual impression is false. In this case, System 2 can reject the impressions from System 1 and form a more abstract belief about the length of the two lines, based not on visual intuition but informed judgement.

So what is the belief-bias effect and how does it come about? Well, the problem arises from the fact that - as Kahneman frequently points out - System 2 is in many cases lazy. This is not true for all; some people are wired to prefer System 1, others to prefer System 2 when engaging problems and puzzles. When talking about students failing a logic experiment by relying on and being tricked by System 1 bias, Kahneman has this to say: "'Lazy' is a harsh judgment about the self-monitoring of these young people and their System 2, but it does not seem to be unfair. Those who avoid the sin of intellectual sloth could be called "engaged." They are more alert, more intellectually active, less willing to be satisfied with superficially attractive answers, more sceptical about their intuitions." (Kahneman, 2011, p.48) The implication here is that more rational and mentally active individuals are

[43] As with many terminologies in philosophy, the usage and meaning of terms shift between authors and eras. In this case, I will do my best to make sure the term "belief" is used consistently.

more likely to engage in System 2 reasoning, while less engaged individuals take the easy road of following gut-feelings and intuitions since they don't require cognitive effort. The range of beliefs that we hold in System 2 is ultimately what drives the interaction between System 1 and 2, which in turn *decides* what impressions are going to be endorsed as future beliefs. Without proper engagement from System 2, the system will simply go with what has worked before, or what seems familiar[44], i.e. the intuitions based on knowledge gained from past experience, which is the domain of System 1. Since we as thinkers usually take the easy cognitive path unless we're focused and on the alert, there is no reason to question impressions or hesitate to endorse them unless they cause some sort of surprise, i.e. a conflict in believability. This is perfectly illustrated in experiments involving the belief-bias effect (Evans, 2003, 2004 [Morley et al.], also initially in Evans et al., 1983).

The belief-bias effect is when the conclusion of a syllogism is deemed as valid whenever the conclusion, on its own, seems plausible and/or in accordance with our beliefs, even if said conclusion does not follow from the premises. It often comes in the form of us bringing in beliefs from outside the context of the syllogism, beliefs through which we can get immediate impressions about the *truth* value of the statement by looking at the conclusion alone, and maybe one of the premises. What we are *not* looking at in such a situation (or rather actively check, which would induce cognitive strain) is whether or not the two premises actually lead to the stated conclusion. What's even more interesting is that when the reverse happens and the conclusion does *not* seem plausible it is on similar grounds rejected for not fitting with our intuitions, even though the conclusion is perfectly valid following the premises. In 1983, Evans, Barston and Pollard created a methodology to highlight this belief-bias effect in a series of logic tests in which subjects were presented with a series of syllogisms, their task being to evaluate whether or not these syllogisms were logically valid or invalid. The tests were purposefully constructed to create the kind of conflict between believability and logic present in belief-bias. As such, there were four kinds of questions: Some questions had *believable* sounding conclusions, as well as a valid logical structure. Others had *unbelievable* conclusions and invalid logical structure (i.e. the conclusion did not follow logically from the premises). Then there were the cross-examples where the statement was *believable* but logically *invalid*, and *unbelievable* but logically *valid*. Note that in the cases of invalid logical structure, there were not obviously so to the extent of ridiculousness. Rather, they were of the kind "No A's are X; Some B's are X; Therefore, some A's are not B's." In context, the A's and B's were things related by context, creating a

[44] Even familiarity through repetition is enough for us to be put at cognitive ease when accepting the truth of statements. Kahneman takes the example of subjects being repeatedly presented with the incomplete sentence "the body temperature of a chicken" being more willing to endorse any statement regarding the body temperature of a chicken (Kahneman, 2011, p.64).

somewhat “disguised” logical flaw that requires attentive cognitive action and isn’t immediately caught - the purpose being to not make things too easy for System 2. An example could be this (taken from Evans, 2003):

No nutritional things are inexpensive
Some vitamin tablets are inexpensive
Therefore, some vitamin tablets are not nutritional

This is an example of a logically valid argument (the conclusion follows from the stated premises) where the conclusion is designed to be *unbelievable*, giving subjects a first impression (System 1) that it is not correct. The results of these tests were that subjects were significantly more likely to accept arguments with believable conclusions over unbelievable ones, showing a clear trend of belief-bias. Even in logically valid arguments, unbelievable conclusions were frequently rejected - results showing a 47% rejection rate in one experiment and a 38% mean rejection rate in another (Evans et al., 1983, p.300-301). Needless to say, the rejection rate was much higher in the cases involving both invalid *and* unbelievable syllogisms. In contrast, acceptance rate was the higher in scenarios with believable conclusions across the board, with a lower rate for invalid arguments. In general, the results speak for the existence of a belief-bias in reading the believability of conclusions. However, it is important to note that Evans (et al.) also noticed a significant increase in acceptance of valid arguments over invalid ones, pointing out that even though belief-bias is present, attentive individuals are able to overcome it (Evans et al., 1983).

Now I will return to our comparison. In a similar manner to what’s described above, visual prediction error minimization (Predictive Processing’s theory of vision, which I deem a good comparison to System 1, especially for the easily imagined visual aspects) is also a belief-based process. After all, it is based on hypotheses and predictions. In order for a person or system to predict something about the world, there has to be a hypothesis – or a belief – positing what is *expected* to be the case about the world, i.e. there has to be beliefs about the world present. One could argue that on the more finely grained levels in the hierarchy, where our hypotheses are no longer accessible to our conscious thought (only the higher level hypotheses are generally deemed to be something that we have conscious access to – this will be further explained in section 8 and 9), any one hypothesis at that level would not qualify for our definition of a held belief. What I mean with this is that we would be generally happy to say that when I expect to see a cup of coffee on my desk when I enter my room, it is because I have a belief that there is a cup of coffee on the desk

within my room. The hypothesis that there will be a cup of coffee on the desk in my room then cascades downwards in the hierarchy, becoming expectations of cups, desks and further down the geometric shapes that constitute these. Ultimately I may expect an ear-shape attached to a cup-shape. I'm not consciously aware of the process within which this hypothesis takes place, but do I believe it? When I rotate an apple to expect to see a bite on the other side, do I have a belief that I will see a jagged edge where red will go into white with patches of light brown? These seem far removed from what people usually consciously believe about the world, but I would argue that these cases still very much highlight beliefs. After all, the hypothesis of an ear-shape attached to a cup-shape is a product of my belief of how a coffee cup looks. My prediction of the colours and shapes in the visual stimuli of a bite in an apple is based on my attained concepts and beliefs about apples I've previously taken a bite of. Even though I do not actively hold these beliefs in my conscious thoughts, they follow from my knowledge about the world, and all relevant parts are baked into my hypothesis of seeing a cup of coffee on the desk. Expecting a 'cup of coffee' and knowing what a cup of coffee is (given that the term has meaning to me) thus involves a lot more that is employed and packed into that expectation, as is evident as the hypotheses cascade and become more granular.

So yes, this confirms what I said at the outset; that much like System 1, visual prediction error minimization is a belief-based system. However, I'm also here to compare how both systems learn. Much like System 1, Predictive Processing systems *learn* over time. In fact I believe I've already made the point previously that one of the best features of Predictive Processing is that prediction error minimization allows for bottom-up learning about the world that constantly updates and tweaks itself as we gain more experiences. While a belief-based system obviously 'learns' as these beliefs get updated, Predictive Processing also has its own learning process taking place within the hierarchical prediction structure. Jakob Hohwy (2013) has a wonderfully illustrative example of this involving being tasked to plug the leaks in an old dam. You will not be able to plug all the leaks in the dam, but your task is to *minimize* the flow (just like Predictive Processing is in the business of minimizing the amount of error signals in its predictions). In time, as you become more proficient you will be able to start seeing patterns in how the leaks occur. When this happens you will be able to anticipate leaks before they happen and tweak the way you plug the dam accordingly - here Hohwy invites us to imagine something akin to a Rube Goldberg contraption that we tweak and add to over time. Eventually this contraption and your tweaking of it becomes very efficient at minimizing the amount of leaks in the dam, and through clever engineering you can make the machine follow rules and depend on long-term patterns (like floods etc.). In this scenario, the lake on the other side of the dam and all the weather conditions affecting its water levels represent the external world. The dam is the

filter through which we perceive the world (the retina) and our contraption is the Predictive Processing taking place in our mind. In this way, we learn about the world through the patterns of the leaks in the dam, and adapt our means of prediction accordingly. Hohwy makes the important point that "you didn't have to *try* to represent [the world]. All you had to do was plug leaks and be guided in this job by the amount of unanticipated leaks. Similarly, all that is needed to represent the world for the human brain is hierarchical prediction error minimization." (Hohwy, 2013, p.63)

5.3.2 – The Problem of Inattentive Blindness

Kahneman, as I've pointed out previously, makes the point about System 1 that it often leads us to irrational or poor judgements. Kahneman points out that System 1 is "the origin of many of the systematic errors in [our] intuition." (Kahneman, 2011, p.60) One example of this is the issue of the disregard of useful data in favour of biases and emotive responses as I presented earlier in Chapter 3. The relevant case mentioned was the task of categorizing a set of descriptions (100 descriptions of 100 people respectively) between a larger group of engineers (70) and a smaller group of lawyers (30). The result was that most subjects would disregard the additional statistical information given (avoiding use of System 2) and instead of adding this probabilistic data, they simply went with the gut feeling that fit certain stereotypes (using System 1 intuitions). Thus, despite the probability of any one characteristic being much higher to belong to a person in the engineer group, descriptions that fit the lawyer "character" better piled up in the smaller group. The emergent pattern is that System 1, in nature of being fast and independent of working memory, presents low-effort answers that, as Kahneman argues, the lazy System 2 preferably avoids picking up and doing anything with. System 2 is active thinking; it involves effort or "cognitive strain". It would make sense that human nature leads us along the path of least resistance, and if System 1 gives us *plausible* answers, this may be enough for many to not bother thinking any further. Be that as it may, the key point relevant in this case is the claim that System 1 usually produces inferior answers. Is this true in a general sense, and how does it compare to Predictive Processing? Well, visual distance judgement and other perceptual operations are part of System 1, and while we may at times be poor at assigning numerical representations to the distances or sizes we perceive[45], people with good hand-eye coordination are quite apt at judging where a tossed ball will land and accurately get in position to catch it. Rather, in

[45] Roy Mackal, a Loch Ness Monster researcher, has tested the size judging capabilities of alleged witnesses of the folkloric beast and discovered that they often, intentionally or not, overestimate the size and speed of objects on the lake surface. This may also be related to what is known as the size-weight illusion which imparts an evolved innate bias to experience more 'throwable' objects as heavier than other objects of equal weight. For more see Zhu & Bingham (2010).

the opposite manner to what Kahneman suggests, it would seem that in cases pertaining to perceiving things, it is the application of attention (System 2) that sabotages System 1. Take for example the case of the world-famous Invisible Gorilla experiment (Chabris & Simons, 2010). When tasked with paying attention to the number of times a basketball is passed between a group of players, people often fail to notice the man in a gorilla suit walking across the field, something one may think would stand out. This diminishing or distracting effect that attention has on the perception of peripheral or unattended changes in the environment is called inattentional blindness. How could such an effect come to be in the case of Kahneman's System 1 and System 2 distinction? The answer to this can potentially be found in Kahneman's claim that while the activity of System 1 is automatic (it cannot be turned off on command), System 2 can program System 1 in a top-down manner to mobilize attention upon the detection of particular events (Kahneman, 2011, p.104). These events could be spotting a specific word on a page, or a hammer in a drawer. In the case of the gorilla, its appearance does not cause any perceptual surprise as those faculties are currently programmed to trigger attention upon the ball passing between players and nothing else.

If we accept this as a plausible explanation for now, what then could be said for Predictive Processing? The way a predictive visual system has been described, one would think that something as unexpected as a man in a gorilla suit in the middle of a game of basketball would cause a massive error signal marching its way up the hierarchy without being explained away by any hypotheses in place. There has to be some sort of power that attention has over the predictive processes that changes the ways its priorities work. One way to look at it is that it can choke the system, indiscriminately or perhaps selectively silencing or prematurely explaining away error signals that are not related to the matter attended to. This does seem unlikely, or at the very least a clumsy explanation, as it would require a prediction of *unpredicted* error signals, a sort of meta-hypothesis to explain away error signals that wouldn't normally be explained away. A more likely and elegant explanation would be that it has something to do with how attention limits data intake, or rather the range of data that is to be parsed. The focus would as such not be on actively ignoring error signals, but affirming the presence of a specific set of hypotheses' confirmations, letting other things fall to the wayside. Something like this is suggested by Hohwy when addressing inattentional blindness, stating that when "the gain on one signal is turned up, the gain on other signals must be turned down. Otherwise the notion of gain is meaningless: weights must sum to one."(Hohwy, 2013, p.199) So, much like how reliable signals can be laterally weighted, so can attentive action in turn cause its own weighting, causing weighted signals to gain greater priority in the message economy and thus reduce priority of other signals (such as error signals about gorilla costumes). This way, the system sacrifices general accuracy in order to improve on specific tasks. Hohwy makes a further distinction

between two different types of attention – endogenous and exogenous attention. Just to clarify: the word "attention" used here does not necessarily or in all cases refer to conscious attention, but rather the attention which the predictive system gives to certain hypotheses and their performance.

Endogenous attention is the kind of attention that takes the driver's seat in cases such as that provided by Chabris & Simons. Endogenous attention takes part in what Hohwy refers to as "active inference" (Hohwy, 2013, p.201). This kind of inference is very high in its predictive precision and aims to pick out a very selective sample of our perceptual intake. In other words, active inference seeks to predict very specific events at the cost of paying less attention to everything else. It also involves sampling the world for specific data, which very well may involve bodily action like looking around and altering the external world to fit our expectations (for example shuffling things aside to look behind them, stoop under a table etc.). A prime example here is paying attention to the passing of the ball in the gorilla experiment. Other examples include looking for scissors and only scissors in a drawer - or a desk - full or cluttered with other objects. I am expecting to see scissors on the table in front of me, but they aren't there. In such a state one is paying less attention to the identity of objects that aren't scissor shaped and may be prone to miss objects that would otherwise be unexpected finds (for example, a person hurriedly looking for their car keys may completely miss the note from their partner very obviously laid out on a surface within the active search area).

Exogenous attention exists in the system as a counterbalance to endogenous attention. Exogenous attention takes part in "perceptual inference" and is, simply put, there to make sure "non-predicted high precision stimuli are not missed." (Hohwy, 2013, p.201) While endogenous attention can be very closely tied to our conscious experience of 'paying attention to things', exogenous attention is further removed from this experience. It casts a much wider net in an attempt to catch any unexpected changes to the environment, exactly the kind of thing that endogenous attention is very bad at, and update our internal model[46] of what we can hope to predict about the world. Unlike active inference where we expect very specific input and will alter the world and our position in it until this prediction is met, perceptual inference is the "passive mode" (though not truly passive) of Predictive Processing, simply sampling all incoming signals, modifying our predicted model of the world in accordance with error signals so as to explain them away. Normally, whenever perceptual inference encounters a great amount of "surprise" in the difference between predictions and incoming error signals (i.e. the state of the world is too different from what we expected) this event will trigger active inference to take over (we start looking around to see what is

[46] This view of an internal model is very explicitly an internal representation of the world from Hohwy's version of the theory. Other theorists may have a softer version where the internal model is merely a set of prediction patterns that are constantly updated without pushing the envelope of internal representations.

going on). Say I'm heading down the street to the train station while looking at my phone. I look up and realise I can no longer see the train station I expect to be seeing. I turn around and realise that I've simply walked past it. In this case, I encounter a high degree of *perceptual* surprise, causing me to *actively* alter my visual perspective of the world (I turn around) until I can see the train station. These two *modes* of attention have their own zero-sum game of balance where increased pressure on the endogenous mode will cause the exogenous mode to suffer for it. At the extreme end of the spectrum, an increase in the gain of a specific set of channels (which is what occurs during active inference), the stimuli incoming in other channels may not have enough gain to trigger our exogenous attention. If that doesn't happen, the system would fail to turn the scales and revert into perceptual inference mode (Hohwy, 2013). If *that* doesn't happen, we as conscious agents are never alerted to the man in the gorilla suit.

Going back to where we last left off with Kahneman, we were aware that System 2 could instruct System 1 to mobilize conscious attention upon certain triggers. We then assumed that while resources were spent toward this cause, they would be diverted from somewhere else, causing the inattentive blindness. Should explanations about the mind's workings coming from Predictive Processing be compatible with the structural model of System 1 and System 2 - such as that we can be happy with events taking place within System 1 functioning much like how they are described to do in a predictive visual system without any major inconsistencies - then the question of *why* selective attention causes inattentive blindness can be answered in both theories by Hohwy's zero-sum answer. In exchange, the question of *how* the visual system is able to change its weighting in such a manner is explained by System 2 instruction. But while Kahneman's model of System 1 and System 2 uses two very separate systems that operate very differently, can the same truly be said for Predictive Processing?

5.4 – General-level Predictive Processing

If System 1 is truly encapsulated, like a module, then one has to argue that the only place and time when information within System 1 is accessed by System 2 would be when said information is leaving System 1 as output and entering System 2 as input; i.e. when processing of visual input is finalized into an answer to our question; "what do I see?" Going along with the comparison I've been making for the past section, one could argue (as I have) that the same holds true for Predictive Processing; that we can describe our predictive visual system as having a System 1-like relationship with the conscious part of our mind. Whether the cognitive product entering our conscious thought from this largely subconscious system can be defined as a mental representation or not is the issue I

will now turn to in the coming section. Another issue that follows from this is the question of how the automatic nature of this System 1-like visual system links together with the rest of our more deliberate (and if the comparison holds: System 2-like) mind. How far does the predictive model go? Is there a structural, even modular divide between our predictive visual system and the rest of our cognitive being? Well, there one could assert the claim that *all* we do across the board is Predictive Processing, proposing that this model of prediction is not just limited to certain domain specific subsystems, and that even the domain-general level of reasoning engages in the same methods of predicting the world and finding confirmations for hypotheses. That is where the *hard* claim for Predictive Processing can be made.

To clarify, when I refer to the "domain-general" level of reasoning I refer to the kind of analytical reasoning that we perform in deliberation and 'slow thinking', or stream of thought. These are the processes of Kahneman's System 2, or the central processing system of Fodor. A criticism that often arises with this kind of divided structure of peripheral and central systems is that it traps us within our own heads. This view of the mind and its interaction with the world creates a border between us and it. We have peripheral systems that process and filter the raw data we receive from the world and the output these create, these impressions of the world are then presented to our more analytical systems. In systems such as vision, it is only the product of these processes, the representational 'image' that we become consciously aware of. The actual processing of the image happens so fast and automatically that it slips our notice, and in this way could be considered a sub-personal process. In this way we become audience to what the peripheral visual system presents to us, creating a somewhat indirect connection to the world. Taking this divided model of the mind, which as I have explained is an entity on the abstract level of explanation, and directly imposing it upon the world of the human brain, to which explanations belong to a *physical* level of explanation, we end up with something akin to what Daniel Dennett calls a Cartesian Theatre (Dennett, 1991). This is meant as a derisive term aimed at the idea that there is a singular place in the brain where all of visual experience happens, as if there's a little observer inside our heads and the brain is essentially one large projector screen. Now, I would argue that while this is indeed a valid criticism against such thinking about the brain, I do not think the same can be automatically applied to the mind. The division between System 1 and System 2 (or modular subsystems and central processing) is a claim about the structure of the mind and not that of the brain (which, I think, is reinforced by my suggestion that computational mind models would work well in conjunction with Extended and Embodied models). Nevertheless, this way of thinking about the mind *does* awaken similar concerns in those who would make the claim for an all-encompassing model for Predictive Processing.

In the hard claim for Predictive Processing, the idea of a divided mind is not really present in the architecture. So far I have talked about the visual prediction error minimization mechanism as if it was this hierarchically structured module, but the idea actually goes quite a bit further than that, and takes a different spin on the idea of the mind. In this view, consciousness exists, of course, but it is not its own system (or the product of a separate system). Rather consciousness is something that could be linked to (but should not be considered identical to) Predictive Processing's idea of attention (Hohwy, 2013). In order to fully appreciate what I call the 'strong claim' of Predictive Processing, we need to understand how consciousness and world interact in such a system. As I mentioned before, when Hohwy introduces his example of perceptually predicting the world being comparable to plugging holes in a leaky dam with a contraption, he mentions that in a prediction error minimization system, there is no "you" tweaking the plugging contraption. This person is really only there for the sake of illustration. No little man is needed to tweak the hierarchical prediction structure, it does so itself.

5.5 – Perceptual Unity

Jacob Hohwy seems to want to push for an all-encompassing theory of Predictive Processing, applying the prediction error minimization mechanism to all aspects of our mental life. However, in a conservative move he limits himself in his book *The Predictive Mind* (2013) to talking about perception and (to a lesser extent) emotions. When introducing the concept of perceptual unity, what is described is the unity of the phenomenal experience we have of our perception. When we experience the world, it is not a number of different conscious experiences, but a single unified whole (Hohwy, 2013, p.209). He derives this idea from the work of Tim Bayne (2010) who argues that there is a unified *total* conscious state which is made up of the mutually connected *specific* conscious states of what we are experiencing (auditory, visual, sensory etc.). This unity can on a grander scale be called a 'unitary phenomenal field' or 'phenomenal unity' but Hohwy, sticking with perception as the main focus, narrows the scope into the term 'perceptual unity'. Much like Bayne's view, it can be summed up as the unitary experience of all the things we perceive. What does and what doesn't constitute a visual (or otherwise perceptual) experience in Hohwy's view is determined by the posterior probability of our hierarchical hypothesis structures. Again, imagine the case of the dam. In the example there was a contraption representing the hierarchical prediction error minimization mechanism which plugged the leaks in the wall, eventually predicting them through the tweaking of the agent. However, above I also mentioned that in reality the agent is only there to illustrate, and that hierarchical system in Predictive Processing has no such agent. This is because the automatic

System 1-like lower levels of the hierarchical prediction mechanisms in Predictive Processing don't have a System 2-like master system like Kahneman's or other classical views do. What we describe on the fine-grained level of Predictive Processing with all its Bayesian calculations is not processes that we are aware of; this much is intuitive through mere introspection of our own cognitive states. So where does the visual experience enter the picture, at what level are we cognitively aware of our hypotheses' success or failure? Hohwy claims that what we are visually experiencing when 'seeing' something is the "interconnected set of currently best performing predictions down throughout the perceptual hierarchy (down to some level of expected fineness of perceptual grain)" (Hohwy, 2013, p.201-202). In a way this functions much like a peripheral / central system divide – experientially, we are treated to the fruits of our peripheral systems' labour. However, this setup is a lot more fluid in its arrangement. There is no clear border where we say "this process is experienced and this one isn't". Rather, our consciousness is attracted by high posterior prediction probabilities and creeps into our different hierarchical structures at various levels of depth. One could claim that attention has a role to play in this. As stated above, attention and consciousness in this model share a link based on common sense – we cannot focus on things that we are not conscious of, and we cannot be unconscious of things that we are attending to (Hohwy, 2013, p.193). What we *can* say about attention in Predictive Processing, that is reminiscent of elements of Kahneman's System 2, is that it relates to the ignition of active inference. As mentioned above, active inference is the process taking part under endogenous attention. This is when we focus on one particular thing, applying special attention to the performance of a singular hypothesis (or at most a narrow range of them). What is reminiscent of System 2 about this is that Hohwy claims this is where operations take place within a global workspace. In this workspace, the best hypotheses are unified into a fixed model of the world, presentable to conscious thought (a representation of the world if you will). What follows, then, is that endogenous attention focuses on its singular task via active inference using the assumption that, for the moment, the world is as it's represented in the global workspace (Hohwy, 2013, p.215). The processes taking place in active inference are thus the processes we are aware of as the conscious thought or "stream of consciousness" when engaged in problem-solving. In this way, endogenous attention and System 2 share two features with Fodor's central processing: they are limited to a focused flow of tasks and they incorporate information from several systems to create a global model of the world upon which they perform their assessments. In System 2, the relationship between it and System 1 is more separated, much like how one can imagine the encapsulation between modular, peripheral systems and central processing in CTM, while in Predictive Processing error minimization has to overcome a Bayesian "threshold" of plausibility to attract or awaken conscious attention.

So if there is a fluid divide between unconscious and conscious cognitive processes in Predictive Processing, insofar as it is threshold-based, and the hierarchical structures are all separate but unified through a conscious 'field', what ties it all together if not a central system? A good way to describe this, I think, is to imagine a sliced up pizza where each slice represents a hierarchical prediction cascade. The highest levels of the hierarchy are in the centre and extend downward/outward toward the crust, beyond which lies the outside world. As one can see there is no room for a central system here, no separate round slice in the middle, yet all the slices meet up and tie together in a central field, a field that represents our conscious experience of our thoughts and cognition. The field does not have a bounded border but extends and retracts its area depending on where attention leads it – toward the highest precision, best performing predictive processes. There is no output here going from a peripheral system to a central one, rather the mechanism simply keeps predicting, sending down expectations and receiving feedback, explaining away or tweaking signals as it goes along.

5.6 – The Case for Old Vision

By now I have thoroughly presented both Marr's Theory of Vision and its apparent rival in contemporary philosophy, the prediction error minimization framework, and shown how the former does indeed not cut it anymore. However, does this mean that we should take the path of radical enactivism and abandon the idea of internal content? The answer is no. For this to be the case, predictive vision either needs to be unviable or would have to play entirely into the hands of enactivism, and the latter isn't possible given the fact that Predictive Processing is highly dependent on an internal *representative* model of the world. As such, it is more compatible with a computational world-view. Philosophers like Hohwy (2013, 2014) quite happily maintain both Predictive Processing *and* internal representational content, which goes against the core tenets of radical enactivism.

In the next chapter, the spotlight on Predictive Processing will continue as I explore the debate to be had between active and passive cognition. There is a clear underlying conflict in old and new theories of mind when it comes to how we cognize the world and where the causal relations lie, and I think this conflict needs to be addressed. In the context of this chapter, Marr's theory is a clear example of passive cognition, where the world comes to us, the light hitting our eyes causing us to see. In contrast, Predictive Processing is a more active approach in that it anticipates what the world will do next, effectively playing damplugger-role, doing its best trying to minimize the amount of surprises hitting our cognitive net. As I said before when first introducing Predictive Processing, there

are many different versions of this theory to consider. In the next chapter I will mainly consider those of Clark and Hohwy due to their different takes on Predictive Processing's relation to internal representation and the E-theories, as well as a third example more generously geared toward enactivism.

Chapter VI – Active vs. Passive Cognition

In the previous chapter we focused on vision, comparing Marr's (1980) theory of vision to Predictive Processing, a more recent theory of perception put forward by various thinkers, exemplified in Clark (2013, 2016) and Hohwy (2013). What marked the greatest difference between these theories is in how they relate to action and interaction with the world. Predictive processing paints a picture of perception where action precedes input. On the other hand, Marr's theory waits for input and then *reacts* to it. This causes a differentiation between what we could call *active* and *passive* perception. This is not to say that Marr's theory is passive in the sense that it doesn't do anything, and is only acted upon by our various visual inputs, far from it. Marr's theory is a very involved process that also, in a way, 'meets' the input with knowledge-dependent processes in order to translate the incoming data into something that we can understand visually. However, unlike Predictive Processing, it does not anticipate or act before the input. That is the true distinction between active and passive that I will discuss in this chapter: In active perception, the action comes *before* the input. By the time our visual data comes in, we have already made our guesses as to what will happen next. By contrast, in passive perception the action comes *after* the input. Without something prodding it, a passive system would not be inclined to act on its own. In this chapter, I will make deeper distinctions between these forms of perception, and how they relate to cognition in general. I will argue that active cognition is a much better model than the passive one, but that they may both in fact be present in a system. Furthermore, I will also reintroduce the opponent to CTM in the form of enactivism. Here, the threat from enactivism becomes particularly clear, as it also champions the view of active perception and cognition, but does not want to support the idea of representations. Representations, as I will show below, are in fact crucial for both Hohwy and Clark's versions of Predictive Processing, and as such we will see to defending them in this incoming clash.

6.0 – Active Inference in Predictive Processing

Looking back at the previous chapter and the comparison between Predictive Processing and more classical models of cognition, it is clear that what really makes Predictive Processing stand out is the lack of passivity. Traditionally, information about the world is seen as something that enters our mind as input – information that we can then use to think *about* the world and act upon it. Predictive Processing on the other hand *meets* the world with predictions about the world utilizing prior

knowledge gained from previous experiences. The information that comes about in our minds about the world is thus not raw imported data but feedback to our active guesses. Add to that the lack of a System 2 (as I explained at the end of Chapter V with the pizza analogy) and it is clear that in this view our experience about the world is not passively received, but rather actively gained through *interaction*. Hohwy likens the processes of prediction error minimization to mirroring the world. In our predictions, we recapitulate the causality of the world around us in order to grasp it and subsequently predict what happens next (Hohwy, 2013, p.228). To reuse Hohwy's example of the dam: Just as how the world beyond the dam causes the various patterns in the leaks that spring, so must the contraption operate to meet these leaks and plug as many as possible. This becomes a game of mimicry of sorts. Imagine an exercise where two people stand in front of each other, one gets to make any movements they want (the world) and the other has to mimic the movements, learning the patterns to know what comes next in order to keep up with the movements (the Predictive Processing system). Hohwy sees this view of our perception via active inference as one of a fragile nature, because it carries with it a great weakness. In the passive observer view, the world tells us how it is. As long as we can rely on our systems that interpret the signals coming from the world, we can rely on our experience of the world to be true (unless, of course, the world is lying to us, but let's not worry about those Cartesian issues). In contrast, in the world of Predictive Processing *we* are the ones asking the questions, the world merely answers. As such what we know about the world is grounded in what our queries are in the first place (Hohwy, 2013, p.225). Imagine two lecture theatres next to each other. Both are teaching separate classes in the same subject with the same topic, but there is one difference: In theatre A, the lecture takes place much like how we are used to; there is a lecturer holding a lecture, teaching the students about the material that they'll be tested on to pass the course. Much as how a passive cognition system can perform actions to probe for specific information (via bodily movement or deploying increased focus and attention) the students in this hall may ask questions but aren't required to. In theatre B the situation is different. Here, the lecturer *only* answers questions. It is thus up to the students to find out what questions they need to ask in order to get the information about the material that they need. Answers to previous questions will guide the students in figuring out what questions to ask next, effectively improving their abilities for question-making over time. Ultimately the students may become efficient and proficient enough in their questioning of the lecturer that they end up with a similar body of information gained as the students in theatre A. However, the path there has been entirely the work of the students, and failure to ask the right questions would've had catastrophic consequences for their ability to pass the course. When applying this view as a model of perception, the fragility Hohwy speaks of comes from the fact that if we fail to ask the right questions, or do not

know what questions to ask due to lack of statistical regularities in the answers we receive, one could surmise that our model of the world would quickly fall apart.

In terms of applying Predictive Processing to further fields than perception, say emotions and thought, this manner of the world falling apart could be a way to describe the occurrence of mental illness, e.g. psychosis, schizophrenia, body image distortions, depression etc. More specifically for perception though, I would like to refer back to the idea of experiential blindness as I've talked about it in an earlier chapter in relation to the work of Noë. More specifically, I'm speaking of the case where subjects with inverted left eye-right eye input – through a set of special glasses – were losing the ability to see the world in terms of expected regularities (Noë, 2004). When these subjects would navigate their surroundings, their perception would go haywire. When turning their heads to discern the locations and movements of various objects in the world, their predictions (which assumed a non-inverted input) were grossly mismatched with the visual information they received. Apparently, this caused their perception of the world to devolve into chaos, to the point where the sizes of objects seemed to fluctuate before their eyes and others would appear and disappear in and out of view unexpectedly. Their visual experience of the world did not match what was actually going on around them, making them effectively blind or as Noë calls it – experientially blind. This, I think, is what we could picture as an example of what would happen if a person's internal model of the world – based on a predictive error minimization mechanism – were to shatter. When we fail to ask the right questions, and when we fail to grasp the patterns of the world, our experience becomes a constant barrage of surprising events. In this scenario we are not blind in the sense that we cannot see, but rather what we see cannot be grasped to form a robust internal understanding of the world, nor allow us to act upon it. At that point the visual system would be of no practical use for our overall cognitive system, and we would be better off closing our eyes.

To have perceptual experiences that we cannot make sense of is a problem for both passive and active cognitive systems, for obvious reasons, but I argue that it is even more devastating for the latter. Unlike a passive system, active systems are *required* to perform this back and forth dance with the world in order for the system to function. As I mentioned above, the lecturer in theatre B does not speak in anything but answers to questions. If the answers are so unexpected and syntactically detached from the questions we ask, how are we meant to deduce what questions to ask next? This effectively causes a confusion that can cripple the whole system, or at least the part of the system where the relevant questions are asked, in this case the visual "pizza slice" (remember how I modelled the separation of various hierarchical systems as a sliced pizza in the previous chapter). One could even argue that mismatching of posterior probability results between pizza slices (for example mismatched relations between expected sounds and images; a kicked football saying

"quack") could cause a similar, though potentially less shattering, confusion. Imagine standing at a train station when suddenly you hear the rushing sound of an approaching train, you can feel the push of the wind as if something large passes you by at high speed. For all intents and purposes, your tactile and auditory senses express the presence of a passing train right in front of you, yet your visual experience is that of an empty railway platform. One could imagine how confusing such a scenario would be, and the need that would arise in our mind to find an explanation barring invisible trains. There are a number of discovered and tested sensory illusions that highlight just how important it is for our mind to reconcile what our different senses tells us about the world. Take for example the McGurk effect, where the visual perception of someone speaking accompanied with a mismatched dub causes a third sound to be experienced that is warped to fit the movement of the lips (McGurk, 1976). The sound of a person saying "ba" accompanied with a video clip of a person explicitly mouthing "va" tricks our auditory system into hearing "va". Another example would be the rubber hand illusion, also generally known as the body transfer illusion. In this scenario, visual input of someone tapping a rubber hand, synched with the tactile sensation of someone tapping *your* hand, will cause the experience of feeling the tap *upon* the rubber hand, as if it was yours (Ehrsson et al., 2004). The illusion could be taken even further. While the experiment generally fails when tapping objects that are not hand-like in shape (or whatever body part is currently being tapped), the right method can still convince the mind, causing the illusion to occur on seemingly any object. Hohwy & Paton (2010) performed experiments attempting to cause the body transfer illusion on a cardboard box. The first attempt to simply go from normal hand to cardboard box failed, the subjects' prior body image was not explained away in favour of the box. However, first performing the normal rubber hand illusion then quickly swapping the rubber hand for a cardboard box managed to maintain the illusion. Not only that, but the subject actually felt like the tapping on their hand was occurring on the box. Hohwy & Paton suggest that this is due to a Bayesian trick where the subject's mind is gradually changing its expectation of where tapping can occur; from our prior-belief-hand to hand-like objects and finally to inanimate objects, allowing the illusion to persist. The leap from the hand the subjects recognised and had prior beliefs of as "their own" to a cardboard box was simply incredulous to explain away without a step in-between, a gradual change (Hohwy & Paton, 2010).

These two examples definitely give credibility to the idea that the mind is willing to trick itself and even reject (or explain away) previously reliable beliefs, such as our arm not being a cardboard box, in order to consolidate our sensory experiences. In active cognitive models, this becomes a necessary move to maintaining the flow of our predictions and keeping our cognition going. It's not a risk-free mechanism, however. Just as failing to ask the right questions would cause our internal model of the

world to break apart, so can erroneous consolidations lead us astray into a twisted view of the world. Hohwy & Paton note that subjects who are presented with the visuotactile ambiguity found in body transfer illusions were willing to "volunteer supernatural explanations that they normally would never entertain, not even as a remote possibility." (Hohwy & Paton, 2010, e9416) This, they argue, could be important for theorists looking into the formation of delusions, while Hohwy on his own notes more specifically a relation to body dysmorphic disorder, anorexia and schizophrenia (Hohwy, 2013, p.236). Be that as it may, if we return to the example of a single hierarchical prediction mechanism, what is it exactly that prevents it from "shattering" if it is as fragile as Hohwy mentions? The broad answer to this question would be that thankfully we live in a predictable world with consistent causalities. The keyword here is "predictable". Evolutionarily speaking, the only reason why a *prediction* error minimization system such as that presented in Predictive Processing *would* be present in human cognition in our day is because we live in a world where such a system can prosper, i.e. a system based on prediction *needs* a world that is *predictable*. If our world was a chaotic existence where nobody knew what to expect next (jumping could cause you to accelerate laterally, objects like billiard balls travel in random directions no matter the direction or force of impact by the cue, etc.) the world would be *unpredictable* and a system based on predicting future events based on past events would be unworkable. Such a world could not be simulated, nor scientifically explored since any kind of inference would be futile. Obviously, this is not the case (at least not on levels above quantum). This, of course, does not say much above the already obvious, nor does it grant the theory any additional credibility. We are already aware that we live in a world that follows certain observable laws of nature and any theory of mind we construct is in some way constructed upon this fundamental fact.

6.1 – Linking Predictive Processing to EEEE-theories.

When dealing with the divide between active and passive cognition the external world takes on different roles in the systems. Simply the difference between Marr's vision and Predictive Processing causes a change where the world is either an informant regardless of our will, or a teacher that only speaks when spoken to. When taking a step back and looking at the other theories, the E-theories for example, we can see similarities in the arguments being made here and in enactivism's rejection of the inert perceiver (Noë, 2004). It would seem that the conflict between action and inaction in cognition is a key element to this debate, not only in how we perceive but in how we utilize the world in our thinking. As the title of this chapter suggests, this is the debate that I will focus on from now on. I will also note at this point that while a lot of this debate will involve a spotlight being shone

on Predictive Processing, predictive error minimization is merely a vehicle for carrying the debate onward. This debate will also involve the place of internal representations in active cognition, which relates to Predictive Processing's possible status as a computational system.

There are a few questions I would like to pose at this time, in relation to what we will explore in this chapter. First of all, *is Predictive Processing with its pro-active approach perhaps fundamentally a basis for an enactivist theory of sensory cognition (or cognition in general)?* I touched upon this at the end of the previous chapter. It would seem that at this stage there are a lot of comparisons that can be made, and a predictive model for perception could certainly fit the bill in many regards as the sensory explanation for an enactivist theory of mind given the right concessions. However, active processing on its own, I would argue, is not enough for it to make the cut. There would have to be more focus on the interaction between organism and world and not merely internal processes. The interaction between internal predictions and the external world would have to be such that the interaction becomes part of the cognitive system, which also invites possible comparisons to two other E-theories – Extended and Embodied cognition.[47] Thus far, we have not greatly explored these elements apart from the general overview of active inference in endogenous attention and its role in altering the external world to meet our expectations. There is also the role of the body and how prediction error can alter what we identify as part of our sensory selves. Second, *are active cognition theories to be preferred over those involving passive cognition?* As we have seen thus far there is a sentiment that passive cognition (or at least passive *perception*) is a lesser or perhaps even unworkable theory. Enactivism outright rejects it as a possibility because of its adherence to a mind without representational content. Is there a way to account for both active and passive cognition in the same theory? Could a prediction error minimization mechanism be postulated as an active module in a System 1 cognitive mechanism answering to a passive System 2 mechanism as initially posited in our comparison in Chapter V or is this theory only workable as part of the "pizza" model of multi-hierarchical predictive cognition? Would this please everyone or ultimately please no one? Finally, in relation to all cases and outcomes, can Predictive Processing be compatibly interpreted as a *computational* theory of mind? After all, a Bayesian brain sounds very much like a brain that could be described as manipulating internal symbols. These are the questions we will ask ourselves moving onward. What I hope to achieve with this chapter is to prove that even if we let go of passive cognition and embrace its active counterpart as the best explanation for how the mind works, we still do not have to give in to radical enactivism. In this next section, I will perform a comparative analysis between two different interpretations of Predictive Processing, followed by a further comparison between Predictive Processing theory and the E-theories' approach to cognition.

[47] Both of which I've described previously in Chapter II.

6.2 – Is there a border between Predictive Processing-minds and the world?

Two of the main proponents of Predictive Processing, whose work I have been referring to in the past chapter, are Andy Clark and Jakob Hohwy. Interestingly, the two take quite different stances. Clark takes a much more radical approach than Hohwy, being more willing to go further into the E-theory connections. In contrast, Hohwy's approach to Predictive Processing is more conservative and, as will prove beneficial to the ambition of this thesis, takes a defensive stance in favour of internal representations (de Bruin & Michael, 2017). As you may have noticed in the previous chapter when explaining exogenous attention and perceptual inference, Hohwy's version of Predictive Processing emphasises an internal *model* of the world. When applying active inference through endogenous attention we gather the best-performing hypotheses to create a probable version of the world which we then act upon (Hohwy, 2013). It is this version of the world that we consciously and experientially "see" - our experienced representation of the world. This, of course, flies in the face of the generally anti-representational sentiments of the more radical E-theories. Clark, on the other hand, takes an approach of substituting parts of these internal representation responsibilities and suggests that we offload them upon bodily and extended processes whenever possible, effectively circumventing the need for internal representations in these offloading cases - though not entirely across the board (de Bruin & Michael, 2017). If anything, this creates a compromise between the more "moderate" E-theories of Extended and Embodied cognition and classical "input-output" models that could be hard for the more "radical" players (i.e. Enactivism) to swallow. So it would seem at the outset that despite the comparable aspects in promoting active over passive models of cognition, Predictive Processing would have to go through yet another level of radicalization before matching the hard-line anti-representational sentiments of enactivist models.

A big difference in Clark and Hohwy's views comes down to the question of the internal representation of the world and what it is we see when we predict the world. We have experiences of the world. In computational interpretations of vision (including Marr's theory), when we see the world, we have, to put it bluntly, an experiential image in our head of what the world looks like.[48] Looking at Predictive Processing, the core of the theory is that we use prediction to *reduce* the amount of cognitive work we have to do to produce this image, unlike Marr's theory where each "frame" of vision needs to be fully processed through the whole rigmarole of steps from primal sketch to 3D representation. We meet the input halfway to see where we were wrong, basically saying to the world "tell me what's new". As I have noted before, this is an argument for viewing

[48] This in turn opens up computational vision to criticisms of Cartesianism, in particular the Cartesian theatre criticism (Dennett, 1991). I will attempt to counter this criticism in Chapter VII.

predictive vision as an attractive theory simply because it introduces a greater element of frugality that better represents what we can expect of an evolved system. However, it does come with certain counter-intuitive elements. A question we might ask ourselves is: "If what we 'see' is the internal representation in our mind, then isn't the predicted part of what we see simply conjured up by our mind?" This may be a worry because in a case where we walk into a room and by some miracle not a single error signal is generated (the visual input is completely matched by our predictions), then no signals would ever ascend the hierarchy. In such a scenario what we see is not actually what hits our eyes, but what was already in our mind. Clark thinks this interpretation is quite unfavourable, as he prefers the experienced world to be the actual world, and further objects to it being the case. Clark does not outright object to internal representations. Rather, he is all for us possessing them in a Predictive Processing framework. However, his approach is distinguished from Hohwy's in that he doesn't believe that is all there is to it (de Bruin & Michael, 2017).

As a reminder, the core argument of Embodied Cognition is that our bodies are vital instruments in aiding our minds' interactions with the world. We use our bodies in ways to offload, scaffold and improve our capacities to think. In this way, bodily movement becomes part of the cognitive system in gathering information about the world. One could say that the mind at this point becomes a distinct entity apart from the brain, since the Venn diagram of "the mind" encompasses both "the brain" and the areas of "the body" involving actions aiding cognition. Extended cognition takes it further, claiming that we can use external objects in similar ways. As mentioned earlier in Chapter II, Extended cognition follows a 'parity principle' according to which if an external tool can be used and interacted with in a way that is systematically and functionally the same as processes going on purely in our heads, then proponents of this view claim that these tools are equally parts of our cognitive process and thus our mind (Clark & Chalmers, 1998). Active inference would seem to be a prime candidate for qualifying the Predictive Processing framework as involving embodied cognition. Active inference (related to endogenous attention) involves us selectively sampling the world to match our preset expectations. An example I've used before is that of searching for a pair of scissors. In our search, we move around, opening drawers, shuffling objects aside. In essence, we are (a) changing what parts of the world enter our perceptual field through the bodily changing of perspective (tilting your head, moving around in a room) and (b) changing the world itself through physical bodily interaction with it (opening drawers, moving obstructions etc.).

Both Hohwy and Clark seem happy with this connection between Predictive Processing and Embodied Cognition. Hohwy, however, has some reservations about how far this view can be pushed and conservatively suggests Predictive Processing should remain secluded from the external world akin to a Markov blanket (Hohwy, 2014). 'Markov blanket' is a term coined by Judea Pearl (1988) and

denotes a field surrounding a node in a Bayesian network. This field includes all the neighbouring nodes (parents, children etc.[49]) necessary to fully predict the probabilistic behaviour of the node in question. This means that even though there may be additional nodes and connections going beyond this field, they are not relevant to the equation at hand. Imagine pool balls impacting upon each other on a pool table. We denote one particular ball as our focus, ball A. All we need to know to predict the behaviour of ball A in a sequence of chain events is the force and angle of the impact of any ball that strikes ball A (the parents) and the force and angle at which ball A in turn through this impact travels and strikes subsequent balls or walls (the children). We do not need to know the manner in which the parent balls were put into motion as long as we have the interactive data between the parents and ball A. While all balls are needed to describe the network as a whole, the causality sequence of ball A only requires the data of its parents and children. If we could put these relevant events within a field, this field would be the Markov blanket. Back to Hohwy; what is relevant for a hypothesis node in the Bayesian network of Predictive Processing is the evidence (incoming signal) that the hypothesis can explain away. In that way, Hohwy argues, the hypothesis is self-evidencing in that it, by explaining away the incoming signal, provides the system with evidence for itself. All that is needed for any hypothesis *H* is found in its own explaining away of incoming signals E_n. As such, *H* and E_n form what Hohwy calls an explanatory-evidentiary circle - or EE-circle (Hohwy, 2014, p.5). Whatever lies beyond these signals are hidden causes which, of course, matter in the larger system but have no direct contact with *H*. This creates an evidentiary boundary between the EE-circle and the hidden causes of E_n beyond. Imagine the dam again: We do not need direct knowledge about the world beyond the dam to learn how to predict the leaks that need plugging. By learning about the patterns and adapting our contraption accordingly we eventually learn to understand the consequences (leaks) brought about by these hidden causes (seasons, floods, droughts, winds affecting water levels etc.). By understanding the consequences, we can infer hypotheses about the hidden causes beyond, and as such develop a working knowledge about floods etc. through self-evidencing our various hypotheses (plugging mechanisms) with incoming evidence (patterns of leaks). The Markov blanket then becomes the dam - a veil through which we are inferentially secluded from the world beyond. In this way, if we wished to understand the world outside of us, this world could only be inferred, with the leaks taking the role as our evidence. It is thus a structural fact of the system that we are inferentially removed from the outside world. Of course, this does not mean that Hohwy's version of Predictive Processing is entirely divorced from

[49] In a network, parent nodes pass on signals to their children. To illustrate with a case of pool balls: The 'parent' of ball A is whatever ball caused A to move by striking it (passing on kinetic energy from the initial blow), and the 'children' are whatever balls A in turn hits while in motion. As the cue strikes the first ball, this creates a network of parent and child balls where the white ball is the master parent, as it was the first ball to have any energy to pass on, and all other balls are its children. 'Parent' and 'child' are thus causal relations in the network, relative to each ball.

the world. As he points out, if that was the case "there would be no prediction error minimization and thus no meaningful perception, attention or action." Instead the system is in constant open dialogue with the world, and through this constant tuning the produced hypotheses are "supervised directly by the truth." (Hohwy, 2014, p.8-9) One could draw a comparison to Searle's (1980) Chinese room argument, in which the system is completely closed off from the outside world and whatever is produced in terms of giving answers to questions we are presented with, is done so without any real understanding of the subject matter. A similar worry could be raised about Hohwy's dam, where one could argue that the man operating the plugging contraption doesn't really need to have any true understanding or concept of the outside world's floods and seasons etc. in order to effectively learn the plugging. Of course, the refutation to this worry would be that in Hohwy's Predictive Processing, the hypotheses made in the higher hierarchy (thus the abstract hypotheses) are predictions *precisely about the world* and would not even begin to function without concepts (and representations) about said world. If we only considered the very granular lower-level interactions between data and predictions, the objection may work out more favourably. In any case, Hohwy argues that despite maintaining the mind as secluded from the world, it is "an open mind, porous to the world." (Hohwy, 2014, p.9)

6.2.1 – External objects as trusted tools of policy

A worry that Hohwy has about including external objects into our Predictive Processing system as part of that very system is that they then become part of the mechanism that is doing the representing. This is a potential problem for his notion of representing the world because we can have representations of those very objects. Taking Otto's notebook for example: we would be able to have an internal representation of the notebook, while simultaneously having the book be part of the system creating this representation. This causes a blurring of the brain-world boundary, effectively placing the notebook outside and inside the Markov blanket at the same time. Of course, there is already an object that is doing the representing while simultaneously being represented, and that object is the brain. Now, the brain is not "external" in the sense that it is outside of our skull, but the evidentiary boundary is not about the cranium but about the divide between our hypothesis-evidence circle and the hidden causes outside. Hohwy himself admits that the brain is, indeed, a hidden cause in this manner, stating that "[f]rom the perspective of PEM, the agent's causal interaction with external objects such as notebooks and smart phones are, as I said at the beginning, modeled by the agent's brain. As such, the brain holds beliefs about these external objects (and about itself), and these beliefs exhaust the role of those objects in the mental states of the agent."

(Hohwy, 2014, p.13) Even if we try to squirm out of this objection by pointing to the *mind* as being the Predictive Processing system (in either a dualist or a level-of-explanation kind of way) we simply have to rephrase the objection stating that we clearly have internal representations of minds or, much more to the point, Predictive Processing. It is thus not clear what the exact problem is with the notion of objects as both content and vehicle in Extended Cognition other than what Howhy refers to as unattractiveness.[50] As Hohwy puts it, Extended Cognition requires an unnecessarily complex system of two overlapping EE-circles; one that is the brain/mind representing the outside world (including the notebook), and one that is the brain/mind plus the notebook, representing the world outside this coupling (Hohwy, 2014). One could see how this approach would be absurd for self-referential representations as you cannot posit a "brain minus the brain" EE-circle. In this case, the brain/mind is necessarily content and vehicle. But if an exception can be made for the brain, why cannot the same exception be made for Otto's notebook in virtue of it qualifying functionally as part of the system?

This may relate to another objection Hohwy has, which is that there is no clear way in which a notebook participates in the hierarchical message-passing economy such that it fills the same functional role in a Bayesian manner. If a notebook or Smartphone was part of the hierarchy, where would one place it? Hohwy is much more content to leave the notebook as a useful tool. A tool we can put trust in, which in turn helps a great deal with our posterior probability. Hohwy suggests there are many different tools we can use in such a way, even other agents. Otto trusts in his notebook that what he finds there represents things that he's written down and is interested in. The trust in the notebook thus becomes an external *weighting* of sorts, helping Otto to improve upon his decision-making. He doesn't have to second-guess the things written in there as things he endorses (memory replacements), because he has formed a bond of trust with the notebook such that prediction errors generated through interactions with it are preferentially treated, as they are expected to be more precise. *Trust* is thus a more tangible factor for including external objects into an agent's mind without extending it. When Otto hears about an exhibit at the museum, and gets the intention to go there, what he is effectively doing in Predictive Processing terms is that he is predicting input of being at the museum and will alter the world (walk to the museum) until he is there. There are many different ways of getting there however; you can go down the road, through the park or take the bus for example. Hohwy describes these different flows of input as policies, some of which we may have preference toward. Part of Otto's policy is always to check the notebook

[50] In representation, there is a distinction between contents and vehicles. If we take an example from linguistics: words and letters are vehicles, while the meaning behind them is the content. In that sense, "dog" and "perro" are two vehicles with the same content. Representations are symbols that carry meaning. The symbols are the vehicles, and their meaning (their reference to the objects and concepts being represented) is the content.

first. In the notebook, Otto has the address of the museum and thus this policy is of crucial aid to Otto successfully being at the museum. Hohwy states that "not using the notebook will make it much harder for action to reliably fulfil the predictions of being at the museum, resulting in a prediction error increase and, in the long run, a difficulty with remaining within the expected states – such as literally getting lost." (Hohwy, 2014, p.14) In the end, Hohwy claims that anyone defending Extended Cognition in Predictive Processing has the responsibility "to define a plausible evidentiary boundary" and that the new EE-circle created by that boundary "should make it clear that prediction error is minimized for a system including the external object to which cognition is extended, and with respect to hidden causes outside this extended boundary." (Hohwy, 2014, p.12)

So while Hohwy wants the mind to stay unextended (but is happy to involve external tools as parts of trusted policies, or alternatively wishes to see a new "extended" evidentiary boundary established) Clark, on the other hand, is much more optimistic about Extended Cognition co-existing in certain ways with Predictive Processing. Clark proposes that the way we treat our environment is that we modify it to suit our cognition. He calls these types of environments "designer environments" and they involve anything from notebooks to computers, Smartphones etc. but also the methods involved like reading, writing, mathematics and even (as an inter-agent process) schooling. To him, we are not lone cognitive agents when we walk out into the world. Rather, we are surrounded by constructs of our own making, both physical (tools) and societal (practices) that work as scaffolding for our thinking and interaction with the world. We have GPS's in our cars, showing us how to get from A to B without having to memorize locations ourselves (effectively removing the need for a predictive schema of how to reach B from A). Our phones store plenty of personal data and alleviate communications with those we seek to speak with (distance is no longer a problem). We also have a schooling system where we teach our young how to interact with this designed environment and the symbols that mediate our interactions with it. All of this Clark suggests forms "symbol-mediated loops into material culture" (Clark, 2013, p.195) which effectively becomes extended EE-circles between our brain and the technological tools and societal practices that we employ. This takes a step further than Hohwy's trust-based model for external cognitive scaffolding and incorporates the environment into our cognitive process. When Clark states that these information loops are symbol-mediated, the most obvious example to draw is that a lot of these connections are language-based. Whether it is text in notebooks or on a tablet screen, or the transfer of digital data for that matter, the information exchange carry meaning through the use of symbols, much like how classical representational content carries information in the form of *mental* symbols. Clark believes that this way we can propose an action-oriented model of the human mind through a foundational framework of Predictive Processing and then letting Extended and Embodied processes

"fill in the explanatory gaps" that are left behind. The reason for Clark to emphasise the incorporation of the external world is that he finds these "designer environments" to significantly improve our overall prediction error minimization (Clark, 2013). This is in stark contrast to Hohwy, who believes that "the neurocentric mind is the one that best minimizes prediction error in the long run" (Hohwy, 2014, p.13).

6.2.2 – A difference of representational opinion?

De Bruin and Michael (2017) suggest that this difference in approach between Hohwy and Clark may come down to a difference in how they define representations. Hohwy seems to be following a very classical approach to representations, which he defines as: (1) representations can be parts of larger representational structures (like the things in the world constituting a more general representation of the world itself); (2) representations represent things other than themselves; (3) a representation can misrepresent that which it represents; (4) representations can be distinguished from the things they represent. What this is basically saying is that representations are not the objects they represent and as such can be mistaken (WOLF can involve false beliefs about wolves) due to inadequacies of the representing system to fully understand the object it is making a representation of (like lack of knowledge).

This also leads to representations being separate entities from the objects they represent. In many ways, this can be perfectly illustrated by a parallel to the relationship between a painting of a subject and the subject itself. The painting depicts the object being painted, and they are as such not the same entity. In virtue of this, the painter can also be biased or mistaken, and misrepresent that which is being painted (maybe a depicted person becomes taller or more handsome in painted form, or a facial feature is off). Finally, representations can be components of larger representations, such as representations of a dog's fur and legs can be lesser parts of the larger representation which is the whole dog. So while Hohwy maintains his views well within these boundaries due to his caution of the evidentiary boundary between world and mind, Clark ends up in trouble in regards to these distinctions. Clark's version of Predictive Processing involves appealing to what he refers to as Action-Oriented Representations (AOS). What makes AOS diverge from the kind of representations we are used to dealing with is that they aren't only descriptors but also *prescriptors*. What is meant by this is that AOS prescribes how the agent is to interact with the environmental features being represented, which of course due to the changing nature of the environment becomes context sensitive. The kinds of representations are as such not abstract copies of objects (such as a representation of a dog can be imagined as a 'virtual' model of a dog inside of the mind) but rather probabilistic schema for how

the hierarchical prediction system is to engage with the world depending on contexts (Clark, 2016). In this case, the representational content of Otto and his notebook when trying to find his way to the museum consists of the action-oriented representations that he is to walk to the address described in the notebook in order to get to the museum. A representation's relating to a football in the context of a football match would involve all of the relevant actions that one does to a football in such a context (kicking, dribbling, passing and scoring). Such representations are better suited for Extended Cognition since the prescriptive content is more easily functionally represented in things like notebooks and Smartphones. Much like Rowland's (2010) version of Extended Cognition, prescriptive representational models in Predictive Processing can reach out into the world to make additional information available (for example Otto's forgotten memories that were deposited into the notebook in text form). Otto has an action-oriented representational bond with the notebook because he is prescribed to write down information before he forgets it and intuitively reaches out to the notebook to retrieve said information when it is relevant. This bond allows Otto to navigate and engage with the world much like he would if all processes went on inside of his head, a functional parallel which according to Clark's view makes Otto-and-notebook coupling an Extended cognitive system (Clark & Chalmers, 1998). This action-oriented take on representations does make the move to Extended Cognition more plausible, but is something lost in the process? De Bruin & Michael argue that this does not alleviate the problem of a representation being both content and vehicle. They suggest that maybe context-dependence can bypass this issue, that content is only to be distinguished from vehicle in the relevant environmental context – i.e. a notebook is in some contexts vehicle and in others content. The problem then is that since these representations are situated in environmental context, it cannot be decoupled from that environment, which goes against the idea that representations are separate entities from the things they represent (de Bruin & Michael, 2017). Here, Clark would either have to move away from the classical definition of representations and form a new one where this isn't a problem, or hope that the brain exception can be made for notebooks as well. What this illustrates is that Hohwy and Clark indeed do have very different definitions of representations going on in their theories, and thus have different conceptual responsibilities to adhere to, which in turn explains Clark's greater openness. The important issue to consider is if this openness to the world while adhering to a representational content structure is the best way to go for active cognition. As we will see, those taking a radical stance toward active cognition would disagree, as they too believe representations are an impediment to openness, but choose the latter rather than the former. In this way, they take the opposite path to Hohwy, who chooses seclusion over throwing away representations.

6.3 – Three shades of active cognition

There seems to be three main stances that can be taken in active cognition: There's what I would call the classical or conservative stance (Hohwy) which concedes certain E-theory elements like embodied processes but generally remains at a brain-centric view with mind structurally secluded from the world. The moderate stance (Clark) is generally happy with internal representations but does not want the internal kind to be all there is. Here, the sympathy for E-theory elements is greater and the moderate stance encourages a direct link between mind and the external world but is careful not to overemphasise internalism over externalism and vice versa. Finally, the radical stance eschews internal content, including representations, to instead focus on the dynamic mind-world link - taking these interactions to not only be supplementing cognition, but defining it. Neither Hohwy nor Clark support this third view and it is not to generally be found in the Predictive Processing framework. Could a radical Predictive Processing framework be posited? Potentially, but then one would have to come up with a way for how a hypothesis-generating system would work without the explicit usage of internal content. This would leave a very mechanical approach that would require new explanations for consciousness and attention, which would probably be better off entirely replaced by enactivist explanations. In short, a worry would be that Radical Predictive Processing would have to be reworked to the point where we may as well not call it Predictive Processing anymore.

Clark (2015) writes about Radical Predictive Processing in contrast to what he coins Conservative Predictive Processing. Conservative Predictive Processing is exactly of the kind that Hohwy presents. This model for prediction focuses on the mind as an arena, secluded from the world and full of vivid representations that effectively replace the world outside. Clark is not happy with this model of the predictive mind, partly because of its adherence to 'classical computationalism' which he claims does not stand up to the challenges that modern research faces. He mentions for example how this classical computationalism held robotics research back, and how much ground was gained once embodied solutions were approached. Pointing to the success of embodied robotic frameworks, Clark compares ASIMO[51] to the work of Collins (et al., 2001) on 'passive-dynamic' walkers. ASIMO is a robot capable of walking, but it does so "by means of very precise, and energy-intensive, joint-angle control systems." (Clark, 2015, p.10) What this means is that ASIMO's walking is a very involved process, requiring a lot of 'thinking' on ASIMO's part on how to angle its joints for each step. By contrast, Steven Collins' robots use their own bodies as tools in creating locomotion, allowing the mechanism of the legs and the pendulum motion to work on their own without much energy expenditure. This, Clark argues, reflects the principle that the load of problem-solving should be spread between brain, body and world, in accordance with Gibson's (1979) concept of ecological

[51] ASIMO is a robot created in 2000 by Honda, as part of a robot development project started in the 1980s.

psychology.[52] This in turn links into the idea of sensing as a way to couple agent with environment (Beer, 2000) and thus, Clark's (2015) theory of action-oriented representations. Clark argues that the Mataric robot (Mataric, 1990, 1992), which registers landmarks in a maze as a combination of sensory input and current motion, displays the use of exactly these types of representations. A narrow corridor represents both the visual image of the corridor, as well as "forward motion". As the robot creates a map of the maze, it is thus full of visual information as well as recipes for action (Clark, 2015). This argument that computationalism is detrimental is, of course, an argument that we have heard before, and exactly the argument that we seek to disprove in this thesis. As we have seen in the previous chapter, Predictive Processing in its general form is by no means a 'classical' theory in the way that Marr's Theory of Vision was. Clark admits to this, making a clear distinction between classical computational models and Conservative Predictive Processing. The trait they both share is that they are "reconstructive" approaches to perception. What this means is that they aim to take in information about the world to internally construct a rich copy of it. However, where these two approaches diverge is in how they do it. To illustrate, let us use Marr's and Hohwy's theories as comparative examples. In Marr's theory, information is fed into the system and processed, in the classical input-output sandwich way, to produce a finished product, which is our internal model of the world. Hohwy's, on the other hand, meets the information flow halfway with an already predicted model which is then altered to explain away the error signals. In this way, Clark claims the internal model is not so much "built up" as it is "activated" by the confirmation of a high posterior probability. Either way, Clark's concern is that this allows the thinker to "'throw away the world' and select her actions and responses by manipulating the inner model instead." (Clark, 2015, p.15) As showcased above in relation to Clark's positive inclination toward "designer environments", he prefers our tools and constructive connectors to the world occur out where the *actual* world is, in the form of embodied action-cycles or symbol-mediated extended processes, rather than as isolated processes in our heads, thus the idea of throwing away the world becomes a big conceptual issue. One could depict this in an analogy of two artists who enjoy making works depicting natural scenes. The Conservative artist in this case would be the one who goes out into the world with a camera (or maybe even just leans out the window of their house), snaps a picture then takes this back inside locked doors, where they then perform their art in a secluded studio with their photograph (their copy of the world) as inspiration. The second artist, the one that Clark would endorse, would eschew the studio as much as possible, and instead take their creativity out into the world itself. They wouldn't rely on photographs for inspiration but instead make use of the scenes they see there and

[52] Gibson (1979) is a very early example of the idea being proposed that vision is not only about the eyes and brain, but also the body situated in the world. It is an interesting insight into the fact that the groundwork for embodied cognition has in fact been around for as long as Marr's theory (1980).

then, perhaps taking to painting a landscape outdoors on top of a hill. This would be the *Radical* artist.[53]

The difference between Conservative and Radical processing is that while they both 'create' their own worlds, the Radical Predictive Processing is doing so through *action*, i.e. action-based couplings with the world. We have already heard about these couplings above in the symbol-mediated EE-circles extending into the world, or similar interactivity-based solutions involving Embodied compatibility. However, Clark's intent with Radical Predictive Processing is to really embrace the action-centric aspect of Predictive Processing and taking a step further, not merely looking for compatibility with Extended and Embodied theories but making the Predictive Processing model into an *Enactivist* theory of mind (Clark, 2015). While Clark does not enjoy the idea of "throwing away the world", neither can he escape from the idea that the world we perceive is in some way not entirely the objective world "out there". Everything we perceive is always filtered through our own understanding of how the world is, what our place in it is and what interactions are available to us. Even with action-oriented representations a human would have a very different perception of a football from a lion for example. Our concepts of the world simply shape our experiences of it, and an account of this is present in any good theory of perception and/or mind. Even Marr's (1980) bottom-up heavy theory had, as I've shown in a previous chapter, accounts for some top-down knowledge dependent elements with which we identify objects as coherent wholes (we can see a car as a singular thing despite the various dissimilar parts that make it up, like wheels, windows, body etc.). On higher, more abstract levels, certain imagery can make us feel repulsed or joyous because of our understanding and cultural (or just human) connection to what is being presented. To see a scene of an injustice happening may make us feel morally outraged, and in the same way a scene of a wedding celebration can fill us with strong positive emotions like happiness (or negative ones like jealousy depending on your personal history, dispositions and situation). To a lion, these scenes may mean absolutely nothing (or maybe it starts thinking about food seeing all those tasty humans). This is all due to our conceptual connections to our environment. To be able to provide an account for these is *especially* important in a theory like Predictive Processing, where so much of the mind-model is focused on perception and prediction errors. Central processing theories that divide the mind into subconscious domain-specific subsystems and a "passive" domain-general central system (like System 1 and 2 respectively) have a much easier time explaining how the things we know and the kind of creatures we are affect our perceptual data output (intuitions). Such theories can simply point toward our modular makeup being different and producing different types of output

[53] In this sense, Clark is radical in terms of predictive processing (as he invents the term in Clark [2015]), but in my spectrum of active cognition, his position ends up being the moderate stance.

(representations) compared to a lion, or point to the fact that System 2-like features are thought to be more advanced in humans and thus lead to a much richer and nuanced understanding of the world, full of abstract concepts that we project onto it. Either way this leads to the idea that the world we perceive is, to an extent, created by us.[54] Predictive Processing of the kind that makes the claim that "all we do is prediction" does not have the luxury of referring to System 2 capacities but needs to find better solutions for how we 'create' this projection. Hohwy keeps to Conservative Predictive Processing, as Clark calls it, and thus maintains the idea of a representational "model" being perpetually maintained and reconfigured in our minds. Clark, wanting to break out of the skull, so to say, lends his faith to action-oriented representations that can connect directly with the world by representing our action-based relationship with different objects. In this way, we react to a wedding, a coffee machine, or a football differently (in descending order of abstract understanding required) from the way a lion does because we have a different set of actions represented in those environments.

Clark compares this active way of creating one's world to a keyboard under a suspended hammer that goes up and down, hitting the key right beneath it. In order to produce meaningful content, in this case sentences, the keyboard repositions itself. The hammer here represents the world impacting upon our senses, and the keyboard is the agent, creating a role-reversal from what we usually would expect. Just how classical computationalism (e.g. Marr's Theory of Vision) involves the world bringing the meaningful content to a passive observer, what we are used to with a keyboard is that it is the hammer (or the fingers of a human) that manoeuvres about in order to create sentences. Clark instead places the responsibility on the agent, creating a framework where we need to actively sample the world through our own structural couplings in order to make the world come alive in our minds. Wanting to remain true to frugality and the principle that "the goodness of a predictive model is determined by accuracy minus complexity" (Clark, 2015, p.15) these couplings need not even be optimal, but merely satisfyingly good at creating high precision predictions while keeping the structure itself as simple as possible. Clark classifies these action-focused cognitive strategies as a typically Enactivist sort, and also thinks that Radical Predictive Processing "has the resources to cash all these enactivist cheques, depicting the organism and the organism-salient world as bound together in a process of mutual specification in which the simplest approximations apt to support a history of viable interaction are the ones that are learnt, selected, and maintained." (Clark, 2015, p.19) It is thus clear that Clark's ultimate aim for Radical Predictive Processing is to be completely (or at least sufficiently) compatible with the E-theories across the board, while still

[54] We could see this as a form of 'worldmaking' comparable to Goodman (1978).

maintaining some sense of representations in this straggling position between conservative and radical.

While Clark's idea of Radical Predictive Processing might seem like a very attractive stance, combining the best of both worlds into this amalgamation theory, it would do little to sway those from the truly radical stance, i.e. those who want nothing to do with representative content whatsoever. While I do not think that these people *need* to be convinced in order for the moderate stance to hold water, there is a sense in which Clark has shaped his theory to seemingly appease the Enactivist crowd. Despite the content not being (entirely) internal, to the Radical Enactivists the problem that representations present has not gone away. Clark admits that even though Radical Predictive Processing may not be engaged with internal content in the *classical* sense, it is still internal content nonetheless – "rich, frugal, and all points in-between" (Clark, 2015, p.20). Clark is not convinced that enactivism needs to do away with internal content, and presents an alternative type of content (as opposed to classical) that perhaps suits enactivism better. His claim is that what people like Varela et al. (1991) - who were at the outset of Enactivism as we know it - were actually rejecting was not internal representations as a whole but the idea of classical internal representations. The point then, much like I have claimed earlier, is that what they are rejecting is an outdated view that can now be replaced by a perhaps more compatible concept in action-oriented representations. Could this then be the ally that we have been looking for in this thesis, a way for internal representations, and as such computationalism, to survive in the 21st century? After all, Clark has done his best to respect the intuitions of Enactivism, for good or bad, and has set up a system that instead of *mirroring* the world is set up to *engage* with it through structural connections with our environment. Still, there are those who are not happy with the inclusion of these action-oriented representations. In the following section, I will look at some arguments against AOR. This will not only serve to show the perils of the moderate stance, but also lead us into the framework from which the radical stance of active cognition (i.e. Radical Enactivism) operates.

6.4 – The rejection of Action-Oriented Representations

Daniel Hutto (2013b) addresses the viability of action-oriented representations. He claims that there are two ways in which a moderate stance could entice Enactivism to embrace the existence of action-oriented representation; the first being to reject the radical stance by appealing to the necessity of representational content in guiding action, and the second involving an attempt to tame Enactivism by positing a moderate stance in which attention is being given to the enacted vehicles of representational content. The first of these strategies he deems insufficient at making an explanation

that couldn't be posited by dynamical explanations alone[55], while the latter strategy, he claims, renders the existence of content superfluous. Clark's stance is of course the latter strategy, and the one we will focus on for now. So why would Clark's Radical Predictive Processing end up with representational content being made superfluous? Hutto identifies that the core of what action-oriented representations in Radical Predictive Processing attempt to do is to provide for a 'conservative' framework onto which Enactivism can be applied. He dubs this framework Conservative Enactive/Embodied Cognition (CEC). For the purposes of what I'm writing about in this chapter, CEC and RPP are denoting the same philosophical 'move', though they are not to be taken as being identical with one another. RPP is Clark's specific theory of Predictive Processing taking on a Radical turn by involving representational vehicles that structurally connect us with the world. CEC denotes the move to apply, once again, representational vehicles to 'ground' Enactive Cognition in content positive theory, without necessarily speaking of Predictive Processing in particular. Hutto in general prefers vehicle-based explanations for AOR (like Clark's) compared to content-based ones, as he claims the vehicle-based ones sidestep the issue of explaining the causal properties of the internal content itself and instead focuses on the dynamics in how representations cause us to act. Just like de Bruin and Michael noted previously, Hutto too recognises that there is a strong connectedness between the content and the vehicles of AOR. The context sensitive way in which Clark's AORs are embedded in the environment causes the representations to be hard to distinguish from what they represent. In effect, the contents become indistinguishable from the vehicles. Hutto belives that logically speaking this is perfectly fine. However, it causes him to question what the point of these representations then is.

To Hutto, there doesn't seem to be any functional purpose for the content of the representations themselves if it is the vehicular traits that do all the heavy lifting. To illustrate, Hutto draws a reference to linguistics where content *does* have a purpose. In linguistics there is the concept of samesay, i.e. two sentences that differ semantically but carry the same linguistic content. These sentences can be of the same language but formatted using alternate synonymous wording or, even clearer as an example, one specific sentence can be spoken in two different languages, yet still mean the same thing. Hutto takes the example of saying 'it is snowing' in English and 'Il neige' in French. "Making an ontological head count, based on this simple observation, it looks as if we have three things here: two well-formed, complex linguistic utterances – each hailing from different natural

[55] Hutto follows the line of thought that the contents of representations don't themselves seem to make any formal or functional difference in the explanation of how a system is caused to act in virtue of these representations (Hutto, 2013). It is thus to be preferred to speak of the dynamic properties of representations rather than what content is present within them. One could liken this to a weight lifter not caring what material two similar weights are made of as long as they're both solid, 'liftable' objects and create the right amount of resistance when lifting. As such the dynamic properties are more important than the molecular makeup.

languages – and one meaningful propositional content that is expressed by both." (Hutto, 2013b, p.146) The reason why samesay exists is because the sentences (vehicles) and their meaning (content) are distinguishable from one another and the contents themselves carry a functional purpose in the ontology of language. AOR, on the other hand, doesn't seem to have any unique functional purpose that isn't already fully explained by their vehicles' dynamic relation to the agent. A football prompts us to engage with it in certain ways according to our representations, but is it the content itself that does the prompting or is it the structural history between us and footballs? Hutto seems to claim that, to supporters of Clark's theory, it is the latter that truly counts. While happy with the logical possibility of such representational content, Hutto ultimately believes that its existence is fills no functional role in cognition.

I would like to point out that if the terminology in terms of classical/conservative, moderate and radical descriptors seem to reach a certain level of ambiguity, that is because there is a great deal of relativity going on with points of reference here. In regards to the three stances of Active Cognition that I've described in this chapter, Hohwy takes a *Conservative* stance with his version of Predictive Processing, due to his adherence to classical representations. From there, Clark is in the *Moderate* position with his action-oriented representations. Finally, the *Radical* position is the one that eschews representations altogether. However, when Clark is writing about his own (from this thesis' perspective) moderate position, it is labelled as Radical Predictive Processing. This is because in his writings, the point of reference is the spectrum of Predictive Processing theories (which is merely a subset of Active Cognition theories). Being described from a perspective where the origin point is a more neurocentric Predictive Processing model (e.g. Hohwy), Clark chooses the Radical descriptor to point at a move from neurocentrism toward the more action-oriented E-theories. Then, from Hutto's point of view, Clark's position, which has now been described as both moderate and radical, receives another descriptor when Hutto refers to it as Conservative Enactive/Embodied Cognition, or CEC (Hutto, 2013b). The reason why Radical suddenly becomes Conservative is because CEC is described in relation to other Enactivist theories, which places Clark's view on the conservative side of the spectrum, a spectrum on the other side of which lies Radical Enactive/Embodied Cognition (REC). When I can, I will keep to the Active Cognition perspective as the default, and hopefully we run a lesser risk of getting terminologically entangled as we go along.

Hutto believes that just like Chalmers' 'hard problem' of consciousness (Chalmers, 1995b), there is a Hard Problem of representational content. The Hard Problem of Content is born out of the fact that, as Hutto claims, content is always assumed to exist at the outset of any inquiry into the representational mind. Hutto points to the likes of Fodor and Dretske, both of whom did a lot of work into representational content in the heyday of computationalism (Dretske, 1988; Fodor, 1975),

and claims that even then the existence of 'informational content' was already presupposed. The problem that comes out of this is that making such ontological assumptions goes against explanatory naturalism, because the existence of mental representations themselves are never explained through any kind of non-linguistic causation that would give plausible cause for their emergence. What the term 'explanatory naturalism' refers to is the philosophical idea that when we seek to explain things about the world, we do so adhering to natural things, i.e. things that we can find evidence for like laws of nature, rather than relying on the supernatural to fill in the explanatory gaps. It is this kind of thinking that opposes the supposition of mind-body dualism, for example. Instead of starting out with the assumption that the mind is somehow separate from the body, we should only reach such a conclusion if there is no other explanation to be achieved by natural means. Hutto's point is then that, with this Hard Problem in mind, if we can end up with a natural explanation for human cognition that eschews representational content, then we may have fooled ourselves into believing in ghosts all along. Thus Hutto makes two arguments, the first showing why action-oriented representations do not bring anything new to the table that isn't already present in enactivism by default, and the second claiming that representations are assumed rather than derived from naturalistic reasoning to the best explanation.

What this highlights is the Enactivist sentiment that representations should be eschewed because of the apparent lack of need for them. What causes the rift between Clark's moderate stance and the radical stance of Hutto's REC is not a rejection of representations on the basis of an argument that *disproves* them as much as it is on the basis of a lack of argument that *proves* them. Enactivism thus turns the argument on its head, putting the burden of proof upon the pro-representationalists via an argument from explanatory naturalism.

6.5 – Is Active Cognition better than Passive Cognition?

Now that we have gone over the three main stances of active cognition and the move from AOR to Radical Enactivism we can clearly see how Active Cognition enters these three stances in expanding degrees of magnitude. Hohwy's theory, the conservative stance, retains a very classical feel to it and promotes active rather than passive cognition through the idea that we *meet* the world rather than wait for it to come to us. The active part of this theory is an *internal* process, even though bodily action can be included to change the input sample. Clark's moderate stance takes a step further and proposes that not only do we *meet* the world, but our representations of the world are contextually situated in structural connections that reach out of our heads, causing cognition to not only be something that goes on within our skulls but still retaining a focus on neurocentricity. This theory

involves the world to a greater extent than Hohwy's, and Clark's hope for compatibility with enactive cognition generates a balanced focus of internal as well as external active cognition. Finally, the radical stance denies the essential status of representational content altogether, stating that all of our cognition can be explained in terms of empirical ways through which we *enact* the world around us. To deny or leave open the *possibility* of content itself is not the big concern for Radical Enactivism; it rather suggests that all of our cognitive processes can be better explained through dynamic engagement with the environment (via embodied processes etc.). What the radical stance thus truly denies is that cognition always necessarily involves manipulation of content. What this means further is that the phenomenology of perception, which Hohwy places in internal models and Clark in AOR, does not need 'contentful' minds (Hutto & Myin, 2012). This final stance sees our brain more as a mediatory tool through which we make our interactions with the world, rather than a cognitive nexus that houses inner models, and that it is this engagement that makes our experience of the world 'come to life'. This position is unique in that instead of creating an active alternative to internal cognition like the two variants of Predictive Processing, it puts all the focus on the external, interactive processes, where the definition of cognition as "active" is the only viable option. All three of these theories have one thing in common though: they want to move away from the idea of Passive Cognition. The question we then arrive at is; why is Active Cognition seen as a better alternative to Passive Cognition? What are our reasons for seeing the passive theories as unworkable? That is what I'll address in this next section.

To start off, let us look at what led to the move to Predictive Processing, which is represented in both the conservative and moderate stances of Active Cognition. One of the reasons for looking at Predictive Processing in a favourable manner is the explanatory virtue of frugality. A system can be frugal in many ways, but I would like to focus on two aspects: the internal mechanism and the information aspect. I will argue that Predictive Processing is better in both ways. As I've already covered previously, when seeking explanations of how our minds operate it is often good to consider how these minds come about in the natural world. A frugal system requires less spending of resources (energy, effort, mechanistic complexity) on the cognitive, and thus proves more efficient in the long run. The probabilistic element to using Bayesian models of prediction also allows for cognitive models that can operate on a level of finding 'sufficiently good' answers to our inquiries rather than having to achieve full certainty. This, in addition to the ability of a Bayesian system to be self-learning and operate from a very basic starting point, makes for a compelling evolutionary history. Frugal systems do not need to 'tell the whole story' in order to arrive at a juncture where a decision can be made. A frugal system can make do with what is available in a set of incomplete data and still arrive at workable conclusions. For example, an input- frugal visual system would be able to

identify a dog without seeing the whole dog. Indeed, even Marr's theory has explanations for identifying interrupted or partial shapes through the use of generic representations to fill in the gaps until enough information has been presented to generate a complete specific token representation (Marr, 1980). To not have an explanation for representing partial objects would be foolish as such a theory would fail to explain capacities that humans clearly possess. However, frugality works in degrees and what constitutes a 'frugal system' is thus relative. In relation to Marr's theory of vision, Predictive Processing is much more input-frugal in that not only can it handle representation of partial objects through prediction, but this prediction also ensures that the system only needs to work with the *unpredicted* input. This means that in a non-chaotic and fairly ordinary (highly predictable) situation, most of the incoming signals will be confirmed and stopped at the very low levels of the hierarchy. Meanwhile, Marr's theory of vision is always processing all the available visual input in order to refresh our recognition of objects in the visual field. This processing takes place in three steps, creating primal sketches, 2,5D sketches and 3D models out of light signals hitting the eyes. In stark contrast to Predictive Processing, meeting the world with projected knowledge or assumption is possible (to fill gaps, or recognition of generic objects and understanding of shadows etc for constructing 3-dimensional shapes) but not the norm in Marr's theory; everything is processed. The early stages in particular "squeeze the last possible ounce of information from an image before taking the recourse to the descending influence of 'high-level' knowledge about objects in the world" (Marr, 1980, p.206). This, if anything, shows the great divide in frugality between Marr and Predictive Processing, highlighting the kind of information-thirsty cognitive structures Predictive Processing wants to move away from.

With this summary of frugality we can start to see *why* the old theory of passive cognition is seen as unworkable and in need of replacing. As shown above, the move to Predictive Processing is in many ways a move away from Marr's theory of vision. One reason for this is that Marr's theory relies on an almost exclusively bottom-up view of perception. However, this fails to properly account for backward connections in the brain. In Marr's theory, any signals going from higher-level to lower-level (i.e. top-down in a mind sense or 'backwards' in a brain sense) are only there as feedback to patch the holes in the rich, driving forward-signal. This does not, however, successfully paint a comparative picture between mind-level explanations and brain-level explanations (where these backward connections are more functionally prevalent than Marr's theory accounts for). Hohwy deems such a model "straightforward" but "probably unworkable" (Hohwy, 2013, p.225). In response, Predictive Processing embraces the functional role of backward connections, taking these to be sharpening the incoming signals. To the mind, these are the predictive structures explaining away incoming signals and reducing the bottom-up input, level-by-level, leaving only the strongest

prediction errors the further up the hierarchy the signal goes. When Predictive Processing embraces the existence of these connections, creating a top-down prediction system that relies on messages being passed both ways in a cascading hierarchy, it effectively reverses the functional picture. Here, the top-down predictions deriving from our hypotheses (based on internal representations) become the 'rich signal', querying the world, and the bottom-up error signals functionally become the "*feedback* on the internal models of the world" (Hohwy, 2013, p.47).

Assuming we want brain and mind to be ontologically connected (and we do) any theory of mind would have to in some way match the workings of the brain. As I've stated in the first chapter, I separate the mind and the brain as levels of explanation for cognition, where mind may encompass more than just the physical brain. However, if our theory of mind would fail to account for certain capacities of the brain (like the functionality of backward connections), then we run the risk of creating an unworkable dualism. As such, for Marr's theory to compete, it would have to update itself to the modern scientific playing field. This could very well be achievable, but would involve revamping the whole three-step information processing structure to account for a greater functional presence of top-down cognition.

Noë (2004) uses a thought experiment to show how enactivism rejects the idea of a passive or inert observer, something which I have brought up in a previous chapter. In light of our various degrees of Active Cognition, would the passive observer fare any better in regards to either forms of Predictive Processing that we've encountered so far? A passive observer, which we can assume is some form of inert mollusc-like creature with a single instrument for visual perception, would take in visual information from the world much like a stationary camera or a completely paralyzed human being with functional eyes. In this way, we are not looking at the passive observer's outward potential for active cognition, as we already know by the defined parameters of the example any embodied activity is impossible. What we are looking for there is the logical possibility of an active internal system to be situated in an inert body. The question we would have to ask is: would it be able to make predictions? The answer to this in turn ultimately boils down to two follow-up questions: 1) Is a passive observer capable of internally active features? 2) Is a passive observer capable of forming internal representations? Unlike enactivism, Predictive Processing does not rely on external action-based cognition for its "active" qualifier, although both Hohwy's and Clark's version include plenty of externally active elements. My argument here, which I hope to prove with this comparison to the inert observer, is that despite these outwardly active elements, the core of Predictive Processing does not need it. In fact, Hohwy's Predictive Processing doesn't focus on these environmental and bodily activities as part of the cognition itself (as can be seen in its clear separation of mind and world). Rather, what makes Predictive Processing a candidate for Active

Cognition is the way in which it approaches and processes information, opting to meet the world instead of waiting for it, which in large part is an *internal* process. As such, as long as the hypothetical mollusc known as the passive/inert observer is capable of *meeting* the world in this Bayesian manner, it is qualified to be cognizing in an active rather than passive way - much like a person suffering from paralysis would – through Predictive Processing. After all, I doubt Hohwy or Clark would disqualify fully paralyzed people as agents capable of perception simply because of their incapacity to move their bodies. To Predictive Processing, embodied action is not a must, but it is a prevalent feature. To be able to sample the world effectively, or apply schema to alter the world until prediction error signals are explained away – both important parts in how Predictive Processing handles self-evidencing and attention – some amount of activity needs to happen beyond the Markov blanket. The one aspect where the passive observer would be highly limited is thus in its sampling capabilities, as sampling often involves embodied alteration of the world (in the case of vision, bodily manipulation of what enters our visual field). Within our range of incoming signals, however, there can also be further manipulation and sampling in the form of applied exogenous attention. There is no reason to believe an inert observer capable of forming an internal model of the world, and possessing an intellect capable of following the patterns of this world, would be hindered in its way of altering the precision in its predictions in the manner required for active inference. Likewise, such a creature should be able to focus its attention upon certain elements within its visual field, creating a similar zero-sum weighting as is present in exogenous attention. However, such a creature would have no capabilities of altering the world in order to explain away error signals. Whenever an error signal presents itself that cannot be explained away by the internal model, there would be no option but to change the relevant hypotheses. This leaves the passive observer at little more than the absolute basics of Predictive Processing, unable to pursue evidence to maintain a hypothesis. If I was tied to a chair, with my head strapped in place, and saw no pair of scissors on the table in front of me, then I would have no way of looking for these scissors in order to confirm my prediction of seeing a pair of scissors. I may still have the higher level belief that the scissors were present somewhere in the room, but without the ability to alter the environment further, the mystery would remain unresolved. An inert organism that has never had this ability though (and thus may not have a concept of a different perspective or world outside its own static field of view) would likely have a much more analogue approach to the existence of the pair of scissors; either they are there, or they aren't. However, if a certain object was to periodically (and reliably) appear in the visual field like the ducks in a shooting gallery, then such a creature could at least be able to develop self-evidencing hypotheses and expectations about the reappearance of this object. The example of the dam comes to mind, where the contraption plugging the holes has no real capacity to alter the world beyond, but is stuck predicting where the next leak will spring. Still, this shows that Predictive

Processing is not incompatible with the concept of the passive/inert observer the way enactivism is, thanks to its focus on internal processes. Given the presence of an internal world model and the necessary cognitive structures, Predictive Processing is no different between imaginary one-eyed molluscs and real-life humans.

However, what would make all the difference is if the mollusc, in virtue of its state and the evolutionary history involved for such a creature, was incapable of forming these necessary internal representations through which to perform predictions about the outside world. Indeed it is obvious that answering the first of our follow-up questions with a "yes" hinges upon the answer to the second question about an internal representational capacity to be the same. Without representations, the whole idea of hierarchical and internal Predictive Processing falls apart. If there is no world represented, then there is no model from which to draw our predictions of future sensory input. In the case of a human quadriplegic - effectively an inert observer, but by accident and not evolution - we would still be confident in the presence of a capacity for predictive cognition since all things internally remain the same, while externally they'd be limited in the way they apply their embodied functions due to their state. Since by Hohwy's (2014) admission brain and world are inferentially separated from one another, it would not be out of line to even posit an example of a brain in a vat, connected via wires to various information-gathering instruments, and state that this brain could logically possess an internal representation of the world as well as hypotheses about what will happen next in its environment. If the mind is shielded off from the world by the illustrative dam, it does not matter if this mind is placed in a bodily vessel capable of movement or not – any such addition is merely a bonus. There is thus nothing stopping us from saying that, the way Predictive Processing is presented, it is *logically* possible for a bodily inert observer to possess the capabilities of both internal representations and hierarchical prediction structures. The evolutionary story, on the other hand, is a trickier question to answer. According to Hutto (2013) we do not even have a good reason to assume representations in the first place, so coming up with an evolutionary explanation for them would be nothing short of speculation on a philosopher's part. Luckily, we do not need to go down this road with the mollusc thought experiment. The purpose of it is, after all, to show how inertness (i.e. the lack of access to bodily action as a way of interacting with the world) affects our perceptual capabilities (Noë, 2004). It is an argument *for* enactivism, trying to make the claim that logically, we cannot have perception without bodily action. What I hope to have shown here is that this is not the case for Predictive Processing, as we do indeed seem to be able to make a logical claim for inert predictive processors.

6.6 – A statement about perception and cognition.

An important thing to remember is that Marr's Theory of Vision, as it exclusively deals with the processing of visual data, is a theory about the *perceptual* aspects of cognition. These perceptual aspects are also the focal point for large parts of Predictive Processing as well as enactivism, coupled with our interactive relationship with the world. As we presented in Chapter V, the focus on vision is very pervasive in the realm of perceptual cognitive theory, and as we have seen in these past chapters much of the general discussion about greatly varied cognitive theories has been mired in the same realm. Yet, the apparent success in the realm of perceptual cognition can then lead us to claims or expectations that the success of prediction would be equally great in all areas of cognition, i.e. that "all we do is prediction". But are these claims really the way to go just because we've done away with passive perception? Remember that in the previous chapter I proposed the idea of Predictive Processing as a modular entity connected to a more classical central cognitive system, an idea which I contrasted to the 'hard' claim of the hierarchical pizza model. Clark, as I have noted before, is a proponent of this all-encompassing predictive view, as is Hohwy despite his adherence to internal models, since these models in the end are hypotheses for active inference. I would argue that the claim that all we do from a cognitive standpoint is prediction, could be a quite counter-intuitive position at first glance and requires a lot of rethinking of how we describe mental activity. Take for example Otto and his schema for self-evidencing his prediction of being at the museum. From a Predictive Processing point of view, when Otto is checking his notebook he isn't described as accessing his memory to remind himself of the address of the museum in order to find it, he is initiating a schema to transform the world in such a way as to present him in the environment that is the museum. Equally, when I am searching for a pair of scissors in my kitchen drawers, I'm also alternatively described as utilizing precision-altering manipulation of the environment in order to get a scissor shape to enter my visual field and meet in order to reduce the prediction error of not seeing one. In both these cases, the idea is that when I desire to find my way to a certain place or whenever I'm looking for a certain object, what I'm doing is prediction. This prediction is an expectation that I have about what kind of signals will enter my senses, and when my hierarchical hypotheses are not matched by incoming signals, I receive error signals. In essence, what we are doing when we walk to these places or look for these items is fulfilling our expectations by trying to stop being wrong. The actions I perform in travelling to the museum or looking inside the drawers are all ways of reducing the error signals. This means that when I *desire* to go to the museum or to find a pair of scissors, what I'm actually doing from a predictive standpoint is that I already assume I'm already there, or that I'm holding them in my hand. The rest of my actions are not actions to fulfil my desires but

actions made to make sure that I'm right in my predictions. Should these actions fail, I will have to adjust my predictions.

The problem with these explanations is that they do not adequately describe the cognitive phenomenology of the agent. It would be strange to say that what goes through Otto's mind when he performs his entire action sequence - of desiring to see the exhibit, checking the notebook for the address, finding his way to the museum then enjoying the exhibit – is that he is going to change his position in the world until he experiences the visual sensation of being in the museum. Otto may not even know what the museum, or the exhibit for that matter, looks like. Granted, the hypothesis of being at the museum may entail more abstract representations to confirm a matching signal rather than representations of minute visual detail (i.e. he's not looking for a picture-perfect match of some visual museum representation but rather a building that constitutes what he believes the museum to look like, for example a sign bearing the name of the location). Nevertheless, the cognitive story changes from "desire to see X" to "find evidence that seeing X is the case". If Predictive Processing was merely a level of explanation, a piece of a bigger picture, this wouldn't be a problem. However, both Hohwy and Clark stick to the idea that this Bayesian predictive model *is* the cognitive story. How then do we explain the phenomenological story of our beliefs and desires? To be fair, both have admitted that it is not clear how Predictive Processing ought to be implemented for more than perceptual models of cognition (Hohwy, 2013; Clark, 2015), but doesn't then making prediction-based claims on central cognitive elements like internal representations and extended cognition jump the explanatory gun? I would argue that if we want a functioning theory of mind, we cannot focus on the singular aspect of perception and hope that the rest will sort itself out later. Rather, we should do our best to incorporate previous successes in creating a bigger picture.

In the next chapter I will make the argument that while the E-theories and Predictive Processing presented in previous chapters as a whole present attractive new possibilities for thinking about cognition, we cannot allow the move to perception-focused theories like enactivism commit us to theories of mind that ultimately fail to tell the whole story. While there may be a gap between philosophical explanations for mind and neural explanation for brain processes, a gap which we've been historically unsure how to bridge, that does not mean that philosophers should stay away from trying to form cohesive theories about cognitive models. As such, I will be looking to form my Compatibilist Computational Theory (or CCT) in order to, akin to Clark's stance in Active Cognition, form a fusion between computationalism and modern perception-focused theories.

Chapter VII – A Compatibilist Computational Theory

In the previous chapter we had a comparative look at active and passive cognition, and defined what it is for cognition to be active: the core idea is that our cognition involves acting before input, effectively making perception a feedback tool that links agent and world in a dynamic relationship. This could take many forms, but I defined three main strands: (1) We are secluded from the world, but predict incoming input using internal predictive models (Hohwy, 2013, 2014). (2) We predict the world, but our representations of the world also involve action-based connections with the world itself, meaning that it is not the internal model of the world that we act upon, but the actual world outside (Clark, 2015, 2016). (3) We 'build' the world through action. It is thus only through action and dynamic relationships that we have cognition and there is no such thing as an internal model of the world or objects within our heads (Noë, 2004, 2009, Hutto & Myin, 2012). My compatibilist view of CTM, which I will seek to justify in this chapter, is in line with position (1) and could possibly make compromises with elements in position (2). However, it is position (3) that remains a hurdle to be overcome. In this chapter, I will seek to do just that, and defend the idea that our mind requires representative content.

7.0 – Outset to forming a comprehensive theory

As we have seen in the previous chapters, there are a number of ways in which modern day philosophers want to move away from what can be called 'passive cognition', the idea that we are an input-output sandwich to which the world is nothing but a source of inputs, and repository for our outputs, and that all the interesting bits are happening in-between in a secluded space. These theories have their reasons, and as we have seen in previous chapters their way of approaching the move away from passive theories takes very different forms. On one end of the spectrum, we have people who want to keep some core elements of the sandwich, but change its essence from that of a cognitive couch potato to that of an active and skilful anticipator of future input, playing a game of mirroring the world and keeping up with perpetual change. On the other side we have people who want to get out of the sandwich and focus their interest on the (not so boring) bun instead of the filling. To these people who endorse an enactive theory of cognition, our mind is not as much a sandwich as it is a fresh new salad, intermingling with the world and approachable from all angles.

Finally, there's the positions to be found in-between, trying to form a compromise between the ends, a brave move that was not necessarily liked by the radical elements of enactivism due to its adherence to representative content.

In the last chapter, I put these approaches under a common canopy that I named Active Cognition. I noted how all these theories approached the problem of passive cognition from a focus on perceptual theories, which I argued was an oversight, as it ultimately fell short of telling the whole story. In this final chapter, I will address this oversight, and try to mend the fences through the introduction of a Compatibilist Computational Theory (or CCT for short). The purpose of this theory is to avoid throwing the baby out with the bathwater, trying to salvage as much as we can from classical module-based computational models of mind. My goal in this chapter is thus to pick out the so called "best-makers" as Hohwy (2013) calls them, though these best-makers will be theories of philosophical inference that provide explanations toward various issues and questions we have about the structure, contents and physical limits of the mind. Through these I will attempt to find a more comprehensive theory of the mind that is both active *and* passive in different aspects. However, this will not only involve finding various patches of explanations and stitching them all together to fill the gaps in our theory about the mind, but will also involve finding *compatible* explanations that together form a whole. My hope is that I will be able to provide a theory or model that brings out the best of the old veteran in computational theory, such as internal representative mental content, which at the same time meets and intermingles with our contemporary predictive and embodied theories of mind-to-world offloading and interaction. A big part of this chapter will be to defend the existence of internal representations, something I deem essential to computational theory. This, as highlighted in previous chapters, will involve tackling Radical Enactivism, a position that I will aim to refute in favour of CCT.

In the previous chapters, I've presented the collective theories of Predictive Processing, a model of cognition that primarily focuses on explaining perception and how we use it to recognise, learn about and interact with the outside world. Despite this, the success of Predictive Processing tempts its proponents to sometimes suggest that this prediction goes further. In this view, our mind is Bayesian to a much larger extent than sub-personal perceptual processes. Hohwy (2014) makes use of Clark and Chalmers' (1998) thought experiment of Otto's notebook to show how intention to visit a museum can be described as the process of self-evidencing a prediction about a state of the world (or rather an agent's relation to a state of the world, in this case being situated in a museum) and how external tools and representations are used to reduce prediction error and increase precision. Furthermore, when applying Clark's vision of Predictive Processing without an evidentiary boundary between mind and world, these processes could also give rise to extended predictive processes.

Predictive Processing has two strong points in support of it: First of all, it reduces the amount of data required to accurately represent the world by predicting incoming input and processing the feedback in the form of error signals. Second, and building upon the first, it fosters a focus on the interactive back-and-forth nature between mind and the world, which sets up a solid framework for enacted, embodied and extended processes. However, when taking the step beyond 'perception as prediction' and predicating a more universal 'cognition as prediction' viewpoint, does the predictive language hold up? In answering this question, the theory I will attempt to introduce will account for the possibility of Predictive Processing in the perceptual domain while still positing the existence of a central processing system. I will do this by suggesting what is effectively a System 1 and 2 divide, where Predictive Processing can be explained as a module, encapsulated at all levels below the highest in the hierarchy. These higher levels could still be argued to be some sort of connectors open to consciousness and System 2 reasoning. One worry of this project is that I'm positing a model which is neither truly active nor truly passive in nature. When suggesting a System 1 and 2 divide, where the former is largely 'active' (in the way described in the previous chapter) and the latter being more traditionally passive, how does one avoid creating a situation where System 2 becomes an isolated homunculus, and System 1 a barrier of sub-personal processes between us and the world? Why we aren't a homunculus sitting inside of our own heads will thus be another of the issues that I tackle in this final chapter.

7.1 – Dispelling dualistic notions of mind and brain

To recap, this thesis has treated the mind as a level of explanation for cognition in the human brain, as well as potentially offloading (or even extending) processes into our bodies and the world. In that regard, I have at times talked about the mind and brain as almost separate entities. That said, I would like to clarify that I do not want to through this evoke any kind of mind-body dualism where the mind is a separate substance apart from our physical body. Rather, I see this explanatory divide as necessary to talk about the mind in the abstract terms that philosophers and many other cognitive scientists do. When we talk about concepts, internal representations, mentalese, memories and modules, we are talking about things that cannot be observed when we peer into the brain. No more do we observe software when peering into a computer's hardware, yet hardware is the physical basis within which the processes of software take place. Likewise, while the brain is the physical organ through which our cognition occurs, it would not suffice for our understanding of human thought to only speak of cognition in terms of neural activity. The question that arises then in many of the modern models of mind presented in previous chapters is that while there is a parallel relationship of

mind and brain as levels of explanation for thought processes, one abstract and one physical, is this parallel mutually bonded in such a way that the realm of the mind is the brain and only the brain?

While there are many ways to define the mind and its various characteristics, I would argue that the main characteristic is that the mind is the sphere within which cognitive processes take place. As we have seen with theories such as extended and enacted cognition, these seek to include processes that bleed into the world outside of our heads as part of our cognitive processes. If where cognitive processes go, the mind follows, then by that our concept of mind encompasses more than just processes going on in the brain. An extended mind is thus an abstract level of explanation for processes taking place in the brain *and beyond*, showing that while the theoretical Venn diagram of brain processes and mind processes have a clear overlap and connection, the alignment is not perfect. Still, this does not make the mind a distinct entity in a dualistic sense. It is not some form of ghost that reaches out with an ethereal limb to possess Otto's notebook. Rather, there are interactive processes going on, processes that some philosophers and cognitive scientists argue are of the cognitive kind that do not stay within the human skull but rather continue into our bodies and our actions and interactions with the world and, in some cases, external symbols. In contrast, a theory such as Hohwy's model for a predictive yet isolated mind retains a much more aligned relationship between mind and brain levels of explanation, as cognitive processes in Hohwy's meaning do not leave and reach beyond the evidentiary boundary of the human skull. As such, the canopy of what we consider being part of the mind broadens and contracts depending on what any particular theory considers to constitute cognitive processes. This means that when discussing the nature of mind and comparing these differing theories, the processes of the mind have to be treated as separate from the processes of the brain. The mind is still a concept used as a level of explanation for abstract processes that we like to call thoughts, memories etc. but does not constitute a dualistic entity.

7.2 – The homunculus of System 2

The idea of the homunculus relates to internal theories of vision and cognition, and is linked to the idea of an internal observer as the end station of perceptual processes (or any other cognitive process carrying information from the external world into our heads). It is a criticism aimed toward such thinking by pointing out that evoking the existence of an internal observer does not actually explain how or why the visual experience takes place, but merely shifts the responsibility onto explaining the internal observer, potentially leading to an infinite regress. Marr's theory lends itself very well to this criticism, with his idea that we gradually construct sketches and models out of our

visual input data, which is then shown to us inside of our heads. The concept known as the Cartesian theatre proposes that there is an area in our minds where all this data is presented to an internal observer, much like the images of a movie are projected upon a screen. This implies that there is an area in the mind where all this perceptual information *comes together*, and that this is where the phenomenological 'you' exists to experience it, situated apart from the rest of the mind structure. The problem that this creates is thus that we are postulating a reality where we are sitting inside of our own heads, like a separate entity, watching what is projected through the retina and visual system. This creates an unnecessary number of entities within our mind: in order to explain how we (the whole being, body, brain and mind) perceive the outside world, we posit yet another observer sitting within the system performing the kinds of observations that we try to explain in the first place. Imagine if your brain truly was a movie theatre, and the 'true' you was sitting inside 'seeing' what your bodily eyes detect upon a big screen, what then goes on in this miniature observer's head? If our answer is that we need additional internal observers, more of these homunculi, to explain how a mind within the mind observes the world, then we end up in a situation of infinite regress: The homunculus will require another homunculus inside of his head, and so on and so on.

The possible way to step away from this kind of issue would be to say that the observer in the movie theatre is the one and only true observer, the nexus of visual experience within the mind that is. That way, everything else is processing, while the theatre is the place where this processing turns into *images,* made available to something akin to a central processing system. This is very much how a classical bottom-up visual system like Marr's (1980) works; the system processes all incoming data via modular structures, effectively creating a movie reel of representational content that is then projected into the central processing system to be experienced. This is the typical example of the passive observer, where the world enters our senses and through modular processing, like some kind of biological projector, we are presented with the output in the form of images and models within our mind. Even so, this still side-steps the core worry of the argument: why do we need a special place to situate the phenomenology of visual experience? As Daniel Dennett (who coined the term 'Cartesian theatre') puts it, this kind of view states that "there is a crucial finish line or boundary somewhere in the brain, marking a place where the order of arrival equals the order of "presentation" in experience because *what happens there* is what you are conscious of" (Dennett, 1991, p.107). The argument relies on the idea that we literally see images within our heads, but if this internal observer is the way in which we hope to give an explanation to our visual experience of the world, then we have just passed the buck by saying "this is where the magic happens" without actually explaining how the magic trick works. We are, in essence, repeating the problem and ending up where we started. Furthermore, Dennett argues that this idea of a situated individual inside of a

brain is a step toward dualism (thus the Cartesian moniker) and that it is something that should be avoided. Indeed as I've explained above, even when arguing for an explanatory divide between mind and brain I have been careful to emphasise that this is not to be taken as any kind of dualistic statement, but merely the idea that when speaking about cognition, there are two 'levels' at which discussion can be held: one in terms of images, prediction, desires and thoughts (mind), and one for the physical workings of the brain. In keeping with researching and explaining the workings of cognition and our minds as a natural phenomenon, any jump to material dualism of brain and mind is an unnecessary sojourn into the realm of the supernatural and unproven. Dennett's criticism mentions consciousness, but I would like to avoid making that the focal point of this problem. Rather, I will rephrase this worry to something more fitting and directly related to the matter at hand: Why do we need a place where the processed images in (for example) Marr's theory of vision are presented? Marr's theory, as interpreted when open to the criticism of a Cartesian theatre, involves an end station for cognitive processes, a box where all the 'product' is deposited after all the perceptual processing has taken place.

Since models like Marr's seem to attract these kinds of worries, would it instead be worth moving away from visual models that rely on presentation of modular output? It would seem easier to just bite the bullet and say that we're not sure about the connection between the workings of the mind and visual experience, and instead posit that these experiences arise from the workings of the system *as a whole*. This is akin to what Hohwy (2013) suggests, that our 'awareness' or personal-level cognition is the attention drawn to the best-performing predictions in the perceptual hierarchy at the time. This I expanded further in Chapter V to what I came to call the pizza model, with consciousness being a fluid and ever-changing 'field' in the centre. In this way, we can discuss mind and perception without having to indulge in problems of consciousness. Furthermore, visual experiences can be seen as emergent instead of situated.[56] To reiterate, Marr's theory of vision is open to the criticism of suggesting a Cartesian theatre in virtue of presenting vision as an assembly line of visual processing, where the end result is an output image which can then be viewed internally and acted upon. But if the assembly line is where all the actual visual processing takes place, why do we need the image to be presented? The answer is simple: Marr's theory has visual processing as a module, which produces output for a central processing system. Hohwy's hierarchical model sidesteps these worries as the back-and-forth of predictions and error-signals sufficiently provides for an internal model or image of the world to be formed while also accounting for how we respond to it. As such, all there is

[56] That is, there is no finished product delivered to a specific location of the brain. Instead, it is the processing of visual input itself that gives rise to visual experience. Contrast a painter who paints a scene and then shows the finished product to an audience (input-output, where the painting takes place in-between), to a dancer who shows an audience his dance through the act of performing it (emergent through the process).

to vision emerges from that singular system. On the other hand, Marr's theory lends itself to a model of mind where cognition tends to happen in two steps: modular and central. Unfortunately, as I wish to propose a model of similar structure using Kahneman's idea of System 1 and 2, I also find myself suggesting a model in which cognition happens in two steps. Unlike open-border systems like Predictive Processing, any theories involving central processing systems do run the risk of getting stuck in the homunculus' swamp. However, I do have an answer that I think dispels any worries about Cartesian theatres.

It would seem at first glance that any system that models itself after the idea of modular perceptual data being presented to a more centralised system can be a target for taking onboard the idea of a Cartesian theatre, thus making systems with defined internal boundaries and clear input-output structures prime targets for this kind of worry. Outside of worries concerning consciousness, the key issue with the Cartesian theatre on a functional level is that there needs to be a purpose for images to be shifted from a modular process to a central one. If Marr's theory, or any similar theory where visual processing takes place in System 1 and passes on the results to System 2, did so simply because what was theorised about the visual system was insufficient to properly explain perceptual cognition, then the theory would indeed have failed. However, simply by positing the existence of a System 2 we do not make the claim that we lack faith in the integrity of any proposed modular systems in System 1, nor that we want to posit an internal observer removed from the rest of the mind. System 2 has a very distinct function, and that is to take representative content from System 1 and build upon it in more abstract and flexible ways. Unlike a homunculus, which requires another homunculus inside of it, System 2 does not need to be explained by a further System 1 and 2 divide. System 2 is needed because there needs to be an explanation for the divide between fast and slow thinking, for sub-personal and personal cognition levels and, as I'm going to show, for predictive and non-predictive cognitive processes.

7.3 – The language of prediction.

In previous chapters, I've extensively gone over the idea, as presented by Hohwy (2013, 2014) and Clark (2013, 2015, 2016), of perception as an active process involving prediction of future input. As I mentioned in Chapter V there is a distinction to be made between the hard claim and the soft claim for Predictive Processing. The soft claim is, essentially, that *some* of our cognitive processes (the main example being perceptual processes) revolve around prediction. All that would be needed for the soft claim to stand its ground would be to defend the hierarchical prediction error minimization model as a solid contender for explaining our visual system. As I've mentioned before, Predictive

Processing does a good job of this, and wins out over older theories like Marr's theory of vision because of its frugality and Bayesian learning capabilities. The hard claim takes the step further and asks if it could be possible that *all* our cognition in some way or another involves prediction. Could we take this step from a soft to a hard claim, proposing that every cognitive process is somehow linked to prediction, and still maintain an equally defensible position? While the idea that we predict the world before our senses confirm or disprove this prediction may sound like a big step away from the idea of a more classical view - that our senses simply provide us with data that is then processed to provide us with an understanding of what is going on around us - it does not challenge us to change much of what we understand about the world and our relation to it. It is still a fact that light hits the retina and signals travel along the optic nerve to the brain. In fact, the idea that we meet the world with predictions goes hand in hand with scientific discoveries such as backward connections in the brain. Additionally, the idea that we to some extent project internal information onto the world to make sense of our input is already present in older theories of vision such as Marr's (1980), where even then he accounts for our understanding of shadows and depth to turn an otherwise 2D sketch into, eventually, an internal 3D model. To him, the roundness of an apple would not be conveyed in a mental image without a framework capable of interpreting shades as depth. As such, the idea of incoming sensory data being 'met' and configured by pre-existing knowledge or expectations was already present many years before Predictive Processing was developed. What Predictive Processing *does* bring to the table, however, is a stronger framework for new trends in cognition such as the idea of embodied (and perhaps even extended and/or enacted) cognitive processes. It also makes for a more efficient and frugal model of how perception works, where not all data needs to be processed to achieve the same goal. Again, this does not fundamentally change our understanding of perception and how we relate to it, but instead improves upon it and opens up a greater number of possibilities.

However, this is not entirely true when one takes the concept of prediction and applies it to further cognitive areas beyond perception. We can imagine *predicting* seeing something before it is seen, like leaving a cup of coffee on a table and expect to see it there when we return. The idea that us accurately predicting seeing a cup of coffee causes a lesser amount of processing to take place does not as much *change* our perspective on what is going on when we're seeing a cup of coffee as it does *expand* upon it. In contrast, let's revisit an example from the end of last chapter and look at what happens in a Predictive Processing framework when Otto wants to go to the museum: Otto has a desire to go to the museum, so he forms a hypothesis of him being there. To aid in avoiding all the influx of error signals (he is not at the museum yet after all) he may employ a schema to change the world around him until it is in a state such that he is standing at the museum. For example, he may

know that City Bus 11 will take him there, and that taking that bus is an action oriented representation for getting to a number of locations known to Otto along its route. When Otto arrives at the museum, the prediction is fulfilled and thus he has now self-evidenced the hypothesis of being at the museum. However, this usage of perception as prediction changes the way in which we interpret the cognitive processes going on in a person making their way to a museum. When Otto is heading toward the museum, he is not likely to think in terms of this prediction-based chain of actions (all of which would be part of sub-predictions on different temporal scales). He would not describe his journey on the bus as trying to find evidence of being at a museum. He would not be concerned wondering why he isn't at the museum, because he would know that the museum is where he is heading, not where he expects himself to be. The description of Otto's actions from a predictive perspective does not match the mental experience of Otto's actions. This can be contrasted with another example of Otto putting a cup of coffee down on a table, turning around for a second only to later find the cup gone. In this case, Otto does indeed have an expectation to see a cup of coffee, and his frantic searching of the room following its disappearance would accurately match the mental states with the predictive processes taking place; he is looking for evidence of the predicted cup of coffee. Unlike such cases of error signal surprise however, Otto has a *desire* to be at the museum, not a belief that seeks gratification. He is not surprised to be at the museum, nor does he actually hold a belief that he is currently at the museum, yet he *does* apparently have a hypothesis of being there. What Predictive Processing is effectively describing in this situation is that Otto has already predicted being there, and is now sorting out the error signals. In this way of describing a journey to a location as "prediction", the language involved challenges us to change the way we think about agency and the flow of actions.

This is not to say that we cannot halt the language of prediction at a certain stage. We can still apply prediction to a lot of other areas without committing fully to the idea that everything has to be linked to cognition. For example, when Otto is making his way to the museum he may very well apply many predictive processes to lessen the amount of feedback his mind will have to deal with. When he walks, he already has expectations about the ground in front of him being solid, of his legs carrying his body at a certain speed. These are basic motor functions stored in his muscle memory and thus nothing that he concerns himself with or pays much cognitive attention to. We may very well describe these processes as involving prediction, and argue that Otto is, in fact, creating an internal model of expectation as to what will happen when he walks down the street or when he steps on the bus. He may expect to hear a certain fee mentioned by the driver, and when already readying the change from his wallet, will be struck by momentary surprise when he hears an unexpected figure instead (maybe the prices increased). We could even go so far as to say that Otto

may indeed predict what it will be like to be at the museum, and that this internal representational model may be adjusted once he arrives due to minor inaccuracies, or completely shattered when he finds the museum closed or demolished. We cannot go so far as to say that Otto already expects himself to be at the museum, but this seems to be exactly what is suggested be self-evidencing hypotheses described by Hohwy (2014). Is this a fair way to describe what Hohwy intends? The hierarchical structure of Predictive Processing starts with the most general and abstract predictions at the top, where the Bayesian model is formed and updated via feedback. Down this hierarchy there are levels, each of which reaches more fine-grained sub-hypotheses which are, effectively, the building blocks of the hypotheses before. In a simplified manner, expecting a car at one level creates hypotheses for expecting wheels, windows and various other parts and shapes on the level below. On the highest level, Otto *is* self-evidencing a hypothesis for being at a museum, but does that entail Otto maintaining a delusional belief of being at the museum before he actually is? One could argue that no, it is the hypothesis that searches for the evidence, not Otto. But then we have to ask, where is Otto in this system? While Predictive Processing has managed to create a system that functionally is able to describe cognitive processes when we interact with the world, we would have to label these processes as *sub-personal* since they do not describe the thoughts and states of the cognitive agent. Even if we were to accept this, however, the theory has then failed to provide us with a description of Otto's personal-level cognitive processes and states. If we posit a brain level of explanation and a mind level of explanation, shouldn't the latter be concerned with explaining these mental states? I would argue that it should. Yet, what we have here instead is seemingly the introduction of a third level of explanation; a predictive one. If that is the case, then we still need to account for the mental states, which would still fall upon the mind level of explanation, i.e. we have not made progress to explaining Otto's personal-level mental states. The problem that I think Predictive Processing has is thus that it fails to tell the whole story. With that I do not mean to say that I criticise it for not explaining every last detail of the complex processes of the mind, I would not expect that of any theory given how the mind and brain are still largely a mystery to us. What I do mean is that it fails to describe the very apparent thoughts and mental states that we as cognitive agents experience. This would be solved by adopting a soft approach to Predictive Processing, saying that the predictive cognitive processes involved in Otto getting to the museum are sub-personal, domain-specific and modular, categorised as System 1 processes, and that Otto's desires and beliefs are instead linked to System 2, which is domain-general. The hard claim does not have this luxury however.

Let's for argument's sake say we accept the predictive language as valid for describing the actions of cognitive agents, would it functionally hold up to describe cognition if we go further and make the

hard claim that *all* cognitive processes are based on prediction? Would cognitive acts such as remembering or calculating mathematics fit within the Predictive Processing framework? I would make the argument that while prediction would sufficiently explain our interaction with the world, prediction still runs into a problem with many of these internal cognitive processes. What I mean with "internal" here are cognitive processes that are not reliant on any kind of off-loading onto the world. This is unlike processes that bleed into the world or rely on external input to function, which would be anything from perceptual cognition (even if it is behind a Markov blanket) to fully extended cognitive processes. These "internal" cognitive processes are purely mental and internal - a priori - like calculating simple maths without the use of external tools like fingers or pen and paper, or recalling memories without the use of photographs or other external factors that may help trigger said memories. How would we go about describing the processes in terms of predictive cognition? To start with memory, a proponent of Predictive Processing could make an argument that the recalling of memory would work much like our search for a cup of coffee or a pair of scissors. This creates an internal situation where a specific memory is "predicted" and searched for, likely employing schema that allows us to successfully recall said memory. If we say, for example, that memory is sequential, and I want to recall what I did yesterday evening after dinner, I may employ the schema of going through my activities from the time just before I had dinner and retrace my steps. Much like figuring out the 20th letter of the alphabet by sequentially going through the letters from A onwards until I identify T as the 20th step in the sequence, I can retrace my steps before dinner onwards to eventually successfully trigger the memory of what I did afterwards. However, while I employ a schema here much like Otto may employ a schema for arriving at the museum, would said schema be predictive? To predict a memory I would already have had to form some form of expectation of what it is that I want to remember, generating error signals until I have successfully arrived at the desired predicted state. This creates a problem as a hierarchical predictive structure already requires us to create a model to base our hypotheses on. While Otto can form a prediction of what being at the museum would be like, a predicted model of a memory would seem to already *be* a memory. When Otto predicts a room he is entering, he is using his previous experiences of the room, i.e. his memory of the room, to form the internal predictive model. It would seem bizarre to say that we use our memory of a memory to predict said memory. Furthermore, it creates the idea that we have an internal hierarchical pyramid that connects to our memory bank. If we take Hohwy's conservative stance on predictive cognition, then we would find that since our "memory bank" now takes the same functional role as the external world does in perception, this internal hierarchical system would be an enclosed sub-system secluded from the rest of the mind. If we suppose memory as something modular, then this would potentially work, as modular systems are classically described as informationally encapsulated (Fodor, 1983).

What about mathematics? Calculating mathematical problems in our heads is a mental action that takes varied amounts of time and effort depending on the complexity of the problem, but nonetheless there seems to always be *some* amount of effort going into these calculations. Unlike memory, the answers to mathematical problems are not triggered by association unless it involves very ingrained problems that we have readily available answers for, such as 2 + 2 = 4 or memorizing multiplication tables.[57] It would seem that when we approach a mathematical problem as an exclusively internal a priori process, there are two main ways in which it would present itself; either the answer is to us uncertain or we already know it. There is a middle ground to be found here where the result of an addition or multiplication could be gauged, though this I would still label "uncertain". With this in mind, could calculating mathematical problem in one's head be described as prediction? Starting with the situation of mathematical problems with known (to us) solutions, one may pose that this, at the outset, sits rather neatly into the idea of mathematics as prediction in the way of the Predictive Processing hierarchical system. While we could expect the hierarchical structure to be quite a bit smaller (we are already dealing with abstract numbers and as such are not looking into the fine-grained details of perceiving certain number-like shapes the way we would perceptually expect numbers on a calculator screen) we could posit our pre-learned answer ("12" to the question "3 x 4") to be our top-level best-performing hypothesis. But then we seem to run into a problem: if I already arrive at the answer through learned association, does any actual calculation take place that could potentially generate error signals? The answer to that question would appear to be no, I would not be doing any actual calculation to arrive at an answer which my prediction would be judged against. This would only be true if I were to make a prediction through association, and then double-check just to see if it was indeed correct. This would constitute a chain of thought processes that could be described as predicting mathematics, but unlike the case for perceptual prediction, this double-checking is not an automatic part of the system but rather an extra step we would have to add. This is not necessarily a problem though, as a proponent for Predictive Processing could simply make the argument that just as how we need to actively check the results of a mathematical problem we predict, so must we actively check a room we predict the contents of by opening the door. Furthermore, when we double-check by performing the calculation, and arrive at a different result than expected (say I misremembered or was mislead to believe that 3 x 4 was 14) then my surprise at this finding would be analogous to the surprise triggered by a large error signal, followed by an adjustment to my internal model. Then what of mathematical calculations where the outcome is uncertain or not readily available from prior experience? Would these be any worse off for their lack

[57] When I attended school, memorizing multiplication tables was one of the methods in which we learned the sums of multiplications with single digits. Effective way of learning or not, I now associate 7 x 7 with 49 and can provide that as an answer without ever actually engaging in any real calculation.

of a concrete hypothesis? To this, I would have to say no, they would not be any worse off. When looking at the application of Predictive Processing to vision, one would have to accept that there are many situations in which we deal with uncertainty in our hypotheses. If I were to wander into a part of town I've never been before, I do not possess a concrete internal model allowing me to wander around the way I would do in a part of town that I know very well. In familiar areas, we hardly need to look around to find our way and if Predictive Processing is correct, our feedback signals would be minimal as we would be more concerned with keeping an eye on traffic than experiencing new landmarks or navigating the streets. In an unfamiliar environment, we would be much more prone to gather as much data as we can, and our expectations would only be those of the very simple expectations we carry about an organised non-chaotic world. These expectations would for example be of a somatosensory kind at their most basic level, and the expectations of an alignment in movement and spatial change in perspective. We would also from a general knowledge of our life and culture as humans hold expectations of seeing cars, buildings and humans instead of more alien structures and creatures. Equally, prediction in uncertain mathematical equations may be able to gauge the rough size or range of expected sums, or the simple expectations of sums to be larger of added or multiplied, and smaller if subtracted or divided.[58] In this sense, we can also form predictions of mathematical equations that we are uncertain of. However, even as we have done this we have failed to answer the question concerning whether the *doing* of mathematics can be described as prediction. While we can insert the hierarchical model of hierarchical prediction error minimization in a form of internal dam-like relationship between our expectations at the top (behind the dam) and the event of a mathematical equation being worked out at the bottom (beyond the dam) there is little in our prediction error minimization that actually contributes to the act of calculating. This manipulation of internal symbols cannot be described by Predictive Processing but requires an external process to turn the cogs and make the effects upon the dam happen. In the case of Predictive Processing as perception, we can describe our bodily manipulation of the world as embodied processes powered by action oriented representations, and that way get around the fact that much of the processing in our predictive cognition takes place outside of the hierarchical cascades of hypotheses and error signals. However, when it comes to 'internal' cognitive processes, as I've chosen to describe them, there is no other way to turn but within the mind, thus asking for other systems to exist beside Predictive Processing. Many of our cognitive processes can be described as involving prediction, but we need other types of processes, something beyond prediction error minimization, in order to actually perform these mental tasks. As such, Predictive Processing cannot tell the whole story of our mind.

[58] This general rule would of course not be true of negative numbers in addition, or the multiplication of fractions lesser than 1. The rule is presented as simple for the sake of argument.

To conclude, it doesn't seem like Predictive Processing would have the capacity to explain all of our cognitive processes, although it's not entirely incapable of explaining some elements of our 'internal' cognition. That said, if prediction could be the methodology of several of our mind's sub-systems, a System 2 would be needed to explain our higher-order cognitive capacities that simply do not follow the predictive methodology or language. As such, the hard claim for Predictive Processing to encompass *all* of our cognition falls short in favour of the soft claim that *some* of our cognition may be based on hierarchical predictive models. The hierarchical predictive model of perception would potentially fit as a modular system within a System 1 and 2 divide. Its heavy reliance on frequent back-and-forth between mind and world, as well as frequent updating of our internal representative models would mean that perhaps the higher levels of the hierarchical structure were more open to (i.e. less encapsulated from) or even part of System 2. In the end, my key argument prevails that throwing out a System 1 and 2 divide, or to posit all cognition as prediction, causes us to lose out on many other aspects of cognition that are better explained through non-predictive means. Predictive Processing is an intriguing theory of vision, but needlessly restrictive and niche to be applied and endorsed as an all-encompassing theory of the mind.

7.4 - Enactivism and representative content.

While Predictive Processing challenges the notion of the classical "input-output sandwich" by proposing a new model that is a lot more interactive and cognitively engaged and dependant on the environment, both Hohwy and Clark's versions of this theory support the idea of representations. As I described in the previous chapter, the big move that Predictive Processing makes is to change our idea of perception, and indeed cognition in general, from something passive and sequestered to something active and world-engaging. However, as both these versions of the theory adhere to the idea of cognition as manipulation of representational content, none of them really go too far away from computationalism - my argument being that computationalism doesn't need to be held to any adherence to passive cognition. In fact, by showing that Predictive Processing is compatible with a computational theory, I have also shown that computational theory is compatible with active cognition. Hohwy especially, with his conservative stance on Predictive Processing, presents a version of the theory that would be fully compatible with computationalist theory of mind that I seek to defend. The loss of the input-output sandwich in favour of a more E-theory compatible framework is only a blessing, showing that computationalism is capable of keeping itself updated with modern developments. However, I divided up active cognition into three categories – conservative (Hohwy's Predictive Processing), moderate (Clark's Predictive Processing) and radical (radical enactivism). The

big argument that remains against computational theory of mind is the denial of representational content made by this third category - radical enactivism. In this section, I will further explore this criticism, but I also hope to show why, similarly to the previous section, radical enactivism fails to tell the whole story when doing away with internal representational content.

Hutto (2013b) rejects the idea of content on the basis that he claims there is nothing about the content in representational symbols that enriches an explanation as to their function, as illustrated in the previous chapter by his example showing the lack of samesay in action oriented representations - and by extension, representations in general. Instead, the interaction of symbols already does the work by itself, making the proposition of content unnecessary. Furthermore, he claims that there is nothing about the symbols in cognitive processes that improves how we understand the functionality of these processes, since in enactivism this functionality is explained through the structural history between agent and the represented object, thus making the proposition of mental symbols equally unnecessary. What Hutto is doing is effectively a reverse kind of reductionism. Instead of reducing something to its most basic parts, he proposes cutting out lower levels of explanation because of their apparent lack of impact on the higher levels of explanation. When we look at boiling water, we do not necessarily need to know about atoms and molecules to understand the basic properties of the water; we know that it boils at 100 degrees Celsius, we know that it bubbles and that it is hot to the touch. However, knowing about the workings of atoms and molecules enhances our understanding of the boiling water, and also provides us with a level of explanation that provides a perspective that cannot be had by looking at it on the macro scale. Hutto's point is that this is not the case for representational content's relation to cognitive theory. In contrast, as I covered in the previous chapter, Hutto argues that positing content in representational symbols is redundant to the theory of representation, and further argues that representations themselves are redundant to the theory of cognition, and that active cognition theories should instead focus on the interactive relations we claim that these representations (or symbols) mediate. Furthermore, Hutto (2013b) makes the strong claim that not only is there no reason to believe in representations, but that their existence is always assumed at the outset of computational theory and thus inferred without evidential backing. Hutto proposes a Hard Problem of Content, claiming that we are unable to provide proof of any natural causes giving birth to representational content.

Hutto's attack on the computational spectrum of active cognition is thus two-pronged, and the mission to defend computational theory would as such involve two goals: The first goal would be to prove that representations have a functional role to play in active cognition, or that there are areas in which radical enactivism comes out lacking, and would benefit from the existence of representations, internal, action oriented or otherwise. The second goal would be to somehow

provide naturalistic evidentiary grounds for positing representations, or alternatively find an explanation as to why this isn't necessary, attacking Hutto's argument directly.

7.4.1 – Against Hutto's argument.

Is there truth to the claim that representations fill no unique functional role that couldn't be explained away through other more enactivist means? To the enactivist theory of mind, our cognition is determined by our dynamic structures from which we enact the world, but what would keep it from falling into the same trap as Predictive Processing in terms of internal processes? An all-encompassing predictive theory of mind fell short in providing answers for how internal processes, such as mathematical calculations in the head, work. Even if they could be described from a predictive perspective, we require something else to get the whole picture. For enactivism, are dynamic structures enough to explain these internal processes? If not, would representational content be useful or even necessary to fill in the gaps? One could say in the defence of enactivism that its focus is on perception, and as such arguments about internal processes being essentially contentful in a representational fashion cannot be used to thwart their claim of non-contentfulness in perceptual or similarly dynamically structured processes. This argument doesn't stand on very solid ground though, for Hutto makes a claim against representation on a *general* level, and as such has to be defended against on a similar level. Hutto's criticism of representational content is not only in regards to perception, but is a criticism against representations as a whole on an ontological level.

Hutto claims that we require a naturalistic explanation for assuming the existence of representations. However, this argument is a double-edged sword. When making an argument against the existence of a functional role for content, Hutto draws upon a parallel between content in cognition and content in linguistics, pointing at the apparent lack of samesay within thoughts as exists in language. With this move, Hutto has also inadvertently confirmed that while he rejects the idea of content in perceptual cognition, he does affirm its existence in linguistics. What Hutto fails to do, however, is to provide any form of naturalistic argument for representational content within linguistics. Furthermore, in his earlier work Hutto makes the claim that "sentences of natural language, as expressed in linguistically mediated beliefs and utterances, are clearly the paradigms of contentful representations, if anything is. It is also quite clear that when philosophers of mind first developed naturalized theories of content their aim was to explain how *mental* representations could have semantic properties of *just the same sort* possessed by linguistic representations." (Hutto, 2011b, p.48) It is made clear in this passage that if Hutto were to support any type of content, it would be in the realm of linguistics. What this also further implies is that Hutto in this case refers to

language as the tool, separate from our minds much like the rules of basketball. Why can thoughts and beliefs be contentful when mediated through linguistic utterances but require special naturalistic explanations when such contentfulness is theorized as an abstract property of cognition? What extra-cognitive properties does the tool of language have that enables it to transform non-contentful thoughts into contentful utterances? If anything, a fair assessment of the requirements for postulating content would require the same level of scrutiny across the board. One argument that we could propose for this is that spoken and written language provides us with vehicles that can vary but carry the same content (samesay). There are, for example, many different languages in the world that all have words for the same things, though these words (taking the role of representations/symbols/vehicles) are varied and only carry their function as content-vehicles to other speakers who know the same language. In that sense, linguistic representations are like lockboxes, and you need to possess knowledge of the correct language (i.e. the right key) in order to gain access to the contents within. Here, Hutto is happy with content because there is intentionality to language that is not fully explained through the symbols alone, since different symbols can carry the same meaning. This explains Hutto's stance of the first prong of his argument; the attack from functionality. However, it does not give any extra credence to his attack from explanatory naturalism. If Hutto wants to attack the notion of internal representations by attacking the ontological notion of content itself, he would have to do the same for linguistics. Hutto claims that the computational theory of mind owes a naturalistic explanation of mental representations because otherwise, as his accusation goes, these representations are simply assumed out of thin air and the theories around them tailored to fit this assumption. Hutto demands no explanatory naturalism for linguistic representations, because he sees language as being intrinsically contentful, 'if anything is'. The argument of the functionality of content stands, but the demand for a naturalistic explanation falls flat when applied only selectively. Hutto writes that "the major philosophical motivation for seeking to naturalize content is the assumption that linguistic contents derive wholly from and are explained by the properties of the mental representations that underwrite them." (Hutto, 2013b, p.147) To this I agree, that when speaking of meaning in language we are ultimately speaking of meaning in our thoughts, and that as such linguistic representation could be linked to representation in our minds. However, he also adds that since linguistic representations come after mental representations in the causal chain, a naturalistic explanation of mental representations "must be done by appeal to wholly non-linguistic factors." (Hutto, 2013b, p.147)

While he has a point, in light of this it is unclear how Hutto thinks that language, being a product of our internal beliefs, intentions and thoughts, simply by taking external shape as gestures, signs and words gives birth to content where there was none before. After all, as pointed to previously, Hutto

seems happy with content in language. Once again, Hutto lays the explanatory responsibility for content on mental representations, but his own theory dodges the bullet by only assuming linguistic representations. Wouldn't the content of these representations require a non-linguistic causal explanation too? One would have to assume that in Hutto's case this causal explanation would be the non-representational enactive mind, though the link has not been provided. In any case, I would like to argue that computational theory of mind does not in fact owe any special explanation for mental representations from an angle of explanatory naturalism. Even if it does, it is not a keystone that needs to be placed before any credibility could be given to theories involving mental representational content. Hutto is correct in claiming that when philosophers first applied the idea of mental representations, they did so out of a desire to create the notion of an internal computable language. This intent is in fact made perfectly clear by Fodor (1975) proposing the idea of mentalese. The core of that mission could really be stated as: Can the properties of spoken language also be applied as properties of our internal thought processes? The existence of representations is thus not as much assumed as it is a result of projecting the idea of a language onto the processes of the mind. We are, in effect, retracing our steps in order to see how far the linguistic model of meaning and content can carry us. As long as the idea of an internal language holds, and as long as we have no reason to doubt linguistic representations, we have no reason to doubt mental representations either. While I agree with Hutto that a naturalistic explanation for the rise of representational content would be a great success for the theory, I do not think that the matter is so pressing that the theory stands or falls on the current lack of a naturalized explanation, despite the theory's otherwise successful career of providing an abstract framework from which we explain how our mind works. A true attack on exclusively *mental* representations would better take the form of an attack on the idea of a mental language, or the idea that thoughts can be at all computational. The whole premise that Hutto pushes, that the idea of mental content came out of nowhere and was never defended, is simply false. The idea of mental content was born out of the idea that we apply language, as we understand it, to thought. The success of such a theory should stand on its own merit. When attacking computational theory, it should be from the angle of its failure to explain how we think (such as proving that certain cognition is non-computable) or that one of its key components, such as mental representations, fails to fulfil a functional role and as such is redundant to the theory as a whole. This latter is the one valid prong to Hutto's argument.

7.4.2 – A dissatisfactory trade-off.

Even if it is not possible to completely assuage Enactivism's doubt concerning the functional relevance of mental representations, to cast the idea of internal content aside in its entirety is a risky move that ultimately arrives at a theory that only really works for the areas of cognition that it is tailored to, instead of a system that is competent on a general scale. Enactivism works well as a theory of dynamic perception, but even there some philosophers (like Prinz, 2006) argue that enactivism even fails to do this. Ultimately, I would argue that enactivism is dissatisfactory as a general-purpose theory of mind, but even as a specialized theory of perceptual cognition its goals have constricting ramifications with respect to what form the rest of cognition could take. Ultimately enactivism fails because it wants to push cognition in one direction, based on its relative success in a very narrow area; perception. Even if one were to say that enactivism is only a perceptual theory of mind, Hutto's argument against representations is set up in such an all-encompassing way that one could not really let the enactivists off the hook without first explaining how the non-perceptual areas of cognition are supposed to work without mental content. Hutto rejects representation in *general*, and that criticism cannot be selectively applied. If the removal of content causes problems outside of the realm of perception, then successfully crafting a perceptual theory of cognition will not be enough if enactivism leaves the rest of the mind out to dry in its unexplained state. That would result in a partial theory of mind, and we would be better off choosing the more comprehensive theory and handle the much easier problem of defending mental representations.

Even though he criticises enactivism for breaking down the borders between mind and world, Prinz (2006) is by no means bound to classical CTM and is a supporter of the view that there ultimately are no central systems in cognition. He notes that there are a great deal of border disputes between perception and 'thinking', and just as I would agree, these borders need to be renegotiated (in the form of embodied cognitive action, for example). However, instead of maintaining that which I am defending; a central system surrounded by peripheral ones, Prinz argues for a simpler model where "[t]hinking is just seeing with closed eyes, and acting without moving." (Prinz, 2006, p.2) In essence, the only real difference between thinking about moving one's arm and actually moving one's arm is the movement of the arm. This can be likened to how mirror neurons work, where seeing an action performed causes similar brain activity as performing the same action does (Rizzolatti & Craighero, 2004), and the kind of simulation-based thinking that we have seen presented by Clark and Hohwy in their respective predictive theories. Indeed, I am generally in agreement with Prinz' view and I do not think our disagreement on the existence of a central system is in any way crucial to Prinz' criticism of enactivism, which takes place on more common ground.

As mentioned all the way back in Chapter II, enactivism replaces the brain-centric 'traditionalist' view of the mind with a trinity, where mind, body and world come together to give birth to the mind through dynamic bonds of interaction. Prinz (2006) recognises this move as effectively breaking down any and all walls between mind and world. This is a very extreme form of anti-seclusion, a stark contrast to Hohwy's (2014) wishes for a Markov blanket. Enactivism relies on action, because it is the only kind of thinking there can be that breathes life to the agent in the world. As such, states of the mind relating to the experience of a world cannot be reliant on the internal alone, but need to, at least in part, be made up of the bodily action itself. A worry that this raises is that phenomenology is not fully explicable in terms of mind states and, by extension, brain states. The argument seemingly takes on the controversial stance that brain activity is not enough to explain phenomenology. This worry is brought up by Prinz (2006) as he points out that, in enactivism, mental states relating to perceptual experience do not supervene on brain states but on *interactions*. Prinz claims that this view of perception is highly unscientific, but also makes it clear that Noë is by no means in favour of mysticism or dualism. However, by still walking this balancing act on the edge of an unscientific theory of mind, Noë simply ends up with a bad theory. At best, Noë is stating that brain states alone, while probably necessary (brain is part of the trinity still), are not *sufficient* for phenomenology to occur. Prinz is unimpressed by this argument and points to many successful experiments in triggering phenomenology in subjects simply by stimulating the brain via electrodes. In particular, Prinz references the work of Wilder Penfield, the Canadian neurosurgeon, where stimulation of subjects could "generate vivid experiences of songs, conversations, views from childhood windows, and detailed, polysensory episodes from the biographical past." (Prinz, 2006, p.16) Be that as it may, there is the even simpler case of dreams and other vivid forms of imagination-based experiences. These, if anything, show phenomenological experiences that do not depend on dynamic interactions with the world; it all takes place inside of our heads (even though the subjects within the dreams may be based on objects in the outside world). To this, Noë concedes that maybe dreams are part of a set of experiences that do in fact supervene on brain states alone. This, Prinz points out, and I would agree, deflates the whole argument. If enactivism can support the idea that some of cognition is not enacted, and that this is perfectly fine, why make the case that perception is? Indeed, I would agree and say that if such concessions are made, what is proposed as a move out of necessity suddenly becomes really just a move out of preference. Not to mention, this kind of hybrid enactivism makes Hutto's stance absolutely untenable. While Noë may back off on the point of dreams, Hutto simply cannot because Hutto denies the idea of mental content, and would have to explain how and why dreams supervene on dynamic interactions.

An argument in defense of Hutto may come from Churchland (2007, 2012) who argues against the idea of computation, and the existence of representation beyond language. He does so by presenting the case of non-human animals, arguing that it would be absurd to think that when they perceive and navigate the world, forming knowledge about it that they later act upon, these beliefs and actions would in their minds be anything like the propositional beliefs of a language with syntactic structure. In fact, it is unlikely that animals have any linguistic capacities at all, given how hard it is (if not impossible) to teach them language of any kind that we humans use (Churchland, 2007). Since the language of thought as presented by CTM is very much a language in this fashion, what we are proposing then is that humans are fundamentally different thinkers from animals altogether. This Churchland thinks is a false position to hold, and that in fact humans ought to have evolved in their cognition much in the same way as other animals. If we accept that animals do no utilise a language of thought, and that humans, just like other animals, are products of evolution, then we have no reason to believe that human cognition is computational in the manner of CTM. This argument supports Hutto's (2013b) criticism from explanatory naturalism.

My response to that argument is this: Churchland would still fall into Hutto's naturalistic trap because he supports the notion that language is representational. Even if only part of our cognition is language, and thus only part of our cognition involves representational content, the existence of any representations *at all* would still need to be naturalistically justified for Hutto's own argument to be satisfied. Hutto's demand for a naturalistic explanation, in a move to take the burden of explanation away from himself and onto the pro-representational CTM, has in fact backfired because we now demand a naturalistic explanation for language itself by admitting that linguistics are contentful. As long as language is an evolved capacity, it doesn't matter how unique it is, since any evolved capacity carries the burden of naturalistic explanation. If we can explain language, then we can explain language of thought. If we want to avoid this, then we run the risk of claiming that language is something special or mystical, which will not help either stance. Thus if Hutto wants to save linguistics, he also *has* to make an argument that similarly justifies representation. The only real argument Hutto has to stand on is thus that, functionally speaking, we could do without representation in favour of dynamic bonds alone. However, this position I've already argued against above.

Aside from that, I would argue that even if we accept the argument that animals do not have representations, *our* human capacities hinting at representation are so very integrated with the system that you cannot easily assume any non-linguistic parts to be non-representational by virtue of

us being animals too. Take for example the prevalence of corollary discharge in the brain.[59] Corollary discharge is not only a way for the brain to inform itself (sub-systems informing the system as a whole) of actions and motions that are about to happen, but it does in fact also *predict input* (Requarth & Sawtell, 2014) in just the way proposed by Hohwy (2013b). What is more, these predictions are contentful, as we can see in how corollary discharge provides the sensory content that we experience during inner monologues (Scott, 2013). This is thus a strong neural-based case *for* the position that the mind simulates the world internally, perhaps even keeping a model that predicts what will happen next. If we are able to not just anticipate, but also predict and in a meaningful way model sensory consequences before they happen, then we have a strong case for representation in mind. It is logical to say that if there's simulation, then there has to be something in the simulation representing the actual. If we accept all this, then Churchland's (2007, 2012) argument about animals falls apart, because corollary discharge does in fact exist in this way in not just humans but in animals as well (Crapse & Sommer, 2008). In that sense, representation comes before language, and the latter is not a requirement of the former. The fact that humans possess language while most animals don't is thus not an argument against computation per se. Instead, we can simply argue that animals possess a form of proto-computation, whereas our evolved capacity for language has made this much more refined. What's more, we even have a case here to argue that it is exactly these computational capacities that made the evolution of language possible.

7.5 – Concluding on a Compatibilist Computational Theory of Mind.

There is an obvious rift created by philosophers and cognitive scientists who, on the one hand, want a dynamic, out-reaching and sensorimotor-based focus on cognition, setting up their arguments around the issues of mind reaching into the world through embodied, embedded, extended and enactive means. On the other hand, we have those who criticize this move, and who want to maintain a division between mind and world. Amongst these opponents, some may be very conservative, while others endorse at least *some* elements of the E-theories but still maintain a firm division between mind and world (Hohwy, 2014). Furthermore, some want to maintain the border between action and perception, while being sceptical of the internal border of modular and central cognitive systems (Prinz, 2006). I argue that there is a lot to be gained by considering the new ways of looking at cognition that the E-theories bring. By looking at the possibility of extending our minds, to cognize through sensorimotor processes or off-loading of cognitive load onto external symbols, we discover a new active way of looking at cognition, and I do believe that active cognition is the way

[59] Corollary discharge is a copy of a motor command signal in the brain that, while the motor command goes to the muscles, instead travels within the brain to inform other systems if the impending motion (Wurtz, 2013).

forward. However, the projects around this new way of thinking, including Predictive Processing and radical enactivism, put all their eggs into one basket; perception. The move of focusing on perceptual cognition has led to success in fields like robotics, and many philosophers like Clark and Noë have seen this as a sign that perhaps the same success would come if theories of mind put their focus in the same area. Yet, by doing this they leave behind many issues relating to our introverted abstract thinking that do not have much to do with the E-theories at all, such as calculation or imagination (day-dreaming). A problem then arises when we draw up general theories that focus on one part of the mind, while ignoring the potential detriment it could have on the rest. We want to turn the input-output sandwich into a dynamic salad, but maybe the best solution would be a compromise; a dynamic open-faced sandwich? The silly term aside, I argue that it would be a perfectly good move to bring back the idea of a fast and a slow cognitive system; a System 1 and System 2 divide between modular or peripheral systems and a central one. Theories about sensorimotor embodied cognition or prediction error minimization often feature elements that we on a personal level are rarely aware of, or if we are these processes seem to happen automatically through muscle memory or instinct. These processes are fast and largely sub-personal, to the point that we may categorise them as modular or peripheral to our internal stream of thought. If we do that, we can then contrast them with slower and more abstract processes, processes that we are more aware of and are putting more effort into actively thinking about, like in-the-head mathematics or just our everyday internal monologue and philosophising. Since we've adopted a divide between fast, frugal and in some cases sub-personal (System 1) and slow, taxing and open to abstract complexity (System 2) these two different ways of representing thought do not clash with one another. Instead, they can all be posited as part of the same structurally divided yet compatible theory of mind.

Predictive Processing involves frugality and Bayesian probability judgements and learning, but in doing so creates a situation that puts greater emphasis on pre-existing internal models. These models are representational in nature, and the way these representations are manipulated and structured in hierarchical cascades makes them wide open to interpretation as part of a computational system. Despite the downfall of Marr's theory, there is thus not much of a problem in adapting predictive thinking into a computational theory of vision. The computational theory of mind has then not been disproven, but rather improved upon and given a new perspective. In this way, even embedded and embodied elements of Predictive Processing remain representational and, since these processes carry such a heavy focus on internal representational models, maintains a brain-centric perspective despite proposing dynamic structures between mind and world. As opposed to the radical enactivist stance, where brain becomes a mediator between body and world, Predictive Processing emphasises the importance of internal cognition in tandem with sensorimotor processes.

It would not be an inappropriate move to then suppose that Predictive Processing, if only applying it as a perceptual system connected to an internal representational model of the world, could be reinterpreted as a Modular Predictive Processing system, the representational model, alongside our endogenous attention, being a two-way gateway that links it to our central processing. If I expect a pair of scissors on a table, and their absence triggers my attention through heavy error signal feedback, I may actively try to further engage my memory in order to work out why this error may be. Depending on the success in this area, I can then consciously apply a new schema to fulfil my prediction. For example, I may remember that I had put the scissors in the other room earlier and forgotten, so I move over there. If I do not remember, I may engage some general schema based on prior knowledge (check the drawers where I usually put them) or sensorimotor processes to reveal the scissors if potentially obscured (pace about the room, feel around corners, move my head to look behind objects). These strategies in themselves may be described as fast or slow, and my application of them may be decided by the probability of where the scissors are. For example, if a Modular Predictive Processing system detects the lack of scissors on a messy table, the error signal may be explained away by a hypothesis that the scissors themselves are obscured by one of the many other objects on the table. In this case my hierarchical system may apply the sensorimotor schema of shuffling objects about without my System 2 ever being particularly engaged. This may be experienced as routine and almost expected, the judgement delivered by System 1 accepted until further failure leads to more drastic error signals, requiring more extensive revision of my internal model representing the scissors as being on the table, as the probability of it being false rises. From here, I may engage my memory and attempt to remember something about the placement of the scissors that I had temporarily forgotten, or maybe engage in a slow and abstract process of coming up with reasons for why the scissors wouldn't be on the table (maybe I use the fact that it is close to a birthday and the memory that my partner was earlier announcing her intentions of wrapping presents, combining these to arrive at a new possible explanation that she had taken them and would know where they are). What I see in this example is a perfect demonstration of how Modular Predictive Processing can engage with System 2 thinking as well as dynamic embodied off-loading in an adaptable, context based and foremost *compatible* fusion of Predictive Processing and a System 1 & 2 divide, all throughout representational and computational. This, I argue, is the basic functionality of CCT. This CCT is still modular, still computational and still brain-centric, but proposes the mind to be an active computational entity that represents the world internally, yet has the potential to off-load into the world. Through embodied cognition, we use our physical movements to heighten our precision in perceptual sampling, or otherwise improve our cognitive capacities (gesturing while talking, pacing while thinking, mirror neuron activity from observing etc.). Through embedded cognition, we use our environment and the tools available to us as scaffolding to improve our

cognitive potential. We may use notebooks to augment our capacity of memorizing, written equations to off-load mathematical calculations or written language in general (not to mention technology) to improve upon our communicative capabilities. All of this maintains representations, be it actual symbols found in written language or the historical dynamic relations represented by the visual (or other sensory representation) of a tool or place. Yet all these connections lead back to the representations present within the agent. A hammer, football or a school are all representations belonging to a specific agent. While we share these representations among a culture, they carry traits specific to each agent, dependent on their unique experiences of the entities represented. A school (even a specific school) may represent pleasant things in one agent's mind, and something horrible in another's. As such, representations of external objects are shared yet personal, much like a language. We manipulate and organise these representations in order to think about the world, making our thoughts computational in nature.

The input-output sandwich implies that what we want to say when we propose brain-centric computational theories is that all that is interesting goes on in the middle, and that the bread shields us from the boring outside, the environmental and bodily processes. This, I hope to have shown, is not a necessary path of computationalism, as even a secluded mind tucked away from the world behind a Markov blanket can still have a meaningful and dynamic relationship with said world. CCT builds upon this conservative view of active cognition held by Jacob Hohwy and transforms it into a compatible theory that includes the work of Fodor and Kahneman, the former updated to a modern perspective.

Conclusion

To summarise, in this thesis I have argued that the Computational Theory of Mind, despite having to move away from its classical roots in face of contemporary developments, can still retain its core features – a language of thought, built from representational symbols carrying meaningful content, able to computationally manipulate these symbols through processes based on a syntactic set of rules, these processes being divided into domain-general and domain-specific categories – and remain a workable theory of mind that is both versatile and full of depth. This is despite popular arguments to the contrary: that our thoughts are non-computable (Penrose, 1994), that CTM's reliance on mental content and representations fulfills no function (Hutto, 2013b), that the passivity on computational perception models fails in comparison to action-based perception (Noë, 2004) and that the idea of the mind as situated solely in the brain embraces Cartesianism (Noë, 2009). I have argued against these arguments and claimed that they are, at best, aimed at an outdated version of CTM, one that I have revised and transformed into a contemporary version, compatible with elements of cognition that avoids these criticisms, while also moving the theory away from its failings in the face of contemporary science.

I have argued for the importance of modularity for a cognitive system to achieve a level of frugality, a trait that is attractive from a naturalistic standpoint to explain how a computational information processing system could have evolved in an environment where we rely on management of resources to survive. This defence of modularity has been important to my case, as I desired to create a theory that can explain seemingly context-based competency in human psychology, as well as serving as a foundation for processes that I want to deem automatic, fast, frugal, domain-specific and inaccessible to our direct awareness. I've compared accounts of modular systems to theories about heuristic psychology (Gigerenzer, 1999, 2008; Tversky & Kahneman, 1974), and argued that the two have a lot in common; to the point where I've suggested certain heuristics are part of modular systems, thus their context-sensitive application and focus on fast and frugal decision-making. I've shown examples of where these modular and heuristic capacities can be observed in human behaviour (Wason, 1968) as well as explained that we have good reason to believe these capacities to be of a computational nature, since they are described in terms of logic terminology (Gigerenzer, 2008) and can be expressed in mathematics and computer science in a formulaic manner (Michalewicz & Fogel, 2000).

I have addressed Marr's theory of vision (1980, 1982), a theory that I identified as a typical example of computationalism applied to perception. This theory, I argued, was the typical image of a cognitive model following the passive, input-heavy version of computationalism, and as such also suffered from the abovementioned criticisms against classical CTM. I reviewed the theory and concluded that while it was an accomplishment to express vision in such a detailed computational and mathematical manner, it was far too input-hungry to escape the recent trends in cognitive science, and the criticisms aimed against classical CTM. However, I presented a contemporary alternative theory of perception in the form of Predictive Processing (Clark, 2013, 2015, 2016; Hohwy, 2013, 2014). This theory still retains the elements of computation and reliance on internal representations, but it revolves around a Bayesian form of hierarchical prediction that allows for quicker perceptual judgements to be made while relying on less perceptual data intake. This creates a computational model for vision where we meet the world through action and prediction. I argue that this does away with the worry that CTM relies on only passive processes (expressed as a worry by Noë, 2004), replacing the passive cognition of Marr's theory with the active cognition of Predictive Processing.

This move to active cognition scores additional points for CTM, as it opens up the theory to more easily incorporate embodied elements to cognition, making a stronger case for CTM as being engaged in the world. The point of being engaged with the world was something that I deemed a good point to defend, as it would reduce the impact of criticism rooted in comparing computationalism to a form of Cartesianism. However, I've maintained that prediction in the manner of Predictive Processes is best kept as a modular system of vision, and that I did not believe in prediction as an all-encompassing method for all of our cognition. I referenced introspective mental processes like memory recall and mathematical calculations as examples of types of cognition that are unlikely to involve prediction as their *modus operandi*. Additionally, I've pointed to how the language of prediction changes the way we view an agent's intentions. Taking the example of a person wanting to go to a museum, if Predictive Processing was an all-encompassing theory, this agent's action would be described in terms of predicting themselves already being at the museum, then altering the world around them through bodily movement in such a way as to make this state of being true (Hohwy, 2014). I've argued that there has to be a level at which an agent's motives and thoughts more accurately describe what we as people actually experience our motives and thoughts to be like when we seek out museums, and as such suggest that we need to have a domain-general level of cognition outside of our predictive capabilities. To this end, I've suggested a System 1 and 2 split between our domain-specific and domain-general capabilities where prediction-based cognition

fits better into the System 1 category than System 2. I furthermore argued that the System 1 category can involve modular systems, since there is a lot of descriptive overlap between the two.

With this, I have created a model for the mind that incorporates both the pro-active elements of embodied and embedded cognition, as well as Predictive Processing. This is a theory that meets the world and creates our internal model through action. Unlike radical enactivism, this new Compatibilist Computational Theory of Mind also supports a rich internal mind model, involving representational content being manipulated through computational processes. It retains features such as modularity, which allows us to explain context-based cognitive competence at certain tasks, as well as an apparent informational encapsulation and limited accessibility related to peripheral and sub-personal thought processes. This complexity is beneficial to the theory, as it allows it to explain more of our cognitive faculties than what the enactivist alternative does, since enactivism mostly focused on action and perception, and does not make a convincing enough account of exclusively internal thought processes.

References

Adams, F., and Aizawa, K., 2001, 'The bounds of cognition', *Philosophical Psychology* 14, no. 1: 43-64.

Adelson, E. H., 1995, 'Checkershadow Illusion', http://web.mit.edu/persci/people/adelson/checkershadow_illusion.html, retrieved 2014-06-02.

Adelson, E. H., 2000, 'Lightness Perception and Lightness Illusions', in Gazzaniga, M. (Ed.), *The New Cognitive Neurosciences*, 2nd ed, pp. 339-351, Cambridge: MIT Press.

Bayne, T., 2010, *The Unity of Consciousness,* Oxford: Oxford University Press.

Beer, R. D., 2000, 'Dynamical approach to cognitive science', *Trends in Cognitive Sciences* 4, no. 3: 91-99.

Block, N., 1995, 'On a confusion about the function of consciousness', *Behavioral and Brain Sciences* 18: 227-47.

Boysen, S. T. et al., 1996, 'Quantity-based inference and symbolic representation in chimpanzees (pan troglodytes)', *Journal of Experimental Psychology: Animal Behavior Processes*, 22: 76-86.

De Bruin, L. & Michael, J., 2017, 'Prediction error minimization: Implications for Embodied Cognition and the Extended Mind Hypothesis', *Brain and Cognition* 112: 58-63.

Carruthers, P., 2006, *The Architecture of the Mind*, Oxford: Oxford University Press.

Carter, W. R., 1990, 'Why Personal Identity is Animal Identity', *Logos* 11: 71-81.

Chabris, C. & Simons, D., 2010, *The Invisible Gorilla – And Other Ways Our Intuitions Deceive Us*, New York: Harmony Books.

Chalmers, D., 1995a, 'Minds, Machines and Mathematics', *Psyche* 2, no. 23: 11-20.

Chalmers, D., 1995b, 'Facing up to the problem of consciousnes', *Journal of Consciousness Studies* 2(3): 200-19.

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

Chomsky, N., 1968, *Language and Mind,* Cambridge: Cambridge University Press.

Chomsky, N., 1988, *Language and Problems of Knowledge: The Managua Lectures*, Cambridge: MIT Press.

Churchland, P.M., 1989, *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science,* Cambridge: MIT Press.

Churchland, P. M., 2007, *Neurophilosophy at Work,* New York: Cambridge University Press.

Churchland, P. M., 2012, *Plato's Camera: How the Physical Brain Captures a Landscape of Abstract Universals,* Cambridge MA: MIT Press.

Clark, A. & Chalmers, D., 1998, 'The Extended Mind', *Analysis* 58, no. 1: 7-19.

Clark, A., 1997, *Being There: Putting Brain, Body and World Together Again*, London: MIT Press.

Clark, A., 2003, *Natural-Born Cyborgs,* Oxford: Oxford University Press.

Clark, A., 2008, *Supersizing the Mind*, Oxford: Oxford University Press.

Clark, A., 2013, 'Whatever next? Predictive brains, situated agents, and the future of cognitive science', *The Behavioural and brain sciences* 36, no. 3: 181-204.

Clark, A., 2015, 'Radical Predictive Processing', *The Southern Journal of Philosophy* 53, Supplement S1: 3-27.

Clark, A., 2016, *Surfing Uncertainty: prediction, action and the embodied mind*, Oxford: Oxford University Press.

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

Collins, S. H., Wisse, M. & Ruina, A., 2001, 'A Three-Dimensional Passive-Dynamic Walking Robot with Two Legs and Knees', *International Journal of Robotics Research* 20(7): 607-615.

Crapse, T. B. & Sommer, M. A., 2008, 'Corollary discharge across the animal kingdom', *Nature Reviews Neuroscience* 9(8): 587-600.

Dehaene, S., 2011, *The Number Sense: How the Mind Creates Mathematics, Revised and Updated Edition,* New York: Oxford University Press.

Dennett, D.C., 1969, *Content and Consciousness*, London: Routledge.

Dennett, D.C., 1991, *Consciousness Explained,* Boston: Little, Brown & Co.

Devlin, K., 2001, *The Math Gene: How Mathematical Thinking Evolved and Why Numbers Are Like Gossip,* New York: Basic Books.

Dretske, F., 1988, *Explaining Behaviour: Reasons in a World of Causes,* Cambridge: MIT Press.

Ehrsson, H. H., Spence, C. & Passingham, R. E., 2004, 'That's my hand! Activity in premotor cortex reflects feeling of ownership of a limb', *Science* 305, no. 5685: 875-7.

Evans, J. St. B. T., Barston, J. L. & Pollard, P., 1983, 'On the conflict between logic and belief in syllogistic reasoning', *Memory & Cognition* 11, no. 3: 295-306.

Evans, J. St. B. T., 2003, 'In two minds: dual-process accounts of reasoning', *Trends in Cognitive Sciences* 7, no. 10: 454-459.

Feferman, S., 1996, 'Penrose's Gödelian argument', *Psyche* 2: 21-32.

Fodor, J., 1975, *The Language of Thought,* Cambridge: Harvard University Press.

Fodor, J., 1983, *The Modularity of Mind,* Cambridge: MIT Press.

Fodor, J. A., 2000a, 'Why we are so good at catching cheaters', *Cognition* 75(1): 29-32.

Fodor, J. A., 2000b, *The Mind Doesn't Work That Way*, London: The MIT Press.

Fodor, J., Pylyshyn, Z., 1988, ' Connectionism and Cognitive Architecture: A Critical Analysis', *Cognition* 28(1-2): 3-71.

Freuder, E., 1976, 'A computer system for visual recognition using active knowledge', *M.I.T. A.I. Lab. Tech. Rep.* no. 345.

Friston, K., 2005, 'A theory of cortical responses', *Philosophical Transactions of the Royal Society* 360B: 815-836.

Friston, K., 2008, 'Hierarchical models in the brain', *PLoS Computational Biology* 4(11): e1000211.

Gibson, J. J., 1979, *The Ecological Approach To Visual Perception,* Boston: Houghton-Mifflin.

Gigerenzer, G., 1999, *Simple Heuristics That Make Us Smart*, Oxford: Oxford University Press.

Gigerenzer, G., 2008, 'Why heuristics work', *Perspectives on Psychological Science* 3(1): 20-29.

Gilbert, H., 1973, *Thought,* Princeton: Princeton University Press.

Goodman, N., 1978, *Ways of Worldmaking*, Indianapolis: Hackett Publishing Company.

Gödel, K. (translated by Meltzer, B.), 1931/1992, *On Formally Undecidable Propositions Of Principia Mathematica And Related Systems,* New York: Dover Publications Inc.

Haidt, J., 2001, “The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgement” in *Psychological Review* 108(4): 814-34.

Haidt, J., 2013, *The righteous mind: why good people are divided by politics and religion,* London: Penguin Books.

Hansen, T., Olkkonen, M., Walter, S. & Gegenfurtner, K. R., 2006, ’Memory modulates color appearance’, *Nature Neuroscience* 9: 1367-1368.

Helmholtz, H. (translated by Southall, J.P.C.), 1860/1962, *Handbuch der physiologischen optic,* Vol. 3, New York: Dover Publications Inc.

Herz, R.S., Eliassen, J., Beland, S.L. & Souza, T., 2004, ‘Neuroimaging evidence for the emotional potency of odor-evoked memories’, *Neuropsychologia* 42: 371-378.

Hinton, G. E., 2007a, ‘Learning multiple layers of representation’. *Trends in Cognitive Sciences* 11: 428-434.

Hinton, G. E., 2007b, ‘To recognize shapes. First learn to generate images’, in Cisek, P., Drew, T., & Kalaska, J. (Eds.), *Computational Neuroscience: Theoretical insights into brain function,* pp. 535-548, Amsterdam: Elsevier.

Hohwy, J., 2007, ‘The Sense of Self in the Phenomenology of Agency and Perception’, *Psyche* 13(1): 1-20.

Hohwy, J., 2013, *The Predictive Mind*, Oxford: Oxford University Press.

Hohwy, J., 2014, ‘The Self-Evidencing Brain’, *Noûs* 48(1): 1-24.

Hohwy, J. & Paton, B., 2010, ‘Explaining Away the Body: Experiences of Supernaturally Caused Touch and Touch on Non-Hand Objects within the Rubber Hand Illusion, *PLoS ONE* 5(2): e9416.

Horgan, T. & Tienson, J., 1996, *Connectionism and the Philosophy of Psychology,* Cambridge: MIT Press.

Hornsby, J., 2000, ‘Personal and sub-personal; A defence of Dennett’s early distinction’, *Philosophical Explorations* 3(1): 6-24.

Hutto, D.D., 2011a, ‘Enactivism: Why be Radical?’, in Bredekamp, H. & Krois, J.M. (eds.), *Sehen und Handeln*, Berlin: Akademie Verlag.

Hutto, D. D., 2011b, 'Philosophy of mind's new lease on life: Autopoietic Enactivism meets Teleosemiotics', *Journal of Consciousness Studies* 18: 44-64.

Hutto, D.D., 2013a, 'Psychology unified: From folk psychology to radical enactivism', *Review of General Psychology* 17(2): 174-178.

Hutto, D.D., 2013b, 'Exorcising action oriented representations: ridding cognitive science of its Nazgûl', *Adaptive Behavior* 21(3): 142-150.

Hutto, D. D. & Myin, E., 2012, *Radicalizing Enactivism: Basic Minds Without Content*, Cambridge: MIT Press.

Jackendoff, R., 2002, *Foundations of Language,* Oxford: Oxford University Press.

Kahneman, D., 2011, *Thinking Fast and Slow,* London: Penguin.

Kirsh, D. & Maglio, P., 1994, 'On Distinguishing Epistemic from Pragmatic Action', *Cognitive Science* 18(4): 513-549.

Knill, D. C. & Pouget, A., 2004, 'The Bayesian brain: The role of uncertainty in neural coding and computation', *Trends in Neurosciences* 27(12): 712-719.

Konev, B. & Lisitsa, A., 2014, 'A SAT Attack on the Erdos Discrepancy Conjecture', http://arxiv.org/abs/1402.2184, Retrieved 2014-10-09.

Lafer-Sousa, R., Hermann, K. L. & Conway, B. R., 2015, 'Striking individual differences in color perception uncovered by "the dress" photograph', *Current Biology* 25: R545-R546.

Lee, S. A., & Spelke, E. S. (2010). A modular geometric mechanism for reorientation in children. Cognitive Psychology 61: 152-176.

Lucas, J. R., 1961, 'Minds, Machines, and Godel', *Philosophy* 36: 112-127.

Maravita, A. & Iriki, A., 2004, 'Tools for the body (schema)', *TRENDS in Cognitive Sciences* 8(2): 79-86.

Marr, D., 1980, 'Visual information processing: the structure and creation of visual representations', *Philosophical Transactions of the Royal Society* 290B(1038): 199-218.

Marr, D., 1982, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, London: MIT Press (2010).

Marr, D. & Poggio,T., 1979, ‘A Computational Theory of Human Stereo Vision’, *Procesdings of the Royal Society of London, Series B, Biological Sciences* 204(1156): 301-328.

Mataric, M., 1990, ‘Navigating With a Rat Brain: A Neurobiologically-Inspired Model for Robot Spatial Representation’, in Meyer, J-A. & Wilson, S. (eds.), *Proceedings, From Animals to Animats: First International Conference on Simulation of Adaptive Behaviour (SAB-90)*, Cambridge: MIT Press.

Mataric, M., 1992, ‘Integration of Representation Into Goal-Driven Behavior-Based Robots’, *IEEE Transactions on Robotics and Automation* 8(3): 304-312.

McCullough, D., 1995, ‘Can Humans Escape Gödel?’, *Psyche* 2(4).

McGurk, H. & MacDonald, J., 1976,’Hearing lips and seeing voices’, *Nature* 264(5588): 746–8.

McNeill, D., 1992, *Hand and Mind: What Gestures Reveal about Thought*, Chicago: University of Chicago Press.

Meltzoff, A. N. & Moore, M. K., 1997, ‘Explaining facial imitation: A theoretical model’, *Early Development and Parenting* 6: 179-192.

Michalewicz, Z., & Fogel, D. B., 2000, *How to Solve It: Modern Heuristics*, Berlin: Springer.

Miller, G. A. (1956), 'The magical number seven, plus or minus two: Some limits on our capacity for processing information', *Psychological Review* 63(2): 81–97.

Moreno-Bote, R., Knill, D. C. & Pouget, A., 2011, ‘Bayesian sampling in visual perception’, *Proceedings of the National Academy of Sciences USA* 108(30): 12491-6.

Morley, N. J., Evans, J. St. B. T. & Handley, S. J., 2004, ‘Belief bias & figural bias in syllogistic reasoning’, *The Quarterly Journal of Experimental Psychology* 57A(4): 666-692.

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

Noë, A., 2004, *Action in Perception,* London: MIT Press.

Noë, A., 2009, *Out of Our Heads,* New York: Hill and Wang.

Olson, E. T., 1997, *The Human Animal: Personal Identity Without Psychology*, New York: Oxford University Press.

Olson, E. T., 2011, 'The Extended Self', *Minds and Machines* 21: 481-495.

Peacocke, C., 1992, *The Study of Concepts,* Cambridge: MIT Press.

Pearl, J., 1988, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, San Mateo: Morgan Kaufmann.

Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V. & Rizzolatti, G., 1992, 'Understanding motor events: a neurophysiological study', *Experimental Brain Research* 91(1): 176-180.

Penrose, R., 1994, *Shadows of the Mind*, Oxford: Oxford University Press.

Penrose, R., 1996, 'Beyond the Doubting of a Shadow', *Psyche* 2(23).

Pinker, S., 1997, *How the Mind Works*, New York: W. W. Norton & Company.

Poggio, T., 1981, 'Marr's computational approach to vision', *Trends in NeuroSciences* 4(10): 258-262.

Prinz, J., 2006, 'Putting the Brakes on Enactive Perception', *Psyche* 12(1).

Putnam, H., 1961, 'Brains and Behaviour' in Block, N., 1980, *Readings in Philosophy or Psychology Vol. 1,* London: Methuen.

Putnam, H., 1967, 'Psychological Predicates', in Capitan, W. & Merrill, D. (eds.), *Art, Mind, and Religion,* Pittsburgh: University of Pittsburgh Press.

Putnam, H., 1975, *Mind, Language and Reality,* Cambridge: Cambridge University Press.

Putnam, H., 1981, *Reason, Truth, and History,* Cambridge: Cambridge University Press.

Requarth, T. & Sawtell, N. B., 2014, 'Plastic corollary discharge predicts sensory consequences of movements in a cerebellum-like circuit', *Neuron* 82(4): 896-907.

Rizzolatti, G. & Craighero, L., 2004, 'The mirror-neuron system', *Annual Review of Neuroscience* 27(1): 169-192.

Rowlands, M., 2010, *The new science of the mind: from extended mind to embodied phenomenology,* London: MIT Press.

Schliemann, A. D. & Carraher, D. W., 2002, 'The Evolution of Mathematical Reasoning: Everyday versus Idealized Understandings', *Developmental Review* 22: 242-266.

Scott, M., 2013, 'Corollary discharge provides the sensory content of inner speech', *Psychological Science* 24(9): 1824-30.

Searle, J., 1980, ‘Minds, Brains and Programs’, *Behavioral and Brain Sciences* 3: 417-57.

Shepard, R.N. & Meltzer, J., 1971, ‘Mental Rotation of Three-Dimensional Objects’, *Science* 171: 701-703.

Shoemaker, S., 1999, 'Eric Olson, The Human Animal', *Noûs* 33(3): 496-504.

Stone, J. V., 2013, *Bayes’ Rule: A Tutorial Introduction to Bayesian Analysis*, Sheffield: Sebtel Press.

Tenenbaum, J.M. & Barrow, H.G., 1977, ‘Experiments in interpretation-guided segmentation’, *Artificial Intelligence* 8(3): 241-274.

Tooby, J. & Cosmides, L., 1992, ‘Cognitive Adaptations for Social Exchange’, in Barkow, J., Cosmides, L. & Tooby, J. (eds.), *The Adapted Mind: Evolutionary psychology and the generation of culture,* New York: Oxford University Press.

Turing, A., 1936, ‘On Computable Numbers, with an Application to the Entscheidungsproblem’, *Proceedings of the London Mathematical Society,* 42: 230-265.

Turing, A., 1950, 'Computing machinery and intelligence', *Mind* 59: 433-460.

Tversky, A. and Kahneman, D., 1974, 'Judgement under Uncertainty: Heuristics and Biases', *Science* 185(4157): 1124-1131.

Ullman, S., 1979a, ‘The interpretation of structure from motion’, *Proceedings of the Royal Society of London,* B 203: 405-426.

Ullman, S., 1979b, *The interpretation of visual motion*, London: MIT Press.

Varela, F.J., 1987, ‘Laying Down a Path in Walking’, *Cybernetics* 2: 6-15.

Varela, F.J., Thompson, E. & Rosch, E., 1991, *The Embodied Mind: Cognitive science and human experience,* Cambridge: MIT Press.

Warren, R.M., 1970, ‘Perceptual Restoration of Missing Speech Sounds’, *Science* 167(3917): 392-393.

Wason, P.C., 1968, ‘Reasoning about a rule’, *The Quarterly Journal of Experimental Psychology* 20(3): 273-281.

Wilkinson, J., 2013, ‘Largest neuronal network simulation achieved using K computer’, http://www.riken.jp/en/pr/press/2013/20130802_1/, retrieved 2016-09-14.

Williams, B., 1973, *Problems of the Self, Philosophical Papers 1956 - 1972,* Cambridge: Cambridge University Press.

Wittgenstein, L. (translated by Anscombe G.E.M.), 1953/2001, *Philosophical Investigations,* Oxford: Blackwell Publishing Ltd.

Whitworth, B. and Ryu, H., 2007, 'A Comparison of Human and Computer Information Processing' in Pagani, M. (ed.), *Encyclopaedia of Multimedia Technology and Networking*, London: IGI Global.

Wolfe, J.M., 1983, 'Influence of spatial frequency, luminance, and duration on binocular rivalry and abnormal fusion of briefly presented dichoptic stimuli', *Perception* 12(4): 447-456.

Wurtz, R. H., 2013, 'Corollary discharge in primate vision', *Scholarpedia* 8(10): 12335.

Yamamoto, S. & Kitazawa, S., 2001, 'Sensation at the tips of invisible tools', *Nature Neuroscience* 4(10): 979-80.

Zadeh, L. A., 1965, 'Fuzzy sets', *Information and Control* 8(3): 338-353.

Zhu, Q., & Bingham, G. P. (2010). 'Human readiness to throw: The size-weight illusion is not an illusion when picking the best objects to throw.' *Evolution and Human Behavior* 32(4): 288–293.